Customized mask region based convolutional neural networks for un-uniformed shape text detection and text recognition

Ravikumar Hodikehosahally Channegowda[1], Palani Karthik[2], Raghavendra Srinivasaiah[3], Mahadev Shivaraj[4]

[1]Department of Electronics and Communication Engineering, K S School of Engineering and Management, Ghousia College of Engineering, Visvesvaraya Technological University, Karnataka, India
[2]Department of Electronics and Communication Engineering, K S School of Engineering and Management, Visvesvaraya Technological University, Karnataka, India
[3]Department of Computer science and Engineering, Christ Deemed to be University, Bengaluru, Karnataka, India
[4]Department of Electronics and Communication Engineering, Ghousia College of Engineering, Ramanagaram, Visvesvaraya Technological University, Karnataka, India

Article Info

Article history:

Received Oct 26, 2021
Revised Aug 7, 2022
Accepted Sep 6, 2022

Keywords:

Deep neural networks
Optical character-recognition
Region based convolutional neural networks
Region of interest
Text detection
Text recognition

ABSTRACT

In image scene, text contains high-level of important information that helps to analyze and consider the particular environment. In this paper, we adapt image mask and original identification of the mask region based convolutional neural networks (R-CNN) to allow recognition at 3 levels such as sequence, holistic and pixel-level semantics. Particularly, pixel and holistic level semantics can be utilized to recognize the texts and define the text shapes, respectively. Precisely, in mask and detection, we segment and recognize both character and word instances. Furthermore, we implement text detection through the outcome of instance segmentation on 2-D feature-space. Also, to tackle and identify the text issues of smaller and blurry texts, we consider text recognition by attention-based of optical character recognition (OCR) model with the mask R-CNN at sequential level. The OCR module is used to estimate character sequence through feature maps of the word instances in sequence to sequence. Finally, we proposed a fine-grained learning technique that trains a more accurate and robust model by learning models from the annotated datasets at the word level. Our proposed approach is evaluated on popular benchmark dataset ICDAR 2013 and ICDAR 2015.

Corresponding Author:

Ravikumar Hodikehosahally Channegowda
Department of Electronics and Communication Engineering, Visvesvaraya Technological University
Belagavi, Karnataka, India
Email: raviec40@gmail.com

1. INTRODUCTION

Text is one of the most expressive means of communication and can be embedded into documents or into scenes as a means of communicating information. Text plays a crucial role in our daily lives and reading it from videos and images are important values for the plentiful applications of the real world like scene text, image retrieval, and recognition [1], office automation, assistance for blind people and geo-locations consist of very convenient semantics for the understanding world. The reading of scene text gives a rapid and automatic way to access the data of textual embodied in the natural scenes that are most powerful representation given by scene text recognition and detection.

Text spotting intends to identify and localize the images of natural text that have been studied in prior techniques [2]. The techniques followed the traditional pipeline [3] treated the processes of text recognition and identification individually in that the text methods are trained the text detector then it is fed into the model of text recognition. This method looks straightforward and very simply but may lead to the performance of sub-optimal for identification and recognition, where two given methods are most relevant and complementary to one another. On the other hand, outcomes of recognition are highly dependent on recognized text. The recognition outcomes are very helpful for eliminating the detections of false positive (FP). Lately, He *et al.* [4] had started to integrate text recognition and detection with the help of an end-to-end trainable-network that contains 2 models like the sequence-to-sequence (Seq2Seq) network and identified a network for removing the text instances for estimating sequential labels for every single instance of text. The main performance developments for text spots are gained by the given techniques, representing recognition and identification model are corresponding in when they are also trained in the learning method of end to end.

In order to perform the text spot through 2 phases such as a detector is utilized to recognize the text in an image and after that, the recognition of text is used on the identified regions. The main drawbacks of these given techniques rely on correlation ignorance and time cost among the text recognition and identification. Thus, many techniques [5] were introduced to combine oriented and horizontal text recognition and identification in the manner of end to end. Anyways, the scene texts are often performing in an arbitrary shape. Instead of defining the text with quadrangles and rectangles, the text snake [6] defines the text with the units of local series that behaves like the snake. This work is mainly focused on text detection in an image, whereas the mask spotter of text is to recognize and identify the shapes of arbitrary text in which the char segmentation is utilized based on text identifier [7], thus char level annotation is needed at the time of training.

This paper represents unified techniques for arbitrary shape text-potting (ASTS) to identify and recognize the arbitrary shapes by considering 3 semantics levels. With the spotter of image mask text and mask region based convolutional neural networks (R-CNN) [8], we can articulate the detection of text as the segmentation of instance task. Unlike the spotter of image mask text, we adapt image mask and original mask R-CNN detection to allow identification at 3 levels such as sequence, pixel, and holistic-level semantics. Particularly, holistic and pixel-level semantics can be utilized to recognize the texts and define the text shapes respectively. Precisely, in mask and detection, we segment and recognize both character and word instances. Furthermore, we can implement the detection of text through the segmentation result on 2-D feature-space. Also, to tackle and identify the text issue of smaller and blurry texts, we contain recognition of text with an attention-based of optical character-recognition (OCR) model based on mask R-CNN to consider the semantics of the sequential level. We use an OCR module to estimate the character sequence through feature maps of the word-instances in Seq2Seq, which is implemented in 1-D feature-space. Lastly, we integrate 2 text detection outcomes based on edit distance among given lexicon and outcomes. In our techniques, texts can be analyzed at 3 semantics levels consisting of the sequential level for recognition mask, holistic level for an identified mask, and the pixel-level for the mask.

