J. Vis. Commun. Image R. 80 (2021) 103261

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Multiple attention encoded cascade R-CNN for scene text detection ☆

Yirui Wu [a,b], Wenxiang Liu [a,b], Shaohua Wan [c,d,*]

[a] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Fochengxi Road, Jiangning District, Nanjing City, 210093, China

[b] College of Computer and Information, Hohai University, Fochengxi Road, Jiangning District, Nanjing City, 210093, China

[c] School of Information and Safety Engineering, Zhongnan University of Economics and Law, South Nanhu Road, Hongshan District, Wuhan City, 430073, China

[d] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Keywords:
Cascade R-CNN
Deep representation learning
Applications to robust image recognition
Multiple attention encoding
Scene text detection
Multi-oriented text

ABSTRACT

Inspired by instance segmentation algorithms, researchers have proposed quantity of segmentation-based methods for text detection, achieving remarkable results on scene text with arbitrary orientation and large aspect ratios. Following their success, we believe cascade architecture and extracting contextual information in multiple aspects are powerful to boost performance on the basis of segmentation-based methods, especially in decreasing false positive texts in complex natural scene. Based on such consideration, we propose a multiple-context-aware and cascade CNN structure, which appropriately encodes multiple categories of context information into a cascade R-CNN framework. Specifically, the proposed method consists of two stages, i.e., feature generation and cascade detection. During the first stage, we define ISTK (Isolated Selective Text Kernel) module to refine feature map, which sequentially encodes channel-wise and kernel-size attention information by designing multiple branches and different kernel sizes in isolate form. Afterwards, we build long-range spatial dependencies in feature map via non-local operations. Built on contextual feature map, Cascade Mask R-CNN structure progressively refines accurate boundaries of text instances with multi-stage framework. We conduct comparative experiments on ICDAR2015 and 2017-MLT datasets, where the proposed method outperform comparative methods in terms of effectiveness and efficiency measurements.

1. Introduction

Scene text detection is still challenging, due to factors like multi-language, arbitrary-orientation and curving situations. Facing these difficulties, researchers have proposed quantity of methods by regarding text as an instance of segment, which is the core idea of segmentation based methods. On the basis of segmentation methods which have achieved significant detection results facing arbitrary-orientation and curving problems, we believe that applying cascade framework and extracting context information can boost detection performance, especially in decreasing false positive detections on complex background.

Essentially, cascade is a classic and powerful architecture, which has been successfully applied to boost performance on various tasks with ideas of multi-stage refinement. It is noted that most of the current network uses a relative low value as IOU threshold, thus causing noisy detections. However, simply increasing IOU threshold will lead accuracy of detection to decrease. To relieve this issue, Cascade R-CNN [1] constructs a sequence of object detectors trained with increasing IoU thresholds, which makes the threshold setting more suitable for training modules of each stage. Therefore, false positive results can be effectively eliminated.

To fully leverage relationship between detection and segmentation, Chen et al. [2] propose Hybrid Task Cascade (HTC) for instance segmentation, which interweaves detection and segmentation as a joint multi-stage processing to achieve better refinement on both tasks. In order to better distinguish the foreground and background, they further add a convolution branch to make full use of spatial context information. Through multi-stage and multi-information fusion, they successfully gain more powerful detection capabilities. Inspired by these two state-of-the-art tasks, we argue that text detection can adopt cascade architecture to boost performance as well, which not only helps prevent overfitting during training due to exponentially vanishing positive samples, but also relieves the burden of manually defining proper IOU parameters to construct effective text detectors.

Abundant context information is essential to deal with complexity brought by ambiguity property of visual scene, where different categories of attention modules are widely used to perform the modeling

☆ This paper has been recommended for acceptance by Petia Radeva.

* Corresponding author at: College of Computer and Information, Hohai University, Fochengxi Road, Jiangning District, Nanjing City, 210093, China.
E-mail addresses: wuyirui@hhu.edu.cn (Y. Wu), lwxhhu@hhu.edu.cn (W. Liu), shaohua.wan@ieee.org (S. Wan).

https://doi.org/10.1016/j.jvcir.2021.103261
Received 23 November 2020; Received in revised form 29 June 2021; Accepted 2 August 2021
Available online 18 August 2021

of context information. For example, SENet [3] builds subnet by convolutional layers to assign different weights across channels, which greatly improves classification accuracy on complex scenes and wins the champion of 2017 ImageNet Classification Competition. SKNet [4] further proposes to dynamically adjust size of the convolution kernel based on multiple scales of input information, which successfully builds kernel-size attention by adopting task-specified size of receptive field. Inspired by these tasks, we argue that appropriately encoding multiple categories of attention information could help accurately detect text by involving additional and informative scene context information, especially in a complex natural environment. Furthermore, general attention modules could be adjusted to further promote text detection with special text-oriented designs.

Based on all these considerations, we propose a task-specified cascade R-CNN structure encoding multiple context-aware information. Specifically, the proposed method generates feature map with abundant context information, i.e., spatial, channel, and kernel-size context information extracted by the proposed ISTK module and NLNet module in cascade form. After generating feature map, a cascade R-CNN structure is applied to detect texts in multiple stages, which are sequentially constrained with higher IoU thresholds to delaminate false positive texts. After conducting experiments on ICDAR 15 and ICDAR 17 MLT datasets, we prove the proposed structure is powerful to generate correct text samples in complex nature scene scenarios, due to modeling of abundant context information and utilization of cascade structure.

