SCALE-INVARIANT MULTI-ORIENTED TEXT DETECTION IN WILD SCENE IMAGES

Kinjal Dasgupta[1], Sudip Das[2*], Ujjwal Bhattacharya[2]

[1]Heritage Institute of Technology, Kolkata   [2]Indian Statistical Institute, Kolkata

kinjal.dasgupta.9@gmail.com, d.sudip47@gmail.com, ujjwal@isical.ac.in

ABSTRACT

Automatic detection of scene texts in the wild is a challenging problem, particularly due to the difficulties in handling (i) occlusions of varying percentages, (ii) widely different scales and orientations, (iii) severe degradations in the image quality etc. In this article, we propose a fully convolutional neural network architecture consisting of a novel *Feature Representation Block* (FRB) capable of efficient abstraction of information. The proposed network has been trained using curriculum learning with respect to difficulties in image samples and gradual pixel-wise blurring. It is capable of detecting texts of different scales and orientations suffered by blurring from multiple possible sources, non-uniform illumination as well as partial occlusions of varying percentages. Text detection performance of the proposed framework on various benchmark sample databases including *ICDAR 2015*, *ICDAR 2017 MLT*, *COCO-Text* and *MSRA-TD500* improves respective state-of-the-art results significantly. Source code of the proposed architecture will be made available at github.

Index Terms— Scene-Text Detection, Deep Features Representation Learning.

1. INTRODUCTION

Although existing state-of-the-art scene text detection models have already acquired promising performances on moderately well-behaved scene image samples, till date, efficient detection of incidental scene texts (such as texts captured by wearable cameras where the capture is difficult to control) remains one of the most challenging tasks in the Computer Vision community. In 2015, the ICDAR Robust Reading Competition [1] introduced a few challenges of processing incidental scene texts and published a related sample database for the first time. COCO-Text [2], another benchmark dataset created at a later period also contains similar samples of complex everyday scenes. Such samples are often motion blurred, non-uniformly illuminated, partially occluded, multi-oriented or multi-scaled. Efficient detection of similar texts in scene image needs further intensive studies. Robustness of the detection procedure largely depends on the distinguishing power

Fig. 1: Red rectangular boxes show widely varying scene texts detected by the proposed framework.

of the feature representation between text and non-text components. Since supremacy of deep learning-based strategies for feature representation of raw images over various traditional hand-crafted filters has already been established [3], our present study is centred around the development of an efficient deep convolutional neural network architecture for the present purpose. The deep architecture developed in this study includes a novel *Feature Representation Block* (FRB) for robust detection of scene texts. This FRB has been suitably designed towards abstraction of information at multiple levels generating features adaptive to changes in orientation and scale. Its Gabor Filter based Convolutional component captures scale and orientation information, the channel-wise attention map enhances important features and attenuates noisy and unwanted background information, the 4Dir IRNN component [4] takes care of contextual information by moving the RNNs laterally across the image while the *conditional random fields* (CRFs) [5] based aggregation component refines inbound information of multiple scales. Figure. 1 shows outputs of the proposed detection framework on two samples of Robust Reading Competition datasets [1, 6] containing incidental scene texts.

2. RELATED WORKS

Detection and recognition of texts in scene images have been studied for a long period. A recent survey of existing approaches and analysis of results can be found in [7]. Among the traditional approaches of scene text detection, stroke width transform was used in [8], stroke feature transform was studied in [9], MSERs had been used in [10] while the well-known FAST corner detector based text fragment detection was reported in [11]. Recent studies of text detection

---

use deep learning based strategies where features are learned automatically with the help convolutional neural networks (CNN). Certain Edge Boxes region proposal algorithm [12] and an aggregate channel features detector [13] had been combined in [14] for detecting words in scene images. This approach made use of the R-CNN object detection framework [15]. The objective of all these studies was detection of texts in natural scene images where the cameraman controls the camera. However, if the capturing device is a wearable one, detection of incidental texts appearing in such images captured without any control becomes more difficult. In the 2015 version of ICDAR Robust Reading Competition [1], a challenge on Incidental Scene Text was introduced, based on a dataset of 1,670 sample images captured using the Google Glass. A similar dataset, called COCO-Text [2], was later introduced in the year 2016. In the next year, Zhou *et al.* [16] proposed a fully convolutional network architecture and a non-maximum suppression strategy which can directly locate a word or a text line of arbitrary orientations in similar scene images. He *et al.* [17] proposed a direct regression strategy to detect incidental texts of multiple orientations having variations in size and perspective distortions. Later Liu *et al.* [18] combined low-level and high-level feature maps produced by their CNN architecture which could detect incidental texts of different orientations. In [19], a pyramid context network had been proposed for localization of text regions in natural scene images. Wang *et al.* [20] used a region proposal network together with a RNN for detection of scene text regions of arbitrary shapes. Mask R-CNN has recently been used in [21] for detection of scene texts of arbitrary shapes.