Finally, we propose a fine-grained learning technique that trains a more accurate and robust model with the help of learning models through annotated datasets at the word level. First, we trained our techniques utilizing a unified framework. Afterward, a trained model is applied to the annotated images at the word level to discover character samples and exploit the annotations at the word level to defeat the false character. Finally, the samples of the found character can be utilized as an annotation of character level and integrated with the word level annotations to train a larger model. This paper is represented: section 2 represents the prior work related to text spotting, recognition, and detection. Section 3 represents the proposed method before analyzing each model.

2. LITERATURE SURVEY

2.1. Text detection in image

Text identification plays an eminent part in the systems of text detection. Several techniques have been introduced to recognize scene text [9]. Neumann and Matas [10] utilizes edge boxes to create proposals and improve candidate boxes with the help of regression. Modifying solid-state drive (SSD) and the R-CNN with modifications, Ren *et al.* [11] introduced to recognize the horizontal words. Recently, the text identification of multi-oriented has become the most important topic.

Yao *et al.* [12] recognized an oriented based multi-scene text with the help of semantic segmentation. Zhang *et al.* [13] introduced techniques that recognize text segments and join them into text instances by link predictions and the spatial relationship, whereas Tian *et al.* [14] regressed the text boxes through the segmentation maps. Shi *et al.* [15] introduced to recognize and connect the text points to create the text boxes, whereas Lyu *et al.* [16] introduced the regression of rotation sensitivity for an oriented scene. Dai *et al.* [17], the region proposal and feature improvement network are recognized to use representations of improved features and instances of text shape to construct a bounding box. The pyramid region of interest (RoI) pooling attention is improvised, which removes the segmentation of text size feature. The network of a bounding box is utilized to remove the curve text.

2.2. Text recognition in image

Shi *et al.* [18] intended to decode the cropped and decoded image region into the character sequences for the recognition of scene text. The recognition of scene text techniques can be parted into the branches word-based techniques, Seq2Seq techniques, and character-based techniques. The character-based detection techniques [19] mostly localize the individual character and detect and connect them into the words. Jaderberg *et al.* [20] introduces the word-based technique which treats text detection as the issue of sequence labeling. Shi *et al.* [21] utilized recurrent neural network (RNN) and convolutional neural network (CNN) to model image features and outcome identified sequences with the help of centralized traffic control (CTC), whereas Lee and Osindero [22] identified the scene text through the attention based Seq2Seq model. Shelhamer [23] mainly focuses on demonstrating the complementarity module of character segmentation and module of spatial attention, instead of introducing the novel module of spatial attention. The spatial attention and the segmentation of integration characters do not minimize the required annotation of the character level for training but also generate a robust model for text shapes.

2.3. Text spotting in image

Prior text spotting techniques divide the spotting process into 2 phases. First, scene text identification was utilized [24] to contain text instances and then utilize the text recognizer to get identified text. The text spotter utilizes semantic segmentation to recognize the spatial attention and arbitrary text shapes for managing the irregular instances of text shapes by assuming global and local textual data [7]. The text dragon defined text shape with the help of quadrangles sequence in order to manage the RoI-slide and arbitrary text-shape, which interact the deep-network and also temporal based classification on text detector [25]. The characters of location labeling are not required. The Wacnet [26] related the shared convolutional network among char and word-level recognition and detection.

2.4. Object recognition in image

The development of object identification, instance/semantic segmentation, and deep learning (DL) have gained more improvement. The recognition and identification of scene text have gained attention in the last few years. Their technique is motivated by the given techniques. Specifically, their technique is also inspired by the help of mask R-CNN of instance segmentation method [8]. Anyways, there are key differences among mask branch technique which is introduced in mask-based R-CNN. The mask branch cannot segment the regions of text also estimate the probability of text sequences and character maps that means their technique can be utilized to detect an instance sequence within the character maps rather than estimating the mask of an object only.

3. PROPOSED METHOD

This section represents introduced method and then defines four main modules such as a connected network for extracting the image feature, text detection in an image for identifying character and word instances, text recognition in an image for segmenting the outcome of text identification. Afterward, we discover a fine-grained learning method for achieving our technique.

3.1. The architecture of the proposed method

Figure 1 represents the architecture of our proposed method. First, the image is faded into a shared network that utilizes the backbone network in order to remove the feature maps by the input image and it shares with the help of 3 subsequent like the operation of region-of-interest (RoI-align) and region-proposed network (PN). Afterward, text identification identifies the semantics of the holistic level to categorize and recognize the character instances and identify by rectifying region proposals utilizing the network of context refinement. Meanwhile, the image text mask evaluates pixel-level-semantics for recognized instances to conduct the segmentation. After that, an image test mask is utilized to define the character mask, and image text shape is applied for primary recognition. Since blurred and smaller characters are expected to suffer through failures, we introduced 2-D feature maps of the word instances into text detection to discover the semantics of seq-level for exact identification. Lastly, an outcome of primary recognition and recognition outputs are joined based on edit distance among outcomes and given lexicon.

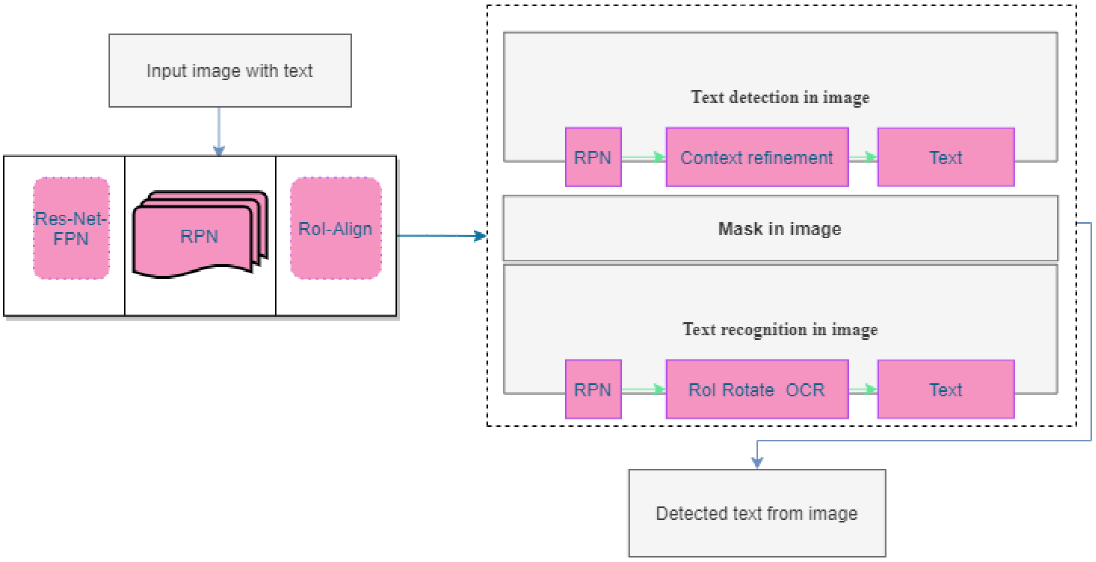

Figure 1. An architecture of proposed work

3.2. Connected network

This network contains risk priority number (RPN) and the backbone network. The feature of the image is removed with a shared and back-bone with a sub-sequent network. The feature-pyramid networks (FPNs) utilize the architecture of top-down to construct the feature map of a higher level at all of the scales through each input image and with the help of the marginal-cost. Hence, we utilize the FPN with the help of Res-Net, which is the backbone of the network.

The RPN is utilized to create the various sizes of the region and the aspect ratios for image mask and sub-sequent R-CNN, and mask-R-CNN. In our method, we implement the RPN to construct the region for three sub-sequent methods. We allocate various phases based on their sizes. Various aspect ratios are utilized in every single phase. RoI pooling assigns a floating number to distinct feature maps, it also leads to the misalignment issue. Then we apply the operation of RoI-align because it computes image-based features on the coordinates in the floating type with a method of the bilinear interpolation, thus it delivers accurate region alignments and the beneficial features to sub-sequent methods.

3.3. Text detection in image

The recognition intends region methods constructed with the help of RPN by estimating their classes and assuming the regression of the bounding box to alter their coordinates. This process is expressed as the identification of holistic-level semantics for region methods. This usually requires estimating two classes like non-text and text. In prior work [27], text recognition intended to recognize both character instances and words. Additionally, in non-text and text, we utilize the character data to train the identification method that helps to learn discriminative representation and improvise the performance of identification. Based on the spatial relationship between character and word instances, an outcome of character identification can be utilized for text identification. This is equal to implementing the character-based text identification technique while conducting text identification.

However, if methods overlap with a true word, existing refinement techniques may also suffer refinement failure. Since methods do not consist of sufficient data for the perception of a holistic object. Furthermore, the module of OCR is sensitive to the outcome of text localization. Therefore, we represent a module of context refinement to argument existing refinement in-text identification. From the refinement of context, the context data gathered from the surrounding regions are added into the representation of a unified context to improvise the identification performance.

The identification of text loss function is defined as (1).

$$Loss_d = Loss_c(\mathscr{p}, q) + \lambda[\mathfrak{u} \geq 1]Loss_l(\mathscr{t}^{\mathfrak{u}}, \mathfrak{v}) \quad (1)$$

Where the log loss is $Loss_c(p, q) = -log\mathscr{p}_{\mathfrak{u}}$ for $\mathfrak{u}$ true class and $\mathscr{p}_{\mathfrak{u}}$ is estimated the confidence score for the given class $\mathfrak{u}$. $Loss_l$ is described as a tuple of the bounding box of true regression targets for $\mathfrak{u}, \mathfrak{v}$ and $\mathscr{t}^{\mathfrak{u}}$ is defined as an estimated tuple.

The masking of text intends to conduct the semantic segmentation and examine pixel-level semantics for character and word instances. The predicted word masks are utilized to give exact shapes and the locations of a word than the identified method. Furthermore, character masks are applied for identification from a 2-D perspective. As represented in Figure 2, we implement the fully convolutional network (FCN) which has a similar image mask in the mask of R-CNN. The given RoI is constructed by the detected method, the RoI-align removes corresponding RoI features from fixed sizes and feature maps. We describe the loss function of image text mask $Loss_m$ as an average loss of binary cross-entropy, which is followed by the help of mask R-CNN. $Loss_m$ can be calculated:

$$Loss_m = -\frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\left[1 - b_n \times log\left(\mathbb{S}(a_n)\right) + (1 - b_n) \times log\left(1 - \mathbb{S}(a_n)\right)\right] \quad (2)$$

where the number of pixels is defined as $\mathcal{N}$, b_n is pixel label ($b_n \in (0,1)$), a_n is defined as estimated output and the function of the sigmoid is $\mathbb{S}(a_n)$.

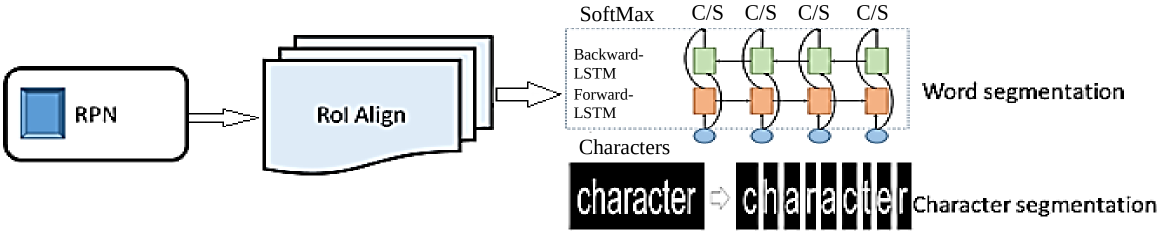

Figure 2. Illustration of the image mask text

3.4. Text recognition in image

The text recognition is intending to construct the module of OCR for discovering semantics of sequential level and estimating the character sequences through the word instances of feature maps. The recognition of text is treated as the Seq2Seq task. Moreover, the attention-based method has established the dependencies of effective modeling without distances in an input sequence. Therefore, we develop an OCR module for recognized text. The architecture of text recognition contains an OCR module and FCN. Like a prior image mask, we apply FCN that contains a deconvolution layer and 4 convolutional to scale up the maps of feature size.

We require to convert an input of 2-D feature map to feature-sequence before estimating character sequence then we implement the RoI-rotate before feeding into the module of OCR. For RoI, we achieve a rectangle with the rotation-angle by computing the minimum area of the rectangle mask which is produced with the help of an image mask. Based on rotation-angle, we also apply the RoI-rotate to alter the feature map-RoI horizontally. The feature maps are scaled to a fixed height and the aspect ratio is not changed with padding. Although an inevitably misleads the image features, we discuss the minimal impact on the performance since the same operation is used for the purpose of testing and training. Since the feature map is very short to reduce the oscillation of the network in the training phase and creates network coverage faster. We re-presented an experiment with an unchanged ratio aspect, but also performance was very poor.

Followed by a classical model of Seq2Seq, we develop an OCR module containing a decoder and encoder. An encoder alters feature-map into feature sequences (FS) and the decoder proposes to evaluate the character sequence by FS. Our encoder model is similar to the extracted FS network in convolutional recurrent neural network (CR-NN) except for an input, which has the feature map in the encoder but an input image in the CR-NN.

The encoder translates the feature-map to the FS via 4 max layers of pooling and 7 convolution layers. Afterward, FS is fed into the long short-term memory (LSTM) of multilayer bi-directional to catch high range dependencies of FS. Lastly, an encoder feeds output FS as context. Based on the Seq2Seq model, we implement RNN to construct the decoder that intends to estimate an outcome of the character sequence and also introduce the LSTM [28].

Let $A^w = \{a_0^w, a_1^w, a_2^w, \ldots., a_{W+1}^w\}$ is defined as the ground truth (GT) for the word instances w and $B^w = \{b_0^w, b_1^w, b_2^w, \ldots., b_{W+1}^w\}$ is defined as the corresponding decoder of the output sequence. The loss of recognition can be calculated by (3),

$$Loss_r = -\frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}}\sum_{a=1}^{W+1} log b_x(a_x) \quad (3)$$

where text number is defined as $\mathcal{N}$ that to be trained and $b_x(a_x)$ defines estimated output probability being a_b at the ath phase.

In order to integrate the loss of detection $Loss_d$ is defined in (1), the loss of image mask $Loss_m$ is defined in (2) and the loss of recognition $Loss_r$ is defined in (4). The loss function of a full multi-task is computed as (4),

$$Loss = Loss_{rpn} + Loss_d + Loss_m + Loss_r \tag{4}$$

where RPN loss is $Loss_{rpn}$.

The unified framework can identify and detect texts considering 3 sematic levels such as sequence, holistic, and pixel-level semantics and also give 2 recognition outcomes. The holistic is text spotting through image mask and text identification implemented from the 2-D perspective and another is an outcome of text recognition from a 1-D perspective. The outcome is robust to not accurate localization in the detection of text. Therefore, intending to improvise the recognition, we select a word with less distance from the lexicon as a final recognition outcome.

3.5. Fine-grained learning method

In order to identify the texts, the proposed method permits us to get promising performance on the basis of weak learning. Based on fine-grained learning, we propose a learning method, intending to train an accurate and robust text spotting method by acquiring through fine-grained learning with the help of trained technicians. First, we have to train the model $\mathbb{M}$ and weak learning $\mathbb{W}$ that has only the annotations of a word. This model is trained with a proposed method by utilizing the proposed method from the fully grained with character and word annotations. Each image has a poor annotation in $\mathbb{W}$ dataset at the word level which is defined as polygons set $\mathcal{G} = \{\mathscr{g}_1, \mathscr{g}_2, \dots., \mathscr{g}_{x,\dots}\}$. The fine-grained learning is to find the samples of character in the dataset $\mathbb{W}$.

The proposed fine-grained learning is represented in Figure 3, first, we apply trained model $\mathbb{M}$ on the dataset of word annotations $\mathbb{W}$. In dataset $\mathbb{W}$, we can get the samples of a candidate set $\mathbb{R}$.

$$\mathbb{R} = \{(p_0, c_0, \mathscr{b}_0, \mathscr{m}_0, r_0), (p_1, c_1, \mathscr{b}_1, \mathscr{m}_1, r_1), \dots.., (p_x, c_x, \mathscr{b}_x, \mathscr{m}_x, r_x), \dots \dots\}, \tag{5}$$

where, $p_x, \mathscr{b}_x, r_x, c_x$ and $\mathscr{m}_x$ is denoted as a predicted category, bounding box, recognition outcome, bounding box, and image mask outcome of $x - th$ sample of candidate character $\mathbb{c}_x$.

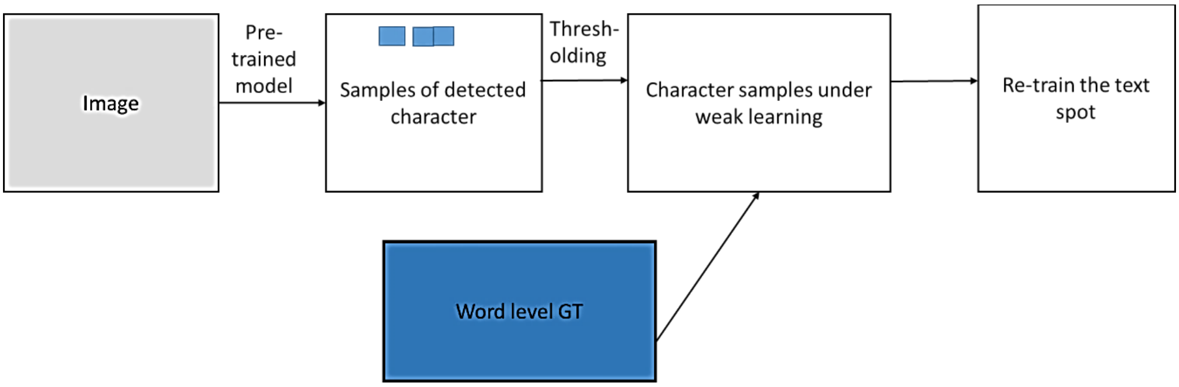

Figure 3. The pipeline of fine-grained learning method

In order to achieve the samples of positive character with an annotation of word-level $\mathbb{W}$ and threshold of confidence score.

$$\mathbb{P} = \left\{(p_x, c_x) | p_x \in C \;\; and \; c_x > S \frac{\mathscr{m}_x \cap \mathscr{g}_x}{\mathscr{m}_x} > W\right\} \tag{6}$$

Where C represents all categories of character to be removed, S defines the threshold of confidence score, which is utilized to detect samples of positive character, $\mathscr{m}_x \cap \mathscr{g}_y$ represents an intersection overlap-of the candidate character $\mathbb{r}_x$ with the level of word GT $\mathcal{g}_y$ and the threshold is W to choose the samples of positive character. The threshold of confidence score S can set less because of constraints given by the annotations of word-level that is also useful for preserving the samples of diversity character. Lastly, recognized samples of positive character W can be used as character annotations and it can be integrated with the word-level-annotations $\mathcal{G}$ to train an accurate and robust model of the text spot.

4. RESULTS AND DISCUSSION

In this section, we will discuss the result and analysis of our proposed approach, which will be evaluated on most popular benchmarks dataset ICDAR 2013 [29] and ICDAR 2015 [30]. The proposed model is compared with state-of-the-art methods. Furthermore, we will discuss experimental environment and parameters, where the experiments were performed at Intel i5-7gen processor, 6 GB RAM, 1 TB solid state drive and GPU of NVIDIA GTX2080Ti at PyTorch platform. The network is optimized by stochastic gradient descent, images were resized to 640 x 640 with batch size of 12. The starting learning rate is considered to be 10-3 and after 100 epochs learning rate is reduced to 10^{-4}.

4.1. ICDAR dataset, detection, and recognition protocol

There are two datasets used: ICDAR 2013 and ICDAR 2015. In ICDAR 2013 the number of training images are 229 (which consist of 849 words) and the number of testing images are 233 (which consist of 1,095 words). In this dataset, the texts placed are horizontal and annotated by the rectangles in words. ICDAR 2015 is given at challenge of ICDAR 2015 “Robust reading competition for incidental scene text detection”, where it contains 1,000 training images (which consist of 11,886 words) and testing images of 500 (which consist of 5,230 words). In image dataset, the text areas are annotated by 4 quadrangle vertices. Here, we provided analysis of evaluation protocols for recognition and text detection, and the process of text detection is evaluated by the ICDAR protocol. In general, text recognition is evaluated using end-to-end recognition procedure or word recognition accuracy.

In order to find the best possible match $t(u, U)$ for a rectangular u in the rectangle set U can be written as (7),

$$t(u, U) = maxU_m(u, u')|u' \in U \tag{7}$$

where w_m shows the match between two considered instances of text rectangles. It can be computed by intersection area divided via minimum area of bounding box (i.e., contain both the rectangles), then the considered evaluation parameters such as precision, recall and F-score can compute as (8),

$$\text{Precision(P)} = \frac{\sum_{u_y \in Y} t(u_y, Z)}{|Y|} \tag{8}$$

$$\text{Recall(R)} = \frac{\sum_{u_z \in Z} t(u_z, Y)}{|Z|} \tag{9}$$

$$-\text{score} = \frac{1}{\frac{w}{P} + \frac{(1-w)}{R}} \tag{10}$$

where Z and Y represent the estimated rectangles and ground-truth set. The u_z and u_y represents the estimated rectangles and ground-truth, w is weight parameters. Text recognition performance is always measured by the accuracy of word recognition. The word recognition accuracy (WRA) is simply defined as the percentage of the recognized text is correct, as in (11).

$$WRA = \frac{\text{Number of words correctly recognized}}{\text{Number of ground truth words}} \tag{11}$$

4.2. Proposed model comparison with state-of-art techniques

Our proposed model has been applied on some images and in Figure 4 shows the sample of texts detected and recognized from ICDAR 2013 dataset. Table 1 shows the recall, precision, and F-score comparison with state-of-art techniques at ICDAR 2013 dataset, ICDAR 2013 dataset comprises a huge amount of text data that stretches across the entire image. As per table we can clearly say that our model has performed considerably well at evaluation parameters.

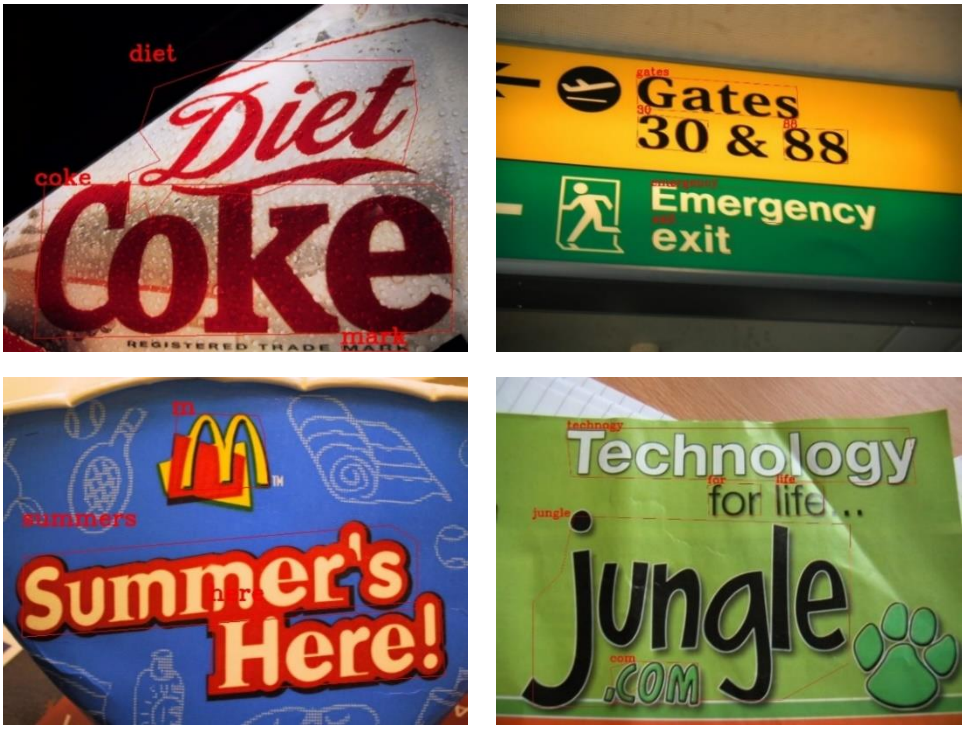

Figure 4. Sample of texts detected from ICDAR 2013 dataset

Table 1. Recall, precision, and F-Score comparison with state-of-art techniques at ICDAR 2013 dataset

Method	Recall	Precision	F-Score
Textboxes++ [3]	74	88	81
Mask TextSpotter [7]	88.1	94.1	91
Connectionist text proposal network [14]	83	93	88
SegLink [15]	83	87.7	85.3
Lyu *et al.* [16]	79.4	93.3	85.8
Texboxes [24]	74	88	81
Fast oriented text spotting [27]			86.96
Fast oriented text spotting-Recg [27]			88.23
Deep direct regression [31]	81	96	86
Rotation region proposal networks [32]	72	90	80
PixelLink [33]	83.6	86.4	84.5
Border ResNet-50 [34]	86.9	87.8	87.4
Border DenseNet-121 [35]	87.1	91.5	89.2
Markov clustering network [34]	87	88	88
Rotation-sensitive regression detector [36]	75	88	81
He *et al.* [37]	**89**	95	91
HAM-ResNet50 [38]	81.28	89.62	85.25
HAM-ResNet50+IRB [38]	82.55	92.3	87.17
HAM-DenseNet-169 [38]	83.47	94.42	88.6
HAM-DenseNet-169+IRB [38]	83.56	94.52	88.7
Proposed method	90.56	97.83	94.05

Proposed model has obtained recall, precision and F-score as 90.56, 97.83 and 94.05, respectively which 1.56%, 2.83%, 3.05% more compared to He *et al.* [37]. Figure 5 shows the sample of texts detected from ICDAR 2015 dataset, samples in training set include 1000 images and 500 images from the ICDAR 2015 dataset. Table 2 shows the recall, precision, and F-score comparison with state-of-art techniques at ICDAR 2015 dataset. Hidden anchor mechanism (HAM)-DenseNet-169+iterative regression box (IRB) [38] has achieve an F-score of 89.21%, and HAM-DenseNet-169+IRB [38] has achieved a precision and recall of 90.69% and 87.77%. Our proposed method achieved an F-score of 89.61% and outperforms HAM-DenseNet-169+IRB [38] by 0.44% in terms of the F-score. In Table 3, we have shown recognition comparison with state-of-art techniques. Our proposed method has got 95.3% recognition accuracy in ICDAR 2013 dataset and got 81.1% recognition accuracy at ICDAR 2015 dataset, which is better than previous techniques.

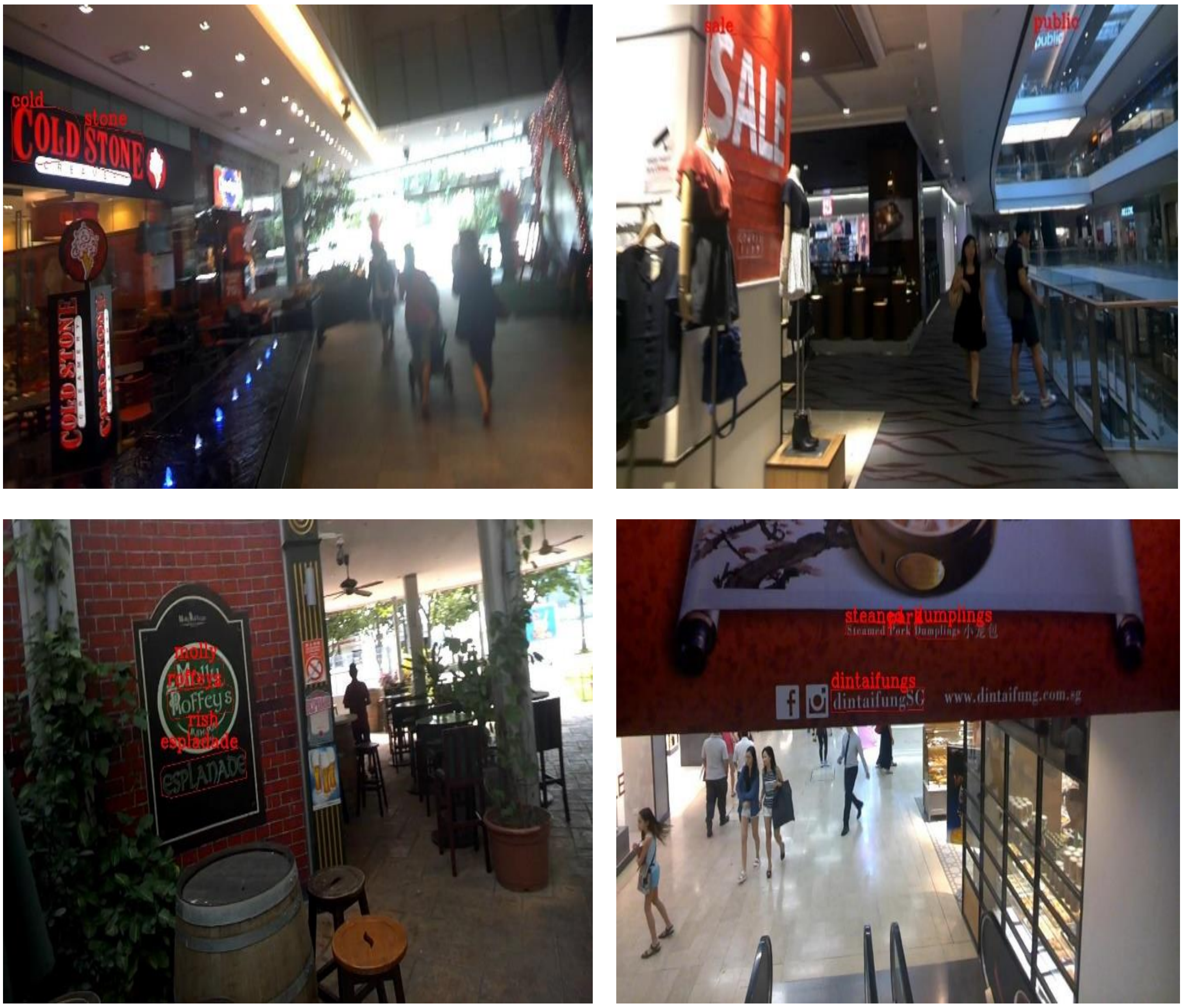

Figure 5. Sample of texts detected from ICDAR 2015 dataset

Table 2. Recall, precision, and F-Score comparison with state-of-art techniques at ICDAR 2015 dataset

Method	Recall	Precision	F-Score
Textboxes++ [3]	78.5	87.8	82.9
TextSnake [6]	84.9	80.4	82.6
Mask TextSpotter [7]	81	91.6	86
SegLink [15]	76.8	73.1	75
Lyu *et al.* [16]	79.76	89.5	84.3
Fast oriented text spotting [27]	82.04	88.84	85.31
Fast oriented text spotting-Recg [27]	85.17	91	87.99
Deep direct regression [31]	80	88	83.8
Rotation region proposal networks [32]	77	84	80
PixelLink [33]	82	85.5	83.7
Markov clustering network [34]	80	72	76
Rotation-sensitive regression detector [36]	79	85.6	82.2
He *et al.* [37]	80	85	82
HAM-ResNet50 [38]	85.84	89.82	87.79
HAM-ResNet50+IRB [38]	85.8	89.9	87.8
HAM-DenseNet-169+IRB [38]	**87.77**	90.69	89.21
Instance transformation network [39]	74.1	85.7	79.5
Efficient and accurate scene text (EAST) [40]	78.3	83.3	80.7
Fused text segmentation networks [41]	80	88.6	84.1
IncepText [42]	84.3	89.4	86.8
Tian *et al.* [43]	85	88.3	86.6
Look more than once (LOMO) [44]	83.5	91.3	87.2
Proposed method	86.82	**92.58**	**89.61**

Table 3. Recognition performance comparison

Methods	ICDAR-13	ICDAR-15
[19]	87.6	
[21]	89.6	
[22]	90.0	
[45]	92.8	80.0
[46]	81.8	
[47]	88.6	
[48]	90.8	
[49]	93.3	70.6
[50]	85.2	
[51]	91.1	60.0
[52]	92.9	
[53]		68.2
[54]	94.0	
[55]	94.4	73.9
[56]	91.5	
[57]	92.4	68.8
[58]	89.7	68.9
[59]	91.0	69.2
[60]	91.3	76.9
[61]	93.9	78.7
[62]	91.9	74.2
[63]	92.9	75.5
[64]		76.1
Proposed method	95.3	81.1

5. CONCLUSION

The demand and requirement of scene text detection and recognition have got cumulative attention in various fields due to its potential. This paper we proposed an optimized detection and recognition approach, text detection in an image is performed for identifying character and word instances. Text recognition in an image for segmenting the outcome of text identification. In addition, we discovered a fine-grained learning method to achieve optimized outcome. For evaluation, the most popular benchmarks dataset ