IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 2, 2019, accepted July 12, 2019, date of publication July 18, 2019, date of current version August 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929819

Convolutional Regression Network for Multi-Oriented Text Detection

JUNYU GAO, QI WANG, (Senior Member, IEEE), AND YUAN YUAN, (Senior Member, IEEE)
School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Qi Wang (crabwq@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant U1864204 and Grant 61773316, in part by the State Key Program of National Natural Science Foundation of China under Grant 61632018, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, and in part by the Project of Special Zone for National Defense Science and Technology Innovation.

ABSTRACT Multi-oriented text detection in the wild is a challenging task due to the variations of scales, orientations, illumination, and languages. The traditional anchor mechanism on generic object detection can only generate horizontal proposals, which cannot be applied to detecting multi-oriented text regions. Considering this, in this paper, we propose a novel convolutional regression network (CRN) to localize multi-oriented text in natural images, which consists of two components: region proposal extractor and text locator. To be specific, we first present a hierarchical deconvolution module (HDM), a text-line and geometry segmentation module (TGM) to segment the multi-oriented proposals accurately, both of which are fully convolutional networks. Then, a classification and regression module (CRM) is adopted to process the proposals and obtain the final localization results. The whole framework can be trained in an end-to-end mechanism which is suitable for detecting multi-oriented texts. The extensive experiments are conducted on three mainstream scene-text datasets, and the experimental results evidence the proposed CRN achieves competitive performance.

INDEX TERMS Text detection, object segmentation, fully convolutional network.

I. INTRODUCTION

Reading text from the natural images has attracted much attention in the field of computer vision because of its numerous applications, such as image retrieval [1]–[4], robot navigation [5], [6], video analysis [7]–[9] and scene understanding [10]–[14]. Usually, accurate text localization/detection [15]–[18] is a prerequisite for effectively understanding text. However, it is still extremely challenging to locate text in natural images accurately due to the diversity of text scales, different text fonts/layouts, variation of background and so on. In this paper, we focus on complicated multi-oriented text detection task.

Previous works related to text detection usually contain many sequential steps, including character detection, character classification, text line construction and word splitting. These multi-step approaches are complicated and the error may be accumulated with the increase of steps. Recently, many methods [19], [20] based on generic object detection frameworks are used to detect texts. But these approaches can be only applied to detect horizontal texts. And their miscellaneous components lead to much more error accumulation.

To tackle these problems, we propose a single framework combining segmentation and detection in an end-to-end manner, which is named as Convolutional Regression Network (CRN). It is a multi-task framework that trains the segmentation and detection simultaneously. The entire process is described as: 1) given an input image, segment the text proposal by a fully convolutional network and obtain the outer rectangles of Region of Interests (RoI); 2) project the outer rectangles to the feature maps and extract the corresponding features by rotated ROI pooling; 3) classify and regress the text location. Figure 1 shows the important intermediate results in text localization.

For obtaining text proposals, we develop two modules to handle multi-oriented text, namely Hierarchical Deconvolution Module (HDM), Text-line and Geometry segmentation Module (TGM). HDM utilizes bidirectional LSTM [21] to encode context information and fuses multi-scale feature maps to get more discriminative feature. For segmenting text

The associate editor coordinating the review of this manuscript and approving it for publication was Ran Cheng.

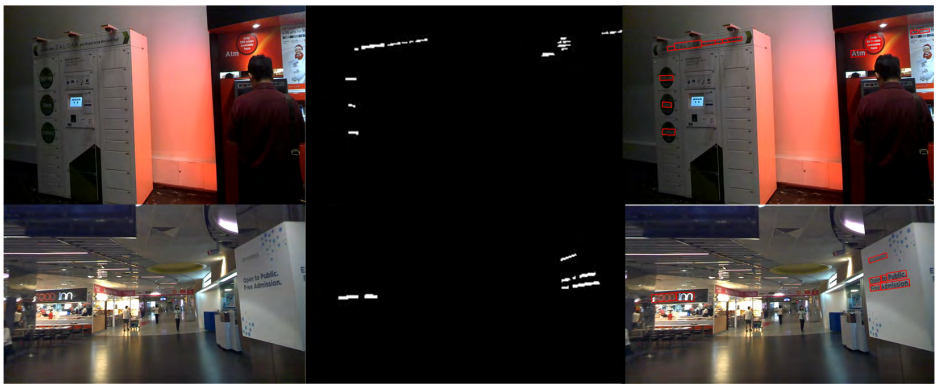


FIGURE 1. From left to right, the pictures represent the original image, segmentation results and its final predict results.

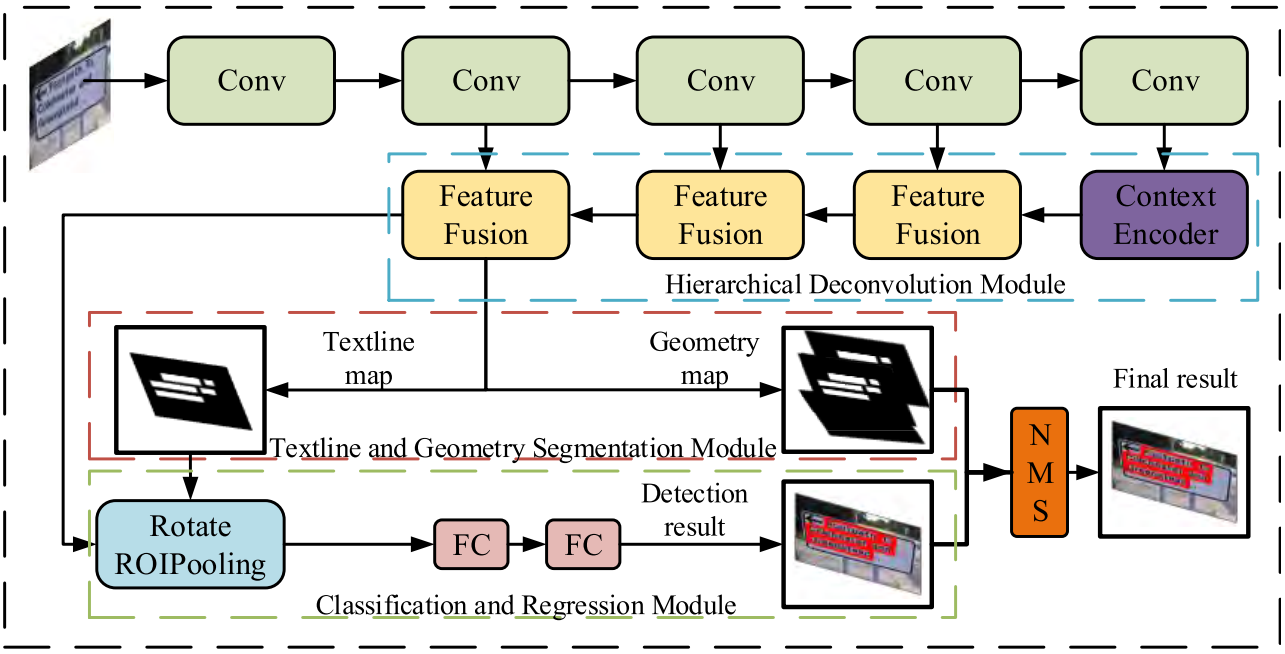


FIGURE 2. The framework of the proposed CRN. CRN consists of four main parts: A VGG-Net Backbone, a Hierarchical Deconvolution Module (HDM), a Text-line and Geometry segmentation Module (TGM), a Classification and Regression Module (CRM). To be specific, HDM encodes context information and fuses multi-scale features to get more discriminative features. TGM consists of two 1×1 convolution layers, generating a text-line map and a geometry map. CRM refines the bounding box of text-line map to get the detection results. The final results are generated by combining the geometry map with the detection results by Non-Maximum-Suppression.

regions accurately, we use not only the region masks that we called text-line map but also the extra five channel masks that we called geometry maps as supervised information of the network. TGM is developed to segment text regions to generate the text-line map and geometry map. The outer rectangles of the connection components in text-line map serve as the proposals of the network. These multi-oriented proposals generated with the segmentation are then projected to the feature map.

To obtain more accurate localization results, we present a Classification and Regression Module (CRM) to refine the proposals from coarse segmentation results. CRM consists of a rotated ROI pooling layer and two fully connected layers. The former can extract the features from multi-oriented regions and the latter output the final detection results. The entire process is illustrated in Figure 2.

In summary, the main contributions of this paper are listed as follows:

1) A novel Hierarchical Deconvolution Module (HDM) is developed which aggregates multi-scale feature maps and context information efficiently.
2) A Text-line and Geometry segmentation Module (TGM) is proposed to segment the text regions as the proposals of the network.
3) A Classification and Regression Module (CRM) is developed to refine the proposals from segmentation results. This is useful and essential for detecting multi-oriented texts.

The rest of our paper is organized as follows. Section II reviews related work about the filed of text detection. Section III describes the proposed approach in detail. Section IV shows the experimental settings and results on three public datasets. Finally, we summarize the work in Section V.

II. RELATED WORK

A. HAND-CRAFTED-FEATURE METHODS

Text detection in the wild has been studied for a few years and many survey results have been released. Methods related to connected component analysis and sliding windows are developed to detect horizontal text. These methods based on connected component analysis often consist of complicated pipelines which could lead to the error accumulation. And they are sensitive to lighting and blurring. Two representative approaches based on connected component analysis are Stroke Width Transform(SWT) [22] and Maximally Stable Extremal Regions(MSERs) [23]. The MSERs algorithm achieves good performance in character detection even on very challenging background, and recently, a number of systems [23], [24] based on MSERs algorithm are developed and achieved high performance. Generally, the connected-component approaches have a great advantage in speed because of low-level feature extraction. However, this approach also generates a large amount of negative samples and some negative samples are hard to be filtered. Different from it, the methods based on sliding windows tend to use a fixed size window to find the most likely text regions. However, the methods based on sliding window often lead to large computational cost. What's more, this method cannot detect text regions completely because of different ratios of text regions. Also, existing approaches based on sliding windows are built on detecting character-level regions, which is unreliable. And it is still difficult to design the discriminative features and train a powerful character classifier. Wu *et al.* [7] propose a novel technique for detecting and tracking video texts of any orientation, which explores gradient directional feature at component level and forms Delaunay triangulation to preserve spatial information.