Our contributions can be concluded as follows:

- We propose a cascade R-CNN structure encoding multiple context-aware information, which involves significant power of cascade framework and context modeling for accurate detection performance, especially in complex natural scene.
- As far as we know, the proposed method firstly involves three different kinds of attention information, i.e., kernel size, channel-wise and spatial attention, to construct task-specified feature map for text detection.
- We specially design ISTK structure to fit with task of text detection, which utilize isolated form of information fusion to generate informative feature map from aspects of channel-wise and kernel-size attention.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work on relative aspects. In Section 3, details of the proposed structure is discussed, including Total network architecture, designs of ISTK and NLNet modules. Section 4 shows our experimental results with several comparative methods. Finally, Section 5 concludes the paper.

2. Related work

We introduce relevant research that inspired us to design the proposed method in the section, including scene text detection and attention model.

2.1. Scene text detection

Due to the wide usage of deep learning models, we can see quantity of mutual applications in different domains [5–10]. Most existing scene text detection methods built on deep learning structures can be divided into two categories, namely regression-based and segmentation-based methods.

Regression-based text detection aims to detect text instance as a common object. For example, Ma et al. [11] propose RRPN, which adds a rotation angle to the detection frame to detect multi-directional text, due to the multi-directionality property of texts. Liao et al. [12] design Textbox++, which not only puts forward the idea of using text recognition to assist text detection, but also proposes a novel function to detect arbitrary text. In order to solve the problem of extremely long text, Zhang et al. [13] present LOMO, which localizes the text progressively with multiple trial times. Most recently, Bai et al. [14] add an obliquity factor to their proposed network, which is able to detect horizontal objects and non-horizontal objects with accurate performance. From the above research, we can conclude that the regression-based text detection methods require additional algorithm design and computing power to solve the problem of rotated texts.

Segmentation based methods obtain masks and bounding boxes according to text instances progressively. Early, Mask R-CNN [15] firstly modify the step of ROI pooling to ROI align on the basis of faster R-CNN [16–19], and then add a mask module for accurate instance segmentation. Later, PANet [20] use method of bottom-up path augmentation to perform tasks of information path shorting, adaptive feature pooling and fully-connected fusion, achieving better mask than former Mask R-CNN algorithm. On the basis of Mask R-CNN, Liao et al. [21] propose Mask TextSpotter, which is capable to detect texts of various shapes and recognize characters. Afterwards, Pixel-link [22] use a convolutional network to perform two tasks, i.e., text/non-text prediction and link prediction, where determined pixels can be utilized to connect link prediction, thus generating convinced shape on curved texts.

Since close text can cause multiple texts to be mistaken for one text, Wang et al. [23] propose PSENet to generate different scales of kernels for each text instance, and gradually expand the minimal scale kernel to the text instance. Owing to the guidance of semantic information, SPCNet [24] propose to involve more context information, resulting in stronger detection capabilities in complex natural scenes. Further, TextSnake [25] regard texts as disks, meanwhile different disks have different radius and directions. In this way, the flexibility of their proposed network is much increased by novel representation of texts. Afterwards, TextFuseNet [26] obtains richer text features by fusing three different categories of features, i.e., character level, word level and global level. Rich features enhance the detection ability and environmental adaptability of their proposed network. Most recently, ContourNet [27] generates more accurate anchors through Adaptive-RPN, and uses Local Orthogonal Texture-aware Module model the local texture information in two orthogonal directions, which successfully reduces false positive results.

The proposed method is a segmentation based method, which is capable to deal with situations of oriented and curved texts. Furthermore, the proposed method tries combination of cascade framework and context information modeling, which successfully improves the accuracy and robustness in locating texts in complex scenes.

2.2. Attention model

People usually focus on objects themselves and ignore the background to obtain important information. Based on such characteristics, attention models have been applied in deep learning area and achieved significant performance in multiple domains.

Researchers firstly try to build spatial and temporal attention, which coincides with the visual principle of humans, i.e., focus on part of scene or a duration within a sequence. With better understanding of structures and functions of CNN network, channel-wise attention is exploited to re-weight conv-layer feature map produced by different layers of CNN, thus offering information on which feature channel is informative during processing. For example, Woo et al. [28] propose Convolutional Block Attention Module (CBAM), which redistributes attention on both channels and spaces to enhance input feature map. Compared with CBAM, Cao et al. [29] propose Global context network (Gcnet), which simplifies non-local neural network and acts to be more efficient and faster in run-time. To obtain more accurate detection results with multiple attention modules, Zhao et al. [30] propose PFAN (Pyramid Feature Attention Network), which adopt spatial attention mechanism for low-level network structures and channel attention mechanism for high-level network.

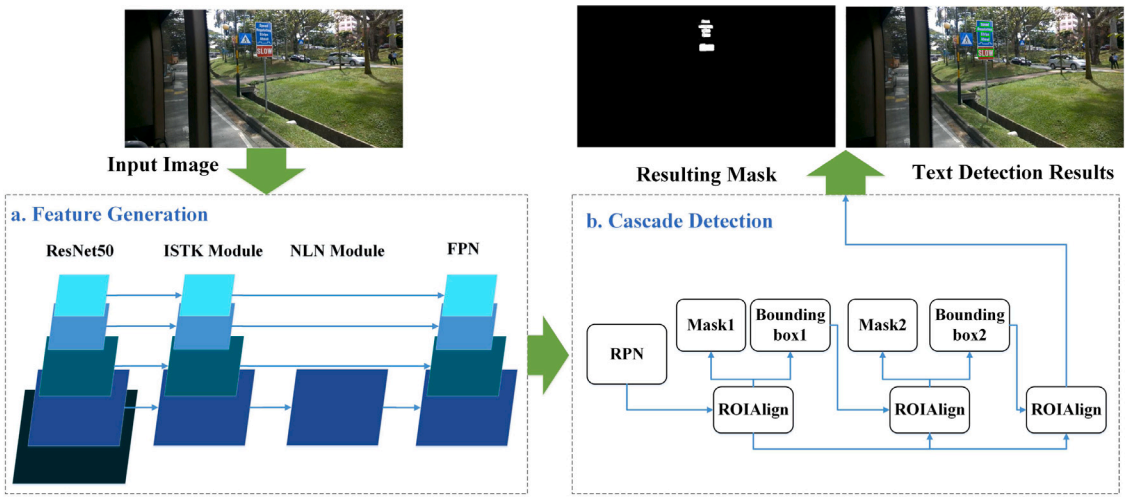

Fig. 1. Network architecture of the proposed method, which consists of feature generation and cascade detection.

In parallel with channel-wise attention, researchers propose receptive field to enhance representation ability of feature map. With the idea of dynamic setting on receptive field, Liu et al. [31] propose RFB (Receptive Field Block) network, which uses dilated convolutions to obtain more noteworthy information on size of receptive field. Further emphasizing on the importance of receptive fields, Li et al. [4] design SKNet (Selective Kernel Network) to assign weights for both channel related information and size of convolution kernel. Most related to the proposed method, Xiao et al. [32] take receptive field, spatial and channel attention information into account, and propose TCAM (Text-Context-Aware Module) to solve the problem of multi-oriented and multi-language, which proves the effectiveness of applying different kinds of context information on task of text detection.

Inspired by these attention models, the proposed method proposes the ISTK module, which not only properly encodes two categories of context information, i.e., channel-wise and kernel-size, but also specially design the module in a isolated form to coincide with inherent characteristics of text in scenes. In order to capture long-range dependencies, we further introduce NLNet to refine low-level feature map by introducing spatial attention information.

3. The proposed method

In this section, we firstly involve text-context-aware characteristics by multiple context modules to generate informative feature map. Based on generated feature map, we construct a cascade R-CNN structure, which performs instance-level text mask segmentation task in a sequential manner, thus performing a higher selective operation against close false positives. We organize this section by first illustrating the total network architecture, and then describing structures of ISTK and NLNet modules in details.

3.1. Total network architecture

As shown in Fig. 1, the overall network structure is mainly composed of two stages, i.e., feature generation and cascade detection, where we perform sequential and step-wise text detection on informative feature map.

Specifically, We first adopt ResNet-50 to extract basic feature representation for further processing. The reason to adopt ResNet-50 as backbone network lies in the fact that texts can be regarded as a special kind of object and ResNet-50 has been proved to be highly effective in classifying different categories of objects with distinguishing feature representation. The whole process can be represented as

$$F = Res_{50}(I), where\ F = \{F_i | i = 1, 2, 3, 4, 5\} \tag{1}$$

where i is the layer index of feature map, I refers to the input scene image containing texts, and function $Res_{50}()$ refers to the operations of backbone network structure, i.e., ResNet-50. It is noted that we define scale factor as 0.5 to deliver scalable feature map from the largest scale (1st) to the smallest scale (5th).

After feature extraction, we refine the extracted feature map through the proposed ISTK modules by introducing context information on channel and kernel size, which can be represented as

$$\tilde{F}_i = f_{ISTK}(F_i), where\ i = \{2, 3, 4, 5\} \tag{2}$$

where $\tilde{F}_i$ is the enhanced feature map with the proposed ISTK module, and function $f_{ISTK}()$ dynamically assigns weights on convolution kernel size and channels based on input feature map F_i. It is noted that we abandon the first scale of feature map with largest size, due to the consideration of computation speed.

Compared with SKNet [4] which addresses the modeling of kernel size context information, we offer separate structures on feature extraction and attention modules, rather than embedding attention module into the structure of feature extraction. By separating ISTK and ResNet-50, the proposed method could not only reuse parameters of the pre-trained ResNet-50 model with less modification and training, but also makes the proposed ISTK module easier to be applied into any existing networks as attention modules.

To capture long-range dependencies among feature map, we propose to enhance the lowest layer of feature map with NLNet module by introducing spatial attention information, which can be expressed as

$$\hat{F}_2 = f_{NLN}(\tilde{F}_2) \tag{3}$$

where function $f_{NLN}()$ refers to operations in NLNet module. NLNet module is capable to model the interaction between pixels, which could be comprehended as globally spatial information, thus introducing spatial context information into the second scale of feature map. The reason to apply NLNet module only on the second scale lies in the fact that low-level feature, extracted by lower layer of ResNet-50, corresponds to color, texture, and other low-level visual features, and high-level feature carries semantical and less information. In order to model spatial context information among long-range regions, it is more suitable to use NLNet to deal with informative and abundant feature map generated by lower layer of the network.