3. PROPOSED METHODOLOGY

The proposed methodology implements a sophisticated neural network architecture capable of detecting instant scene texts. The network architecture is shown in Figure. 2. It is a fully-convolutional neural network consisting of a block for extension of input image in 4 orientation channels as in [3] followed by a ResNet-18 [22] based feature extraction block modulated by Gabor Orientation Filter (GOF), an FRB containing multiple convolutional Gabor Orientation filters and a *Mult-scale Feature Refinement Module* (MFRM), a decoder block of a series of deconvolution operations and finally a detection branch similar to the one used in [16].

3.1. Feature Representation Block (FRB)

The *Feature Representation Block* (FRB) receives the output of the $Conv_{3\_2}$ layer of backbone Resnet-18 architecture modulated by Gabor Orientation Filter used for enhancement of the robustness of traditional convolution filters towards various image transformations such as scale variations and rotations. The FRB consists of multiple Gabor Convolutional layers arranged in multiple rows and a *Multi-scale Feature*

Refinement Module (MRFM) to make the learned feature representation more abstract. Feature maps computed at a higher layer of the FRB is downsampled before entering into the Gabor Convolutional layers of the next row. We also considered kernels of different sizes in each row to tackle the enormous difference in the scales of scene texts. A larger kernel helps to capture more information of Gabor Orientation while a smaller kernel is capable to capture more locally distributed information.

3.2. Multi-scale Feature Refinement Module (MFRM)

MFRM consists of a channel-wise attention map, a 4Dir-IRNN component and an aggregation component modulated by CRFs. Outputs $(N \times U \times H \times W)$ (where N, and U denote respectively the numbers of convolution and orientation channels, H and W indicates the height and width of input image) of the last three Gabor Convolutional layers one in each row of FRB are fed as input to the MFRM. We have chosen $N = 512$ and $U = 4$. Note that for every channel k, the network produces rank 3 matrix Γ_k: $U \times H \times W$ (where $k \in [1, 2, \dots 512]$). We concatenate $\Gamma_k, \forall k$ such that the final output matrix will be $N_U \times H \times W$, where $N_U = NU$. Thereafter it has been moved through a $Conv_{1\times1}$ reduce the number of channels and fed into the distinct position for every feature. The channel-wise attention module that assists to different channels of the feature map to enhance the essential knowledge, while attenuating noise and background data before transferring it to task-definite layers of the model. We first pass the feature map $\mathcal{F}_{Conv3\times3}$ originating from the last convolution block into a $Conv_{1\times1}$ to produce a $\mathcal{F}_{1\times1}$ of same size and number of channels. $\mathcal{F}_{1\times1}$ is then passed through sigmoidal activation function and multiplied with $\mathcal{F}_{Conv3\times3}$ component-wise to produce the feature map $\mathcal{F}_{att}$. Mathematically,

$$\mathcal{F}_{att} = \mathcal{F}_{Conv3\times3} \otimes \sigma(Conv_{1\times1}(\mathcal{F}_{Conv3\times3})) \qquad (1)$$

We follow the 4Dir-IRNN component [4] in this framework for computing the contextual features. The 4Dir-IRNN component is a sort of RNN composed with ReLUs. The total number of channels in coming from FRB distributed by a factor of 4 and for each 128 channel, we implement a $Conv_{1\times1}$ and fed toward the 4Dir-IRNN block. Inside the IRNN we have four RNNs which entirety moves laterally across the image in the 4 directions: up, down, left, right. The input-to-hidden is analysed by $Conv_{1\times1}$ and shared the convolution layer with 4 recurrent transitions. A distinct direction of RNN will move to collect the contextual features from every 128 channels. The hidden-to-output is merged into a single convolution layer, concatenation followed by a $Conv_{1\times1}$ along with a ReLU activation function to generate $\mathcal{F}_{IRNN}$. The aggregation block is a multiple features merging unit reinforced with CRFs. Different features transferred from FRB, $\mathcal{F}_{att}$ and $\mathcal{F}_{IRNN}$ into the aggregation unit. The CRFs [5] present
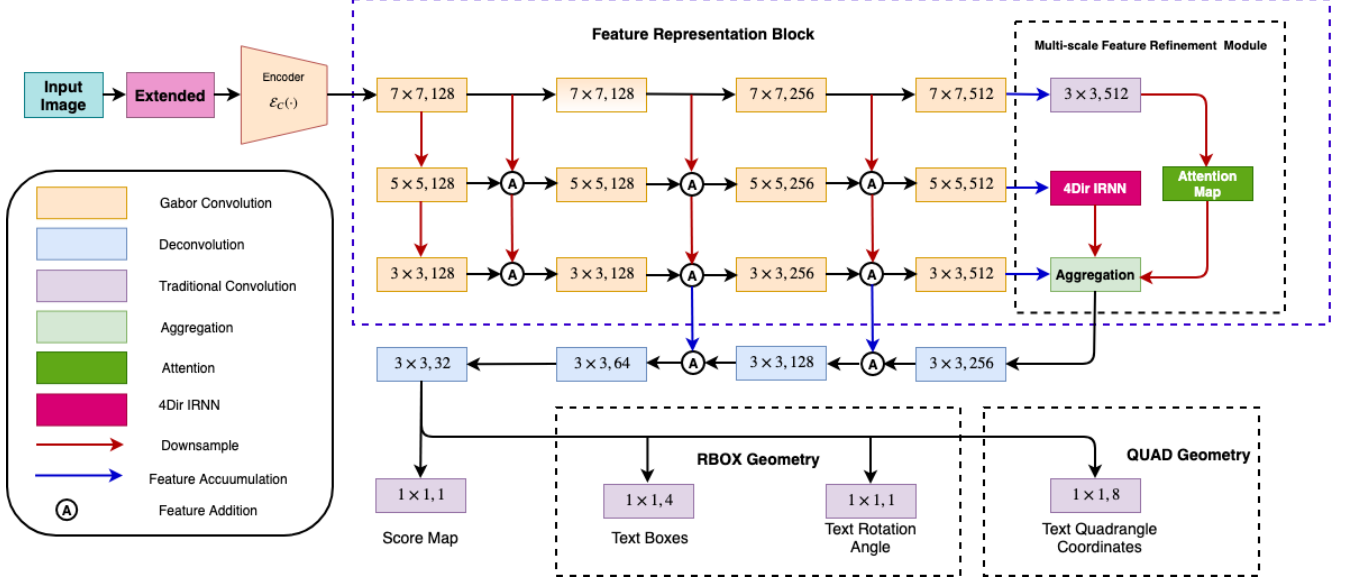
Feature Representation Block

Multi-scale Feature Refinement Module

| Input Image | → | Extended | → | Encoder $\mathcal{E}_c(\cdot)$ |

$7 \times 7, 128$ → $7 \times 7, 128$ → $7 \times 7, 256$ → $7 \times 7, 512$ → $3 \times 3, 512$

$5 \times 5, 128$ → $5 \times 5, 128$ → $5 \times 5, 256$ → $5 \times 5, 512$ → 4Dir IRNN → Attention Map

$3 \times 3, 128$ → $3 \times 3, 128$ → $3 \times 3, 256$ → $3 \times 3, 512$ → Aggregation

$3 \times 3, 32$ ← $3 \times 3, 64$ ← $3 \times 3, 128$ ← $3 \times 3, 256$

Legend:
- Gabor Convolution
- Deconvolution
- Traditional Convolution
- Aggregation
- Attention
- 4Dir IRNN
- Downsample
- Feature Accumulation
- Feature Addition

$1 \times 1, 1$ — Score Map

RBOX Geometry

$1 \times 1, 4$ — Text Boxes

$1 \times 1, 1$ — Text Rotation Angle

QUAD Geometry

$1 \times 1, 8$ — Text Quadrangle Coordinates

Fig. 2: Schematic diagram of the proposed architecture: it contains (i) a ResNet-18 used as feature encoder $\mathcal{E}_c(\cdot)$, (ii) a Feature Representation Block (FRB), (iii) a Deconvolution Block and (iv) a Detection Branch. The FRB has a Multi-scale Feature Refinement Module (MFRM) consisting of a 4Dir IRNN, a channel-wise attention block and a CRFs-based aggregation block.

in the aggregation component helps to refine each incoming multi-scale feature by combining the corresponding knowledge received from other features. We simply conserve those $(N \times H \times W)$ learned features from the aggregation block and shared into the decoder via $\mathcal{C}onv_{1\times1}$.

3.3. Feature Learning

Our primary goal is to make a model which is invariant to learn different deformation. Instead of relying on the large diverse dataset, we consider a different approach *Mask and Predict* [23] along with pixel-wise blurring strategy for training to learn diverse deformation. The strategy in our proposed network endeavours to explain the difficulty of text detection in cases of occlusion, motion blur and different lighting conditions of the input image. In this approach, we gradually increase a certain percentage of pixel-wise blurring in the input image. Also, we use curriculum learning [24] by manually arranging images from simpler to difficult ones.

3.4. Feature Decoder

The feature map containing rich information ejected from the aggregation block is fed into a $\mathcal{D}econv_{3\times3}$. In each feature merging point Figure. 2, the incoming feature map from FRB is operated in the same way as described in section 3.2 for its addition with the current feature map. After the two such successive feature additions followed by $\mathcal{D}econv_{3\times3}$ of 256 and 128 channels, two more $\mathcal{D}econv_{3\times3}$ of 64 and 32 channels are applied before feeding it to the detection branch.

3.5. Detection

The design of the detection branch is similar to the output layer of the network of [16]. It consists of several $\mathcal{C}onv_{1\times1}$ operations of 1, 4, 1, 8 channels of feature maps where the first one provides score map and the following two provide RBOX geometry while the last one provide QUAD geometry. The Loss Function used for the detection branch is as follows,

$$\mathcal{L} = \mathcal{L}_s + \lambda_g \mathcal{L}_g, \qquad (2)$$

Where $\mathcal{L}_s$ is Loss of the Score Map, $\mathcal{L}_g$ is Loss of the Geometry and λ_g consider the significance between two losses.

4. EXPERIMENTAL DETAILS

For comparison purpose, the proposed text detection framework has been trained on four publicly available benchmark datasets: *ICDAR2015* [1], *ICDAR2017* [6], *COCO-Text* [2] and *MSRA-TD500* [25]. Numbers of training samples of these four datasets are respectively 1000, 7200, 43686 and 300 while the volumes of the respective test sets are 500, 9000, 10000 and 200.

4.1. Training Details

We employ Curriculum Learning [24] with respect to progressively increasing pixel-wise blurring, masking percentage and complexity of images to enable the feature learning invariant to image degradation, occlusion and image complexities. We

Fig. 3: Qualitative detection results of the proposed framework on the samples which have texts of widely varying scales, multiple orientations, multiple scripts, motion blurred, non-uniformly illuminated, partially occluded etc.

Table 1: Comparative detection results of the proposed framework and different SOTA models on various benchmark sample databases.

Methods	ICDAR 2015			ICDAR 2017 MLT			MSRA-TD500			COCO-Text		
	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score
EAST	78.3	83.3	80.7	-	-	-	67.4	87.3	76.0	32.4	50.3	39.4
He *et al.*	80.0	82.0	80.9	-	-	-	-	-	-	-	-	-
FOTS	87.9	**91.8**	89.8	62.3	81.8	70.7	-	-	-	-	-	-
Xie *et al.*	85.8	88.7	87.2	68.6	80.6	74.1	-	-	-	-	-	-
Wang *et al.*	86.0	89.2	87.6	-	-	-	**82.1**	85.2	83.6	-	-	-
Qin *et al.*	87.9	91.6	89.7	-	-	-	-	-	-	-	-	-
SCUT‿DLVClab	-	-	-	74.1	80.2	77.0	-	-	-	-	-	-
Lyu *et al.*	-	-	-	70.6	74.3	72.4	76.2	87.6	81.5	32.4	61.9	42.5
Ours	**89.2**	91.3	**90.2**	**73.9**	**88.6**	**80.5**	81.6	**88.2**	**84.7**	**50.6**	**71.5**	**59.2**

train the Gabor Convolutional layers in the proposed framework with 4 orientation channels and let the scale setting remain 4. Momentum Optimizer is used with an initial learning rate of 0.01 and momentum of 0.9 during training. The learning rate is decreased by a factor of 10 after each subsequent 15k iterations. Data Augmentation is also employed to improve the robustness of the proposed framework. The training of the network has been carried out on a computer with two NVIDIA P6 GPU.

4.2. Evaluation Results

For performance evaluation of the proposed method, simulations have been done on several benchmark datasets. Among these *ICDAR2015* and *ICDAR2017 MLT* have image samples containing texts affected by motion blur, low-resolution, variable lighting and multi-scaled texts. Simulations done in [18] ignored image samples suffered from blur during training. However, our model has been purposefully trained on the entire training sets. Results of our extensive simulations have been compiled to construct Table 1.

5. CONCLUSION

In this article, we have presented a novel *Feature Representation Block* (FRB) of a fully convolutional neural network for efficient abstraction of features representing scene texts. For effective learning, we used curriculum learning strategy with respect to percentages of pixel-wise blurring and image sample complexities. Additionally, we used the *Mask and Predict* strategy to enable the network to produce satisfactory results in cases of partial occlusions. We obtained extensive simulation on four benchmark datasets which include *ICDAR 2015*, *ICDAR 2017 MLT*, *MSRA-TD500* and *COCO-Text*. Also, compared these results with the available SOTA performances. The proposed framework can efficiently detect incidental scene texts suffered by uneven lighting condition, blurring, partial occlusions and varying scales. In future, we shall study development of a strategy capable of detecting incidental curved texts. Also, we shall study development of a method to predict occluded texts from the visible partial information.

6. REFERENCES

[1] Dimosthenis Karatzas et al., "ICDAR 2015 competition on robust reading," in *ICDAR*, 2015, pp. 1156–1160.

[2] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie, "COCO-Text: Dataset and benchmark for text detection and recognition in natural images," *arXiv:1601.07140*, 2016.

[3] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu, "Gabor convolutional networks," *TIP*, vol. 27, no. 9, pp. 4357–4366, 2018.

[4] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick, "Inside-Outside Net: Detecting objects in context with skip pooling and recurrent neural networks," in *CVPR*, 2016.

[5] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin, "Crowd counting with deep structured scale integration network," in *ICCV*, 2019.

[6] Nibal Nayef et al., "ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt," in *ICDAR*, 2017, vol. 1, pp. 1454–1459.

[7] Yingying Zhu, Cong Yao, and Xiang Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 19–36, 2016.

[8] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, 2010, pp. 2963–2970.

[9] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *ICCV*, 2013, pp. 1241–1248.

[10] Lukas Neumann and Jiri Matas, "A method for text localization and recognition in real-world images," in *ACCV*, 2010, pp. 770–783.

[11] Michal Busta, Lukas Neumann, and Jiri Matas, "FAS-Text: Efficient unconstrained scene text detector," in *ICCV*, 2015, pp. 1206–1214.

[12] C. Lawrence Zitnick and Piotr Dollár, "Edge Boxes: Locating object proposals from edges," in *ECCV*, 2014, pp. 391–405.

[13] Piotr Dollár, Ron D. Appel, Serge J. Belongie, and Pietro Perona, "Fast feature pyramids for object detection," *TPAMI*, vol. 36, pp. 1532–1545, 2014.

[14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks.," *IJCV*, vol. 116, no. 1, pp. 1–20, 2016.

[15] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[16] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, "EAST: An efficient and accurate scene text detector," in *CVPR*, 2017.

[17] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu, "Deep direct regression for multi-oriented scene text detection," in *CVPR*, 2017, pp. 745–753.

[18] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan, "FOTS: Fast oriented text spotting with a unified network," in *CVPR*, 2018.

[19] Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li, "Scene text detection with supervised pyramid context network," in *AAAI*, 2019, vol. 33, pp. 9038–9045.

[20] Xiaobing Wang, Yingying Jiang, Zhenbo Luo, Cheng-Lin Liu, Hyunsoo Choi, and Sungjin Kim, "Arbitrary shape scene text detection with adaptive text region representation," in *CVPR*, 2019.

[21] Siyang Qin, Alessandro Bissacco, Michalis Raptis, Yasuhisa Fujii, and Ying Xiao, "Towards unconstrained end-to-end text spotting," in *ICCV*, 2019.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[23] Perla Sai Raj Kishore, Sudip Das, Partha Sarathi Mukherjee, and Ujjwal Bhattacharya, "ClueNet: A deep framework for occluded pedestrian pose estimation," in *BMVC*, 2019.

[24] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[25] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*, 2012, pp. 1083–1090.