B. CNN-BASED METHODS

CNN models taking advantage of computing high-level image features have significantly advanced performance on generic object detection in the past few years. Text detection as the subtask of object detection, also benefits from the development of generic object detection [19], [25]–[31]. Region-based Convolutional Neural Network (RCNN) [32], Single Shot Detector(SSD) [33] and segmentation-based Fully Convolutional Network [34] are used to detect texts in natural images due to their great performance and high speed on object detection. Although region proposal based models like R-CNN have already been a state-of-the-art object detector, it cannot be applied to text detection directly without modifications because of the text uniqueness. SSD-based method is efficient and robust in generating text proposals. But it has the shortcomings that it can only generate the horizontal proposals when adopted in text detection. To solve this problem, some researchers develop a rotation proposal method [27] which is able to predict the orientation of text lines. Some modifications of generic object detection framework are developed to detect texts recently. Reference [19] supplements SSD with "textbox layers" to generate bounding-boxes with larger aspect ratios. To combine context information, Tian *et al.* [20] proposed a Connectionist Text Proposal Network(CTPN) to detect text lines with a sequence of text components. However, CTPN becomes futility when meeting multi-oriented texts. Segmentation-based methods [25], [26] attract more and more attention recently because of their pixel-wise classification capacity. This pixel-wise classification is powerful to detect multi-oriented texts. For example, [25] and [26] segment text lines with Fully Convolutional Network(FCN). However, pixel-wise segmentation is not very accurate, which needs other filter strategies to filter noise, and the complicated pipelines in segmentation-based method also result in error accumulations. He *et al.* [30] propose a method for extracting features and locating text line boundaries, which adopts a main-stream fully convolutional network with bi-task predictions.

III. METHODOLOGY

This section describes details of our Convolutional Regression Network(CRN) model, which is an end-to-end framework combining segmentation and detection. Our detector consists of four main parts: a VGG-Net Backbone, a Hierarchical Deconvolution Module (HDM), a Text-line and Geometry segmentation Module (TGM), a Classification and Regression Module (CRM), as shown in Figure 2. The convolutional component is inherited from VGG-Net. To capture more discriminative features, HDM uses top to bottom strategy to aggregate features from different convolutional stages. TGM consists of two 1×1 convolution layers, generating a text-line map and a geometry map. The outer rectangles of the connection component from text-line maps are proposals by our model. Then a rotated ROI pooling layer is developed to process these multi-oriented proposals. Finally, CRN is used to refine the proposes. To be specific, HDM, TGM and CRN are described in Section III-A, III-B and III-C, respectively. In Section III-D, the label generation process is introduced.

A. HIERARCHICAL DECONVOLUTION MODULE

In a CNN model, convolutional features from lower layers contain local image details, while the features from higher layers contain high-level semantic information. To fully utilize different levels of features, we proposed a hierarchical deconvolution module, as shown in Figure 2. Features from the later four stages are aggregated to segment the text regions. The text regions usually contain

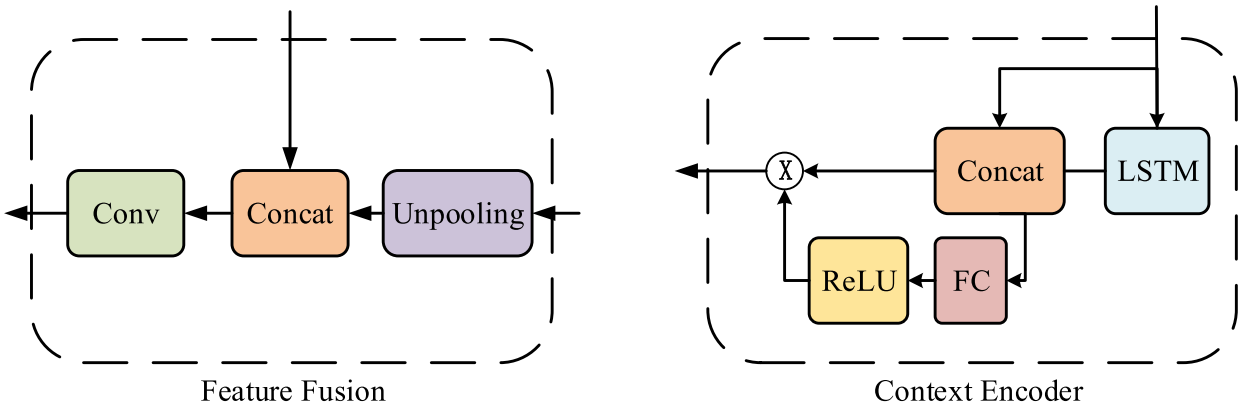


FIGURE 3. The demonstration of Future Fusion (FF) and Context Encoder (CE) in HDM. FF concatenates the features from top to bottom. CE encodes the context information with bidirectional LSTM.

semantic information. To utilize the context information, we develop a context encoder block, as shown in Figure 3. The bidirectional LSTM is used to get the context information. The number of hidden units in bidirectional LSTM is set to 256. So we get 512 extra channels containing context information after LSTM encoder. The 512 extra channels and the original feature channels are concatenated together. Then the fully connected layer is applied to learn the weights of feature channels. These learned weights are used to do channel selections. The context encoder can be presented by the following equation:

$$output = ReLU(fc(concat(LSTM(X), X))) * (concat(LSTM(X), X)). \quad (1)$$

Here, X represents the input feature map. The output of context encoder is then used as the following feature map. The details of hierarchical deconvolution module are shown in Table 1.

B. TEXT-LINE AND GEOMETRY SEGMENTATION MODULE

The text-line and geometry segmentation module consists of two 1×1 convolution layers. The inputs of the two 1×1 convolution layers are both from the outputs of the hierarchical deconvolution module. The first 1×1 convolution layer has only one output channel, which outputs the text-line map. The problem of generating text-line map is that if two text regions are to close, the segmentation could not separate them. To avoid this problem, we prepare the ground truth of text-line maps which are zoomed in half from the bounding box of the texts. The second 1×1 convolution layer has five output channels, representing the geometry map. Geometry map contains five channels, which represent the position information of the pixels in the text-line map. The first four channels represent the distance between the pixel and the four edges of the bounding box. The last channel represents the rotated angle of the bounding box. Thus the geometry map contains the information of rotated box and can output the bounding box of text directly. To segment text-line map, we use class-balanced cross-entropy introduced in [35],

TABLE 1. Details of hierarchical deconvolution module.

block	layer	kernel	channel
context encoder	Bi-LSTM	-	256
feature fusion	Unpooling	-	-
feature fusion	Concat	-	-
feature fusion	Conv	3x3	512
feature fusion	Conv	3x3	512
feature fusion	Conv	3x3	512
feature fusion	Conv	1x1	512
feature fusion	Unpooling	-	-
feature fusion	Concat	-	-
feature fusion	Conv	3x3	256
feature fusion	Conv	3x3	256
feature fusion	Conv	3x3	256
feature fusion	Conv	1x1	256
feature fusion	Unpooling	-	-
feature fusion	Concat	-	-
feature fusion	Conv	3x3	128
feature fusion	Conv	3x3	128
feature fusion	Conv	3x3	128
feature fusion	Conv	1x1	128
feature fusion	Unpooling	-	-
feature fusion	Concat	-	-
feature fusion	Conv	3x3	64
feature fusion	Conv	3x3	64
feature fusion	Conv	3x3	64
feature fusion	Conv	3x3	32
text-line map	Conv	1x1	1
geometry map	Conv	1x1	5

given by

$$\begin{aligned} L_s &= balancedxent(Y_{pred}, Y_{gt}) \\ &= -\gamma * Y_{gt} log Y_{pred} \\ &\quad -(1-\gamma)(1-Y_{gt}) log(1-Y_{pred}), \end{aligned} \quad (2)$$

where L_s represents the loss function of the text-line map, Y_{pred} is the prediction of the text-line map, and Y_{gt} is the ground truth of text-line map. The parameter γ is the factor

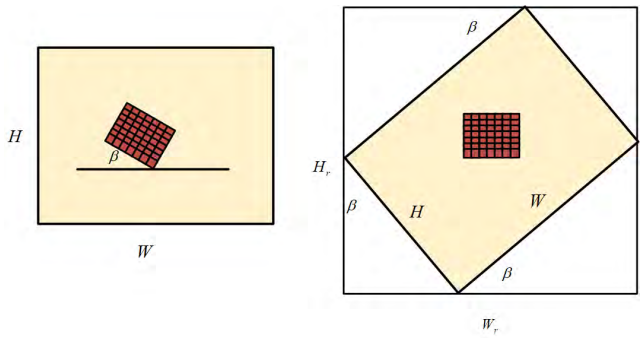


FIGURE 4. The detailed information of Rotated ROI pooling. The red rectangle represents the region of interest. The yellow rectangle is the original image. *H* and *W* represent the height and width of the image. H_r and W_r represent the width and height of image after rotating.

that balance the positive and negative samples. It can be computed by the following equation:

$$\gamma = 1 - \frac{\sum_{y_{gt} \in Y_{gt}} y_{gt}}{|Y_{gt}|}. \tag{3}$$

As for geometry map, we use the following loss function:

$$L_{geo} = -logIOU(R_{gt}, R_{pred}) + \beta(1 - cos(\theta_{gt} - \theta_{pred})), \tag{4}$$

where θ represents the rotated angles of the texts. R represents the region of texts. $IOU(R_{gt}, R_{pred})$ is formulated as:

$$IOU(R_{gt}, R_{pred}) = \frac{R_{gt} \cap R_{pred}}{R_{gt} \cup R_{pred}}. \tag{5}$$

The width and height of the intersected rectangle $|R_{gt} \cap R_{pred}|$ can be computed by the following equation:

$$w = min(d_2^{gt}, d_2^{pred}) + min(d_4^{gt}, d_4^{pred}), \tag{6}$$
$$h = min(d_1^{gt}, d_1^{pred}) + min(d_3^{gt}, d_3^{pred}), \tag{7}$$

where d_1, d_2, d_3 and d_4 represents the distance from the position of a pixel to the top, right, bottom and left boundary of bounding box respectively. The union area $|R_{gt} \cup R_{pred}|$ is define as below:

$$R_{gt} \cup R_{pred} = R_{gt} + R_{pred} - R_{gt} \cap R_{pred}. \tag{8}$$

C. CLASSIFICATION AND REGRESSION MODULE

The classification and regression module consists of a rotated ROI pooling layer and two fully connection layers. The traditional ROI pooling can only process the horizontal proposals while the proposals from the segmentation results could be in arbitrary orientation. To extract the features from multi-oriented regions, we propose a new ROI pooling layer called Rotated ROI pooling layer. The rotated angles come from the main axis of the outer rectangles of the connection component in text-line map. So each proposal has its own rotated angle. The feature channel gets rotated according to the angle, which is shown in Figure 4. The center of rotating is the center of the feature map. The transform matrix can be presented by the following equation:

$$TransformMatrix = \begin{bmatrix} cos(\alpha) & -sin(\alpha) & \theta_w \\ sin(\alpha) & cos(\alpha) & \theta_h \end{bmatrix}, \tag{9}$$

where θ_w, θ_h are the translation parameters. The values of this two parameters can be computed by the following equations:

$$\theta_w = \frac{(W_r - W)}{2}, \tag{10}$$
$$\theta_h = \frac{(H_r - H)}{2}, \tag{11}$$
$$H_r = W * sin(\alpha) + H * cos(\alpha), \tag{12}$$
$$W_r = W * cos(\alpha) + H * sin(\alpha), \tag{13}$$

where W_r,H_r are the width and height of the feature map after rotating, W,H are the width and height of the original feature map and α is the rotated angle. The proposals can be transformed to the horizontal orientation through the transform matrix (Eq.9).

The loss function of CRN consists of four parts which is given by:

$$Loss = \lambda L_{seg} + L_{geo} + L_{cls} + L_{reg}, \tag{14}$$

where L_{seg} is the loss function of the text-line map which has the same format of the equation(2). L_{geo} is the loss function of the geometry map with the format of the equation(4). L_{cls} represents the loss function of the classification in the classification and regression module. The format of L_{cls} can be represented by the following equation:

$$L_{cls}(p_{gt}, p_{pred}) = -log(p_{gt}p_{pred} + (1 - p_{gt})(1 - p_{pred})), \tag{15}$$

where p_{gt} is the true label of the proposals, p_{pred} is the predicted label of the proposals. Lastly, L_{reg} represents the loss function of the regression in the classification and regression module. The format of L_{reg} can be computed by:

$$L_{reg}(t_i, t_i^*) = Smooth_{L1}(t_i - t_i^*), \tag{16}$$

where t_i represents the predicted coordinate of the text, while t_i^* is the target coordinate of text regions.

D. LABEL GENERATION

Inspired by [36] and [37], we generate the masks to train the segmentation network. We consider that the case where the geometry is a quadrangle. For a quadrangle $Q = \{p_i | i \in \{1, 2, 3, 4\}\}$, where $p_i = \{x_i, y_i\}$ are vertices of the quadrangle in clockwise order. To shrink Q, we first compute a reference length r_i for each vertex p_i as

$$r_i = min(D(p_i, p_{(i\%4)+1}), D(p_i, p_{((i+2)\%4)+1})), \tag{17}$$

where $D(p_i, p_j)$ denotes the L_2 distance between p_i and p_j. We first shrink the two longer edges of a quadrangle and then the two shorter edges. For each edge, we shrink it by moving its two endpoints inward along the edge by $0.25r_i$. The label generation of the text-line map is illustrated in Figure 5.

FIGURE 5. From top to bottom, the pictures represent the original image, generated text line mask and shrunk text-line mask.

The generation process for the geometry map is illustrated as the following. We first generate a rotated rectangle that covers the text region with minimal area. For each pixels with positive score from text-line map, we calculate its distances to the four boundaries of the text box and put them into the first four channels. The rotated angle of the text is put into the last channel of the geometry map.

IV. EXPERIMENTAL RESULTS

A. DATASETS

To compare our method with existing methods, we conducted experiments on three public benchmarks: ICDAR2013 [38], ICDAR2015 [39] and MSRA-TD500 [40].

ICDAR2013 includes 229 natural images for training and 233 images for testing which are labeled in word level. It is worth noting that the texts in ICDAR2013 are horizontal.

ICDAR2015 has 1,000 training images and 500 testing images which are annotated with four vertices. We generate inclined rectangle from the quadrangle annotated in ICDAR2015.

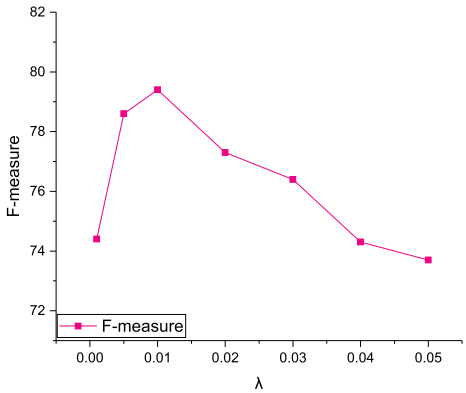
MSRA-TD500 dataset includes 300 training images and 200 testing images, which is annotated with text line level which is different from the annotation of ICDAR2015. To be specific, the annotations of images consist of both position and orientation of text instance.

B. IMPLEMENTATION DETAILS

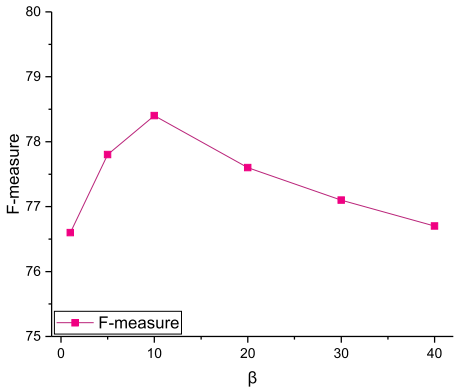
For the text-line map labels, we fully follow the settings of [36] to generate them.

The training process of our method consists of two stages. The first stage is to train the segmentation process including generation of text-line map and geometry map. The second stage is to train all the process including segmentation and detection. During the first stage, we uniformly sample 512×512 crops from images to form a minibatch of size 16. Learning rate of ADAM [41] starts from 1e-3, and decays to one-tenth every 20,000 iterations. On the second stage, learning rate of ADAM is fixed as le-4. In the ADAM optimization, the weight decay for L2 penalty is set as 1e-4.

The training and evaluation are performed on NVIDIA GTX 1080Ti GPU using TensorFlow framework.



(a) Line chart of performance under different lambda in Eq. 14.



(b) Line chart of performance under different beta in Eq. 4.

FIGURE 6. The performance under the different loss weights on ICDAR2015 dataset.

C. ABLATION STUDY ON ICDAR2015 DATASET

In this section, we take ICDAR2015 dataset as example to further analyze and discuss our proposed CRN.

1) PARAMETER ANALYSIS FOR λ AND β

Here, we analyze the effects of λ in Eq. 14 and β in Eq. 4 for detection results. We train CRN on the joint of ICDAR2013 and ICDAR2015 training data and evaluate performance on ICDAR2015 test data.

As for the λ, it is set as 0.001, 0.005, 0.01, 0.02, 0.03, 0.04 and 0.05. Figure 6a illustrates the F-measure values under different settings. From it, we find that the F-measure is the best when λ is set as 0.01.

For the setting of β, we conduct the experiments under 1, 5, 10, 20, 30, 40 and 50. The F-measure values are

TABLE 2. The comparison results of different combinations of each module in HDM.

Methods	CE	FF	F-measure %
NoHDM			76.2
onlyCE	✔		77.2
onlyFF		✔	78.6
HDM	✔	✔	79.4

TABLE 3. Runtime comparison of mainstream methods and ours.

method	Backbone	Runtime (ms)
Yin et al. [43]	-	1400
Yin et al. [24]	-	800
Wu et al. [7]	-	50
Zhang et al. [26]	VGG-16	2100
TD-ICDAR [40]	-	7200
EAST [36]	VGG-16	150
Mask Textspotter [44]	Res-50	208
PixelLink [42]	VGG-16	333
CRN(ours)	VGG-16	**80**

reported in Figure 6b. When β is set as 10, the performance achieves the best of 78.4% F-measure.

2) THE EFFECT OF HDM

In Section III-A, the HDM is proposed that consists of Context Encoder (CE) and Feature Fusion (FF). Here, we analyze the effects of HDM's modules. Table 2 lists the results of different module combinations on ICDAR2015 Dataset. From it, we find the improvement of FF is better than CE, which shows that the multi-layer features play a more important role than contextual information in a Fully Convolutional Network. When CE and FF are introduced together, the best F-measure (79.4%) is obtained.

3) COMPARISON OF COMPUTATION SPEED

Here, we compare the inference time with some mainstream methods in Table 3. From it, we find that the proposed method takes the least amount of time (80ms per image). Compared with Zhang *et al.* [26], EAST [36] and PixelLink [42] that adopt the same backbone, the runtime of our method outperforms that of them.

D. PERFORMANCE ON ICDAR2015 DATASET

Table 4 shows experimental results on ICDAR2015. According to it, we can see that our method achieves 79.4% of F-measure and 78.4% of recall, which are the best in all compared CNN-based methods. At the same time, 80.4% of precision is the second place, which is slightly less than the result of EAST [36] (80.5% of precision). Compared with hand-crafted features, [7] only need 50ms to detect texts. However, its performance on ICDAR2013 and MSRA-TD500 is less than method. In general, our proposed CRN is the best compared with the current mainstream methods.

TABLE 4. Performance on ICDAR2015 dataset. The red, blue and green fonts represent the first, second and third place in the corresponding column, respectively.

method	Recall%	Precision%	F-measure%
AJOU [45]	46.9	47.3	47.1
NJU [24]	36.3	70.4	47.9
Yin et al. [43]	32.1	49.6	39.0
StradVison2 [39]	36.7	77.5	49.8
CNN MSER [39]	34.4	34.7	34.6
Yao et al. [25]	58.7	72.3	64.8
CTPN [20]	51.2	74.2	60.9
Zhang et al. [26]	43.1	70.1	53.6
WordSup [46]	77.0	79.3	78.2
SSTD [47]	73.0	80.0	77.0
EAST [36]	72.8	80.5	76.4
SegLink [48]	73.1	76.8	75.0
CRN(ours)	78.4	80.4	79.4

In order to show the performance of the proposed method more intuitively, some typical visualization exemplars are shown in Figure 7. The red boxes are the detection results of our CRN. From it, we find our model can use rotated rectangles to effectively localize the multi-oriented text in complicated scenes.

E. PERFORMANCE ON ICDAR2013 DATASET

Table 5 reports the detailed metrics (Recall, Precision and F-measure) of the proposed CRN and other mainstream methods on ICDAR2013 dataset. From it, CRN achieves the second place on three metrics. At the same time, we also find CTPN [20] is the best in terms of the three metrics. It is a method elaborately designed for horizontal text detection. Thus, it is able to perform best than other methods. Even so, CRN obtains a competitive results, which are very close to CTPN's performance. In the leader board on ICDAR2015 dataset (see also Table 4), CRN's results is far better than that of CTPN.

F. THE PERFORMANCE ON MSRA-TD500

Here, we evaluate our method on the MSRA-TD500 dataset, which consists of multi-lingual text. Considering that MSRA-TD500 dataset only includes 300 training images and 200 testing images, we apply the following strategy to collect enough training data. Given an image, it is resized using four random scale ratio between 0.25 to 5. As a result, we obtain 1,500 training images (five times the original amount). Then the resized image is randomly cropped into patches with different sizes. As illustrated in Table 6, our method achieves 63% of recall, 84% of precision and 72% of F-measure, which are the best in the corresponding column. Nevertheless, we find that the performance on MSRA-TD500 is poorer than results on ICDAR2015/2013 dataset. The main reason is that MSRA-TD500's training images is too few for a deep CNN. In addition, MSRA-TD500 contains English and Chinese texts simultaneously, so it is a challenging dataset.

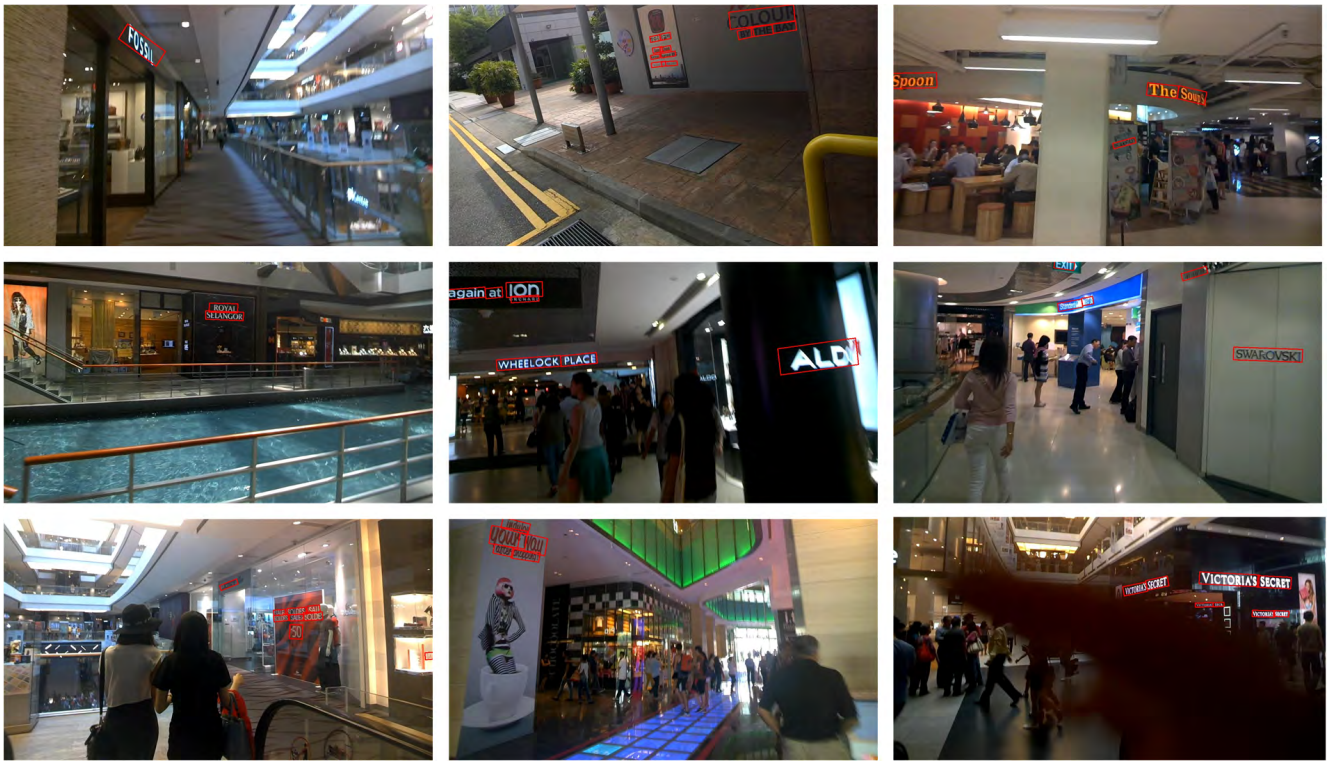


FIGURE 7. Visualization results of multi-oriented text detection on ICDAR2015 dataset.

TABLE 5. Experimental results on ICDAR2013. The meanings of red, blue and green fonts are the same as Table 4.

method	Recall%	Precision%	F-measure%
Shi et al. [49]	63	83	72
SFT-TCD [50]	75	82	73
MSERs-CNN [51]	71	88	78
Yin et al. [43]	68	86	76
Fastext [52]	69	84	77
TextFlow [53]	76	86	81
Wu et al. [7]	76	70	73
CTPN [20]	83	93	88
Yao et al. [25]	80	89	84
Zhang et al. [26]	78	88	83
Seglink [48]	83	88	85
TextBoxes [19]	74	88	81
CRN(ours)	80	90	85

We also follow EAST [36] to exploit extra data [53] to train a model. Finally, we attain a better performance than the original result.

Experimental results on the ICDAR2015 dataset and MSRA-TD500 dataset demonstrate that our method has advantages on multi-oriented text detection. Using segmentation to get proposals overcomes the difficulties that the methods based on anchor mechanism encounter. In summary, our text detector achieves the state-of-the-art results especially for multi-oriented text data.

TABLE 6. Experimental results on MSRA-TD500.

method	Recall%	Precision%	F-measure%
TD-Mixture [40]	63	63	60
TD-ICDAR [40]	52	53	50
Yin et al. [43]	61	71	66
Kang et al. [54]	62	71	66
Yin et al. [24]	63	81	71
Wu et al. [7]	63	70	66
EAST [36]	61	82	70
CRN(ours)	63	84	72
CRN(ours) + extra data	**67**	**86**	**75**

V. CONCLUSION

We propose a new text detection framework combining segmentation and detection. To detect multi-oriented texts, we replace anchor mechanism with segmentation to generate text proposals. Using the segmentation methods to produce text proposals overcomes the multi-oriented layout problems while the anchor mechanism only generates horizontal proposals. The hierarchical deconvolution module in our framework concatenates context information and different levels of feature maps from top to bottom, which is useful for segmentation task. What's more, a novel classification and regression module is developed to refine the bounding box from the text-line map. The experimental results demonstrate that our method achieves good performance on multi-oriented text detection.

REFERENCES

[1] A. Raza, H. Dawood, H. Dawood, S. Shabbir, R. Mehboob, and A. Banjar, "Correlated primary visual texton histogram features for content base image retrieval," *IEEE Access*, vol. 6, pp. 46595–46616, 2018.

[2] Q. Qi, Q. Huo, J. Wang, H. Sun, Y. Cao, and J. Liao, "Personalized sketch-based image retrieval by convolutional neural network and deep transfer learning," *IEEE Access*, vol. 7, pp. 16537–16549, 2019.

[3] Y. Song, J. Lei, B. Peng, K. Zheng, B. Yang, and Y. Jia, "Edge-guided cross-domain learning with shape regression for sketch-based image retrieval," *IEEE Access*, vol. 7, pp. 32393–32399, 2019.

[4] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1457–1470, May 2017.

[5] C. S. Karagöz, H. I. Bozma, and D. E. Koditschek, "Coordinated navigation of multiple independent disk-shaped robots," *IEEE Trans. Robot.*, vol. 30, no. 6, pp. 1289–1304, Dec. 2014.

[6] Z. Wu, J. Li, J. Zuo, and S. Li, "Path planning of UAVs based on collision probability and Kalman filter," *IEEE Access*, vol. 6, pp. 34237–34245, 2018.

[7] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1137–1152, Aug. 2015.

[8] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017.

[9] B. Zhao, X. Li, and X. Lu, "HSA-RNN: Hierarchical structure-adaptive RNN for video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7405–7414.

[10] B. Zhao, X. Li, X. Lu, and Z. Wang, "A CNN–RNN architecture for multi-label weather recognition," *Neurocomputing*, vol. 322, pp. 47–57, Dec. 2018.

[11] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.

[12] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[13] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[14] B. Zhao, X. Li, and X. Lu, "CAM-RNN: Co-attention model based RNN for video captioning," *IEEE Trans. Image Process.*, to be published.

[15] X. Ren, Y. Zhou, Z. Huang, J. Sun, X. Yang, and K. Chen, "A novel text structure feature extractor for Chinese scene text detection and recognition," *IEEE Access*, vol. 5, pp. 3193–3204, 2017.

[16] P. Yang, F. Zhang, and G. Yang, "A fast scene text detector using knowledge distillation," *IEEE Access*, vol. 7, pp. 22588–22598, 2019.

[17] W.-Y. Pei, C. Yang, L.-Y. Meng, J. Hou, S. Tian, and X.-C. Yin, "Scene video text tracking with graph matching," *IEEE Access*, vol. 6, pp. 19419–19426, 2018.

[18] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, Jul. 2019.

[19] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI*, 2017, pp. 4161–4167.

[20] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 56–72.

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

[22] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2963–2970.

[23] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[24] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1930–1937, Sep. 2015.

[25] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," 2016, *arXiv:1606.09002*. [Online]. Available: https://arxiv.org/abs/1606.09002

[26] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," 2016, *arXiv:1604.04018*. [Online]. Available: https://arxiv.org/abs/1604.04018

[27] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," 2017, *arXiv:1703.01086*. [Online]. Available: https://arxiv.org/abs/1703.01086

[28] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: https://arxiv.org/abs/1706.09579

[29] Q. Yang, M. Cheng, W. Zhou, Y. Chen, M. Qiu, W. Lin, and W. Chu, "IncepText: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection," 2018, *arXiv:1805.01167*. [Online]. Available: https://arxiv.org/abs/1805.01167

[30] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5406–5419, Nov. 2018.

[31] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 19–35.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[35] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1395–1403.

[36] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," 2017, *arXiv:1704.03155*. [Online]. Available: https://arxiv.org/abs/1704.03155

[37] J. Gao, Q. Wang, and X. Li, "PCC net: Perspective crowd counting via spatial convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[38] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.

[39] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[40] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1083–1090.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[42] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6773–6780.

[43] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.

[44] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 67–83.

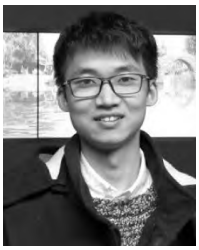
[45] H. I. Koo and D. H. Kim, "Scene text detection via connected component clustering and nontext filtering," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2296–2305, Jun. 2013.

[46] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "WordSup: Exploiting word annotations for character based text detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4950–4959.

[47] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3066–3074.

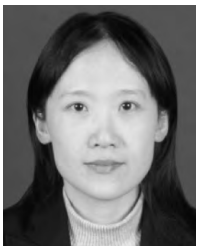
[48] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. CVPR*, vol. 3, Jul. 2017, pp. 3482–3490.

[49] C. Shi, C. Wang, B. Xiao, Y. Zhang, and S. Gao, "Scene text detection using graph model built upon maximally stable extremal regions," *Pattern Recognit. Lett.*, vol. 34, no. 2, pp. 107–116, 2013.
[50] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1241–1248.
[51] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 497–511.
[52] M. Buta, L. Neumann, and J. Matas, "FASText: Efficient unconstrained scene text detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1206–1214.
[53] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4651–4659.
[54] L. Kang, Y. Li, and D. Doermann, "Orientation robust text line detection in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4034–4041.



JUNYU GAO received the B.E. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2015, where he is currently pursuing the Ph.D. degree with the Center for Optical Imagery Analysis and Learning. His research interests include computer vision and pattern recognition.

QI WANG (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, the Unmanned System Research Institute, and the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



YUAN YUAN (M'05–SM'09) is currently a Full Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and also conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.