After processing with ISTK and NLNet modules to introduce context information, generated feature map are fed into FPN (Feature Pyramid Network) for further feature refinement, which could help network

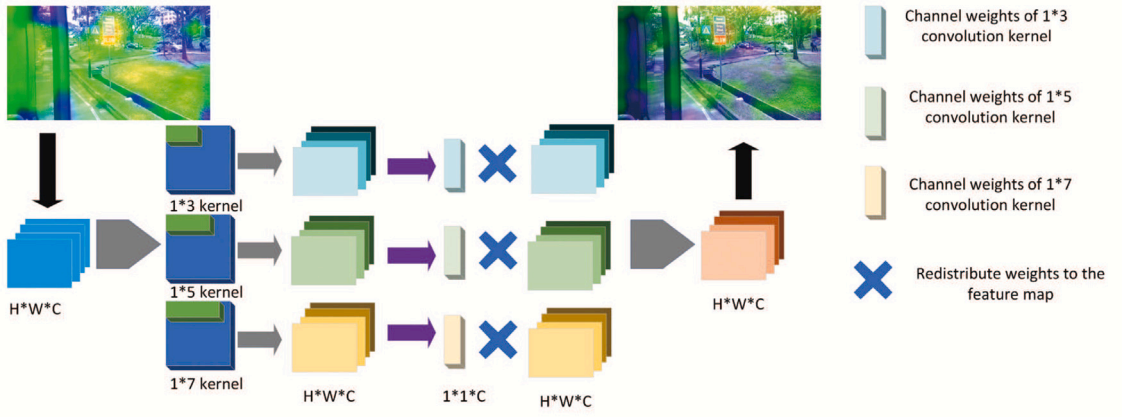

Fig. 2. Architecture of the proposed ISTK module, i.e., multiple attention module, which shows the reconstruction of feature maps through different channels and different convolution kernels.

detect text instances with different scales. The whole process can be represented as

$$P = f_{FPN}(\hat{F}_2, \tilde{F}_3, \tilde{F}_4, \tilde{F}_5), where\ P = \{P_j | j = 2, 3, 4, 5\} \tag{4}$$

where function $f_{FPN}()$ refers to operations of FPN.

After all processing steps in feature generation stage, refined feature map are regarded as input of cascade detection, where RPN (Region Proposal Network) firstly locates possible anchors of text instances b_0:

$$b_0 = f_{RPN}(P) \tag{5}$$

where function $RPN()$ refers to operations of RPN.

After generating anchor candidates b_0, we utilize cascade R-CNN structure to progressively compute bounding boxes. It is noted that cascade R-CNN structure regard the generated bounding box in one stage as anchors, which is fed into ROIAlign module located in the next stage. Such design with increasing IoU thresholds, could help the total network be sequentially more selective against close false positives. We represent processing steps of Cascade R-CNN as follows:

$$m_k = f_{M,k}(R_a(b_{k-1}, P)) \tag{6}$$

$$b_k = f_{B,k}(R_a(b_{k-1}, P)), where\ k = \{1, 2, 3\} \tag{7}$$

where m_k and b_k represent segmentation mask and bounding box generated in the kth stage respectively, function $R_a()$ performs align operations on regions of interest to generate new feature map based on detected bounding box in last step b_{k-1} and feature maps P, and function $f_{B,k}()$ and $f_{M,k}$ denote the box head and mask head to generate corresponding bounding box and mask for the kth stage respectively. After three stages of cascade processing, false positive results caused by the environmental impact of natural scenes can be eliminated. It is noted that we modify classification task of general cascade structure into the process of mask generation, since the core of text detection task is to located text.

The loss function of proposed network is represented as follows :

$$Loss = L_{RPN} + L_{mask,3} + L_{cls,k} + L_{reg,k}, where\ k = \{1, 2, 3\} \tag{8}$$

where the loss function can be divided into four parts, i.e., RPN, mask, classification and regression. Since the generated masks of the first and second stages in the cascade network have no effect on the final results, they are not included in the calculation of loss function.

3.2. Design of ISTK module

In this subsection, we mainly describe the construction steps of the proposed ISTK module.

Inspired by SKNet [4], we propose ISTK (Isolated Selective Kernel) module to enhance feature representation by dynamically assigning different weights to feature map generated by different convolution kernels, where we show its structure in Fig. 2. Compared with feature fusion stage in SKNet, we specially design the proposed ISTK in an isolate form for the task of text detection. The reason to apply isolate form lies in two aspects. Firstly, fusion will lead to information loss, which could result in low representation ability of the generated feature map. Secondly, low-level feature map are essentially important to locate text anchors, such as shape, texture, color and so on. Fusion in higher level leads to abstraction of information, thus generating semantical meanings of feature map. Meanwhile, fusion in lower level of feature map will result in misunderstanding of feature map due to loss of information amount caused by fusion.

Moreover, we modify size of convolution kernel from $n * n$ to $1 * n$, where n refers to the width of kernel and the modified kernel size fits with the long aspect ratio property of text instances. With such specific design, unique text characteristics can be ensured to be well enhanced, thus offering guarantees on robust and accurate text detection. Above all, specific designs in the proposed ISTK module help the proposed method to enhance distinguish ability in classifying texts or not. To prove the effectiveness of ISTK module, we show detection samples of using ISTK module and original SKNet in Fig. 3, where we can observe ISTK module helps generate more promising detection results.

We design ISTK module with three steps, i.e., split, process and select. During the first split step, input feature maps are processed by using convolution kernels of 1*3, 1*5, 1*7 respectively, which compute feature maps with different receptive fields:

$$K_{i,\lambda} = f_{conv,\lambda}(F_i), where\ \lambda = \{1, 2, 3\} \tag{9}$$

where λ refers to index of convolutional kernels with different sizes, i.e., $\{1 * 3, 1 * 5, 1 * 7\}$, function $f_{conv,\lambda}()$ refers to convolutional operation with predefined kernel sizes. Since spatial space occupied by the utilized 1*n convolution kernels are already small enough, the proposed method does not apply dilated convolution to process the convolution kernels.

After extracting feature of different receptive fields with different convolutional kernels, we further dynamically calculate weights w_i based on feature maps generated with different receptive fields:

$$w_i = softmax(\bigcup_{\lambda=1}^{3} f_{fc}(f_{avg}(K_{i,\lambda}))) \tag{10}$$

where function $f_{avg}()$ and f_{fc} represent operation of global average pooling, and two fully connected operations, respectively. After processing of operations in Eq. (10), we will obtain feature map with 1*1*C size. In the generated feature map, we could use $C_{i,\lambda,l}$ to represents a specific feature channel, where l refers to index of feature channel and the corresponding convolution kernel size is 1*3 if λ equals 1. After

Fig. 3. Sample detection results by utilizing either ISTK module (Left) or SKNet module (Right), where red rectangles refer to false positive or false negative results.

processing by the softmax function, the corresponding weight for $C_{i,\lambda,l}$ can be calculated as

$$w_{i,\lambda,l} = \frac{e^{C_{i,\lambda,l}}}{e^{C_{i,\lambda=1,l}} + e^{C_{i,\lambda=2,l}} + e^{C_{i,\lambda=3,l}}} \tag{11}$$

In this way, weights for feature channels computed by different convolution kernels can be obtained for further processing.

Based on the calculated weights, we first multiply weights with the feature map to encode context information and then perform feature fusion to get the final feature map, which could be represented as

$$\tilde{F}_i = relu(sum(w_i * F_i)) \tag{12}$$

where F_i refers to generated feature map by ResNet-50, and function $relu()$ refer to ReLu activation function.

3.3. Design of NLNet module

In this subsection, we utilize Non-local Neural network (NLNet) to encode spatial context information into the extracted feature map.

As shown in Fig. 4, we design NLNet module to model spatial context information between pixels in feature map with long distance. In fact, pixel-level spatial attention is hard to abstract, since receptive field can only process local information so that pixels in the feature map can only be associated with local neighboring pixels for context modeling. To solve the problem of pixel-level spatial context modeling, convolution operations in NLNet module are all equipped with 1×1 convolutions. Choosing 1×1 convolution kernel can not only extract information with high effective, but also reduce dimensions of output features and the total amount of parameters.

Following the description on structure of NLNet module, we could calculate the middle representation of feature map $F_{2,m}$ wotj

$$F_{2,m} = f_{1c}(\tilde{F}_2) * softmax(f_{1c}(\tilde{F}_2) * f_{1c}(\tilde{F}_2)) \tag{13}$$

where function $f_{1c}()$ refers to 1×1 convolution kernel to modify dimension for processing. It is noted that multiply any two branches in NLNet module is to build the connection between any these two points of the feature map, which can be regarded as a procedure of modeling spatial context information in long range.

Afterwards, we process $F_{2,m}$ with the following equation:

$$\hat{F}_2 = f_{1c}(F_{2,m}) + \tilde{F}_2 \tag{14}$$

where we use additional 1*1 convolutional kernel to reshape $F_{2,m}$ as the same size as input feature map $\tilde{F}_2$. In this way, the extracted long-range spatial context information can be successfully encoded into the feature map.

4. Experiment

In this section, we first introduce our dataset. Then, we conduct ablation experiments to show the effectiveness of the proposed structure. Then, we carry out comparison experiments and associated analysis with several latest text detection methods. Finally, we offer implementation details for readers' convenience.

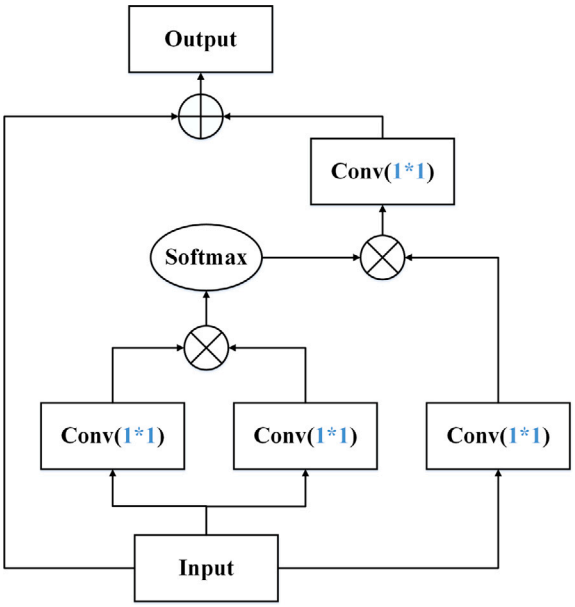

Fig. 4. Architecture of the proposed NLNet module.

Table 1
Performance comparisons with different structure designs on ICDAR2015 and ICDAR2017-MLT datasets.

Method	Precision	Recall	F-measure
Cascade Mask R-CNN	86.7	79.7	83.1
SKNet	**88.8**	81.1	84.8
ISTK (n*n)	**88.8**	81.8	85.1
ISTK (1*n)	88.4	**82.1**	**85.2**

Table 2
Performance comparisons with different structure designs on ICDAR2017-MLT dataset.

Method	Precision	Recall	F-measure
Cascade Mask R-CNN	76.8	59.6	67.1
SKNet	75.5	**61.2**	67.6
ISTK (n*n)	**77.8**	60.3	67.9
ISTK (1*n)	76.6	61.1	**68.0**

4.1. Datasets

In this paper, we use the ICDAR2015 and the ICDAR2017-MLT dataset, where ICDAR2015 dataset contains 1000 training and 500 test pictures, ICDAR2017-MLT includes 7200 training, 1800 verification, and 9000 test pictures. It is noted that ICDAR2015 dataset contains only English texts, meanwhile ICDAR2017-MLT dataset contains multiple languages. Moreover, we could observe the scene category in ICDAR2017-MLT dataset is more diverse and complex than scenes in ICDAR2015 dataset, leading us to believe that ICDAR2017-MLT dataset is more challenging for text detectors to locate text instances than ICDAR2015 dataset.

4.2. Ablation experiments and analysis

We design different structure designs for ablation experiments in Tables 1 and 2, where Cascade Mask R-CNN refers to only apply general cascade structure on text detection without feature map enhancement, SKNet performs feature enhancement with branch feature fusion and n*n kernel size, ISTK(n*n) and (1*n) refer to the proposed method with different settings on kernel size of convolutional filters.

Table 3
Performance comparisons with different stage parameter designs on ICDAR2015 and ICDAR2017-MLT datasets.

Method	Dataset	Precision	Recall	F-measure
Two-stage	ICDAR2015	87.8	82.3	85.0
Three-stage	ICDAR2015	88.4	82.1	85.2
Two-stage	ICDAR2017	76.3	61.2	67.9
Three-stage	ICDAR2017	76.6	61.1	68.0

Table 4
Comparisons of detection performance on ICDAR2015 dataset.

Method	Precision	Recall	F-measure
He et al. [33]	85.0	80.0	82.0
Textbox++ [12]	87.8	78.5	82.9
EAST [34]	83.2	78.3	80.7
PixelLink [22]	85.5	82.0	83.7
SegLink [35]	73.1	76.8	75.0
DMPNet [36]	68.2	73.2	70.6
FoTs [36]	**91.0**	**85.2**	**88.0**
The proposed	88.4	82.1	85.2

Table 5
Comparisons of detection performance on ICDAR2017-MLT dataset.

Method	Precision	Recall	F-measure
He et al.[33]	76.7	57.9	66.0
Lyu et al. [37]	83.8	55.6	66.8
TDN SJTU2017 [38]	64.2	47.1	54.3
SARI FDU RRPN [11]	71.2	55.5	62.4
FoTs [39]	**81.0**	57.5	67.3
The proposed	76.6	**61.1**	**68.0**

From Tables 1 and 2, we can observe that both ISTK and SKNet achieves better performance than Cascade Mask R-CNN, which proves the effectiveness of attention models by introducing informative and abundant context information. Comparing between ISTK(n*n) and SKNet, we could find an improvement in F-measure by utilizing isolate form rather than feature fusion, which proves that retaining the whole information of low-level feature map contributes to accurate localization of texts. Moreover, utilizing 1*n kernel greatly improves recall and slightly decreases precision, when comparing performance between ISTK(n*n) and (1*n). This phenomenon can be explained by the fact that long aspect ratio design leads text instance to be easily located in complex background, meanwhile long objects like sticks would be easy to be misclassified as text as well.

In Table 3, we further compare detection performance of the proposed method on both ICDAR2015 and ICDAR2017-MLT datasets, using different stage parameter, i.e., 2 and 3 stages. In three-stage training, we use 0.5, 0.6, and 0.7 as thresholds, meanwhile we use 0.5, 0.6 as thresholds. We can clearly observe that three-stage parameter setting performs better than two stages, represented by F-measure. However, the recall performance slightly decreases, due to an additional stage of processing.

4.3. Comparative experiment and analysis

We offer comparisons between the proposed method and several latest methods in Tables 4 and 5. It is noted that the proposed method achieve the best F-measure on both ICDAR2015 and ICDAR2017-MLT datasets, which proves the effectiveness of combining strength of cascade structure and context modeling. In fact, the proposed network not only utilize cascade structure to boost performance by eliminating false positives, but also encodes rich context information, i.e., kernel size, channel-wise and spatial attention, to accurately locate texts in complex natural scenes by constructing ISTK and NLNet modules.

Table 6
Comparisons of FPS on ICDAR2015 dataset.

Method	FPS
PixelLink [22]	7.3
Lyu et al. [37]	3.6
TextSnake [25]	1.1
FoTs [39]	7.8
The proposed	7.5

It is noted that PixelLink [22] achieves a slightly higher recall than the proposed method on ICDAR2015 dataset. The reason lies in the fact that the core idea of PixelLink is a classification task on text related pixels and links, which is easier to train and achieve higher accuracy than the proposed method. Meanwhile, both FoTs [39] and Lyu et al. [37] achieve a much higher precision than the proposed method on ICDAR2017-MLT datasets. It could be explained by the fact that FoTs [39] introduce text recognition to improve performance of text detection, which utilize more input information for detection than the proposed method. Lyu et al. [37] perform more precise detection by dividing text instance into corners. Their proposed method is much complicated in structure and requires relatively high computation cost. The result of Fots on ICDAR2015 is 88.0, which is better than our proposed model. Essentially, Fots not only uses ICDAR2013, ICDAR2015, ICDAR2017-MLT, and Synth800k to train the model, but also uses the OHEM to learn difficult samples. The proposed method only selects 1000 images in ICDAR2017-MLT and ICDAR2015 images for training, which is the main reason that we achieve worse results than Fots. Since the main difficulty of ICDAR2017-MLT lies in the fact of multiple languages, adding additional datasets could not be helpful to improve results on ICDAR2017-MLT. Therefore, the proposed method is better than FoTs when testing on ICDAR2017-MLT dataset.

In order to better show the effect of ISTK, we use heat maps to visualize the experimental results in Fig. 5. Based on the results represented at the second row of Fig. 5 where we directly using feature maps obtained by Cascade Mask R-CNN to display the heat map, the text information can be accurately obtained. However, many areas which are not related to text are also marked in red as text regions, which can be regarded as false positive regions. Due to the lack of context information, the ability to isolate the background of Cascade Mask R-CNN is not effective. Although the results of ISTK shown in third row have errors in marking the non-text areas as red, the size of areas is much smaller than that achieved by Cascade Mask R-CNN. Essentially, ISTK uses a convolution kernel with a different aspect ratio to reshape the feature maps. Therefore, the non-text area in the third row displays a relatively long red heat map, which shows the enhanced ability of ISTK module to capture the shape of the long aspect ratio after using the 1*n convolution kernels.

We offer FPS for comparisons in Table 6. Since ISTK uses the Dilated convolution kernel and 1*n convolution kernel, the entire model is lightweight and the detection speed is faster than many text detection models. We show detection samples achieved by the proposed method and Cascade Mask R-CNN in Figs. 6 and 7, where we can notice involving multiple categories of context information could greatly improve performance on text detection.

4.4. Implementation details

All of these experiments are performed on four GTX 1080 GPU cards. We use SGD optimizer for training by setting the momentum as 0.9 and weight decay as 0.0001. We set parameters of initial learning

Fig. 5. The original images, heat maps achieved by Cascade Mask R-CNN and the proposed method are shown in the first, second and third row.

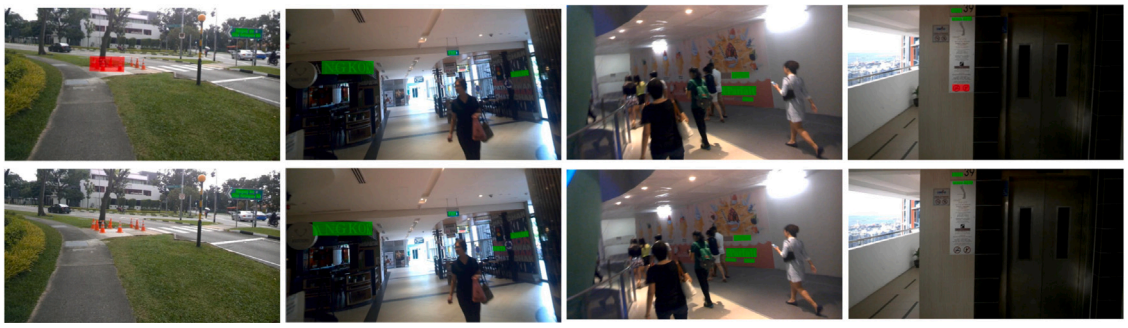

Fig. 6. Detection samples achieved by Cascade Mask R-CNN (the first row) and the proposed method (the second row) on ICDAR2015 dataset.

Fig. 7. Detection samples achieved by Cascade Mask R-CNN (the first row) and the proposed method (the second row) on ICDAR2017-MLT dataset.

rate as 0.01 and the minimum learning rate as 0.0001. The three thresholds adopted by Cascade Mask R-CNN is 0.5, 0.6, and 0.7, respectively. We use SmoothL1Loss as the loss function to make the entire training process smoother, meanwhile we use gradient clip to prevent gradient explosion.

5. Conclusion

In this paper, we propose a multiple-context-aware and cascade CNN structure to detect multi-oriented and Multi-lingual text in natural scene images. Due to the extracted rich context information and cascading structure, text can be detected more accurately by reducing false positive results caused by complex natural environment. Experiments show the proposed method achieves better performance than several latest methods on public datasets. Our future plan is to explore idea of unsupervised learning to relieve the concern on overfitting performance of text detection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by National Key R&D Program of China under Grant No. 2018YFC0407901, the Fundamental Research Funds for the Central Universities under Grant No. B200202177, 31412111303, 31512111310, the open project from the State Key Laboratory for Novel Software Technology, Nanjing University, under Grant No. KFKT2019B17, and the National Natural Science Foundation of China under Grant No. 61702160, 62172438.

References

[1] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: Proceedings of Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[2] K. Chen, J. Pang, J. Wang, et al., Hybrid task cascade for instance segmentation, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 4974–4983.
[3] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, 2017, CoRR abs/1709.01507.
[4] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[5] X. Xu, Q. Wu, L. Qi, W. Dou, S.-B. Tsai, Z.A. Bhuiyan, Trust-aware service offloading for video surveillance in edge computing enabled internet of vehicles, IEEE Trans. Intell. Transp. Syst. (2020) 1–10.
[6] X. Xu, Z. Fang, L. Qi, X. Zhang, Q. He, X. Zhou, Tripres: Traffic flow prediction driven resource reservation for multimedia IoV with edge computing, ACM Trans. Multimed. Comput. Commun. Appl. (2020) 1–10.
[7] Z. Zhang, Z. Tang, Y. Wang, Z. Zhang, C. Zhan, Z. Zha, M. Wang, Dense residual network: Enhancing global dense feature flow for character recognition, Neural Netw. 139 (2021) 77–85.
[8] C. Chen, Z. Liu, S. Wan, J. Luan, Q. Pei, Traffic flow prediction based on deep learning in internet of vehicles, IEEE Trans. Intell. Transp. Syst. (2020) 1–14.
[9] Y. Ji, H. Zhang, Z. Zhang, M. Liu, Cnn-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances, Inform. Sci. 546 (2021) 835–857.
[10] S. Wan, Y. Xia, L. Qi, Y.-H. Yang, M. Atiquzzaman, Automated colorization of a grayscale image with seed points propagation, IEEE Trans. Multimed. (2020).
[11] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Trans. Multimedia 20 (11) (2018) 3111–3122.
[12] M. Liao, B. Shi, X. Bai, Textboxes++: A single-shot oriented scene text detector, IEEE Trans. Image Process. 27 (8) (2018) 3676–3690.
[13] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding, X. Ding, Look more than once: An accurate detector for text of arbitrary shapes, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 10552–10561.
[14] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G. Xia, X. Bai, Gliding vertex on the horizontal bounding box for multi-oriented object detection, 2019, CoRR abs/1911.09358.
[15] K. He, G. Gkioxari, P. Dollár, R.B. Girshick, Mask R-CNN, in: Proceedings of International Conference on Computer Vision, 2017, pp. 2980–2988.
[16] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Proceedings of Neural Information Processing Systems, 2015, pp. 91–99.
[17] Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, B. Menze, Knowledge-aided convolutional neural network for small organ segmentation, IEEE J. Biomed. Health Inf. 23 (4) (2019) 1363–1373.
[18] Z. Gao, Y. Li, S. Wan, Exploring deep learning for view-based 3d model retrieval, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 16 (1) (2020) 1–21.
[19] S. Ding, S. Qu, Y. Xi, S. Wan, Stimulus-driven and concept-driven analysis for image caption generation, Neurocomputing (2019).
[20] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[21] M. Liao, G. Pang, J. Huang, T. Hassner, X. Bai, Mask textspotter v3: Segmentation proposal network for robust scene text spotting, in: Proceedings of European Conference on Computer Vision, 2020, pp. 706–722.
[22] D. Deng, H. Liu, X. Li, D. Cai, Pixellink: Detecting scene text via instance segmentation, in: Proceedings of AAAI Conference on Artificial Intelligence, 2018, pp. 6773–6780.
[23] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, Shape robust text detection with progressive scale expansion network, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 9336–9345.
[24] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, G. Li, Scene text detection with supervised pyramid context network, in: Proceedings of AAAI Conference on Artificial Intelligence, 2019, pp. 9038–9045.
[25] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, C. Yao, Textsnake: A flexible representation for detecting text of arbitrary shapes, in: V. Ferrari and M. Hebert and C. Sminchisescu and Y. Weiss (eds). Proceedings of European Conference on Computer Vision, 2018, pp. 19–35.
[26] J. Ye, Z. Chen, J. Liu, B. Du, Textfusenet: Scene text detection with richer fused features, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020, pp. 516–522.
[27] Y. Wang, H. Xie, Z. Zha, M. Xing, Z. Fu, Y. Zhang, Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11750–11759.
[28] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of European Conference on Computer Vision, Vol. 11211, 2018, pp. 3–19.
[29] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of International Conference on Computer Vision Workshops, 2019, pp. 1971–1980.
[30] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in: Proceedings of Computer Vision and Pattern Recognition, 2019, pp. 3085–3094.
[31] S. Liu, D. Huang, Y. Wang, Receptive field block net for accurate and fast object detection, in: Proceedings of European Conference on Computer Vision, Vol. 11215, 2018, pp. 404–419.
[32] Y. Xiao, M. Xue, T. Lu, Y. Wu, S. Palaiahnakote, A text-context-aware cnn network for multi-oriented and multi-language scene text detection, in: Proceedings of International Conference on Document Analysis and Recognition, 2019, pp. 695–700.
[33] W. He, X. Zhang, F. Yin, C. Liu, Multi-oriented and multi-lingual scene text detection with direct regression, IEEE Trans. Image Process. 27 (11) (2018) 5406–5419.
[34] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, East: An efficient and accurate scene text detector, in: Proceedings of Computer Vision and Pattern Recognition, 2017, pp. 2642–2651.
[35] B. Shi, X. Bai, S.J. Belongie, Detecting oriented text in natural images by linking segments, in: Proceedings of Computer Vision and Pattern Recognition, 2017, pp. 3482–3490.
[36] Y. Liu, L. Jin, Deep matching prior network: Toward tighter multi-oriented text detection, in: Proceedings of Computer Vision and Pattern Recognition, 2017, pp. 3454–3461.
[37] P. Lyu, C. Yao, W. Wu, S. Yan, X. Bai, Multi-oriented scene text detection via corner localization and region segmentation, in: Proceedings of Computer Vision and Pattern Recognition, 2018, pp. 7553–7563.
[38] N. Nayef, F. Yin, I. Bizid, et al., ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT, in: Proceedings of International Conference on Document Analysis and Recognition, 2017, pp. 1454–1459.
[39] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, J. Yan, FOTS: Fast oriented text spotting with a unified network, in: Proceedings of Computer Vision and Pattern Recognition, 2018, pp. 5676–5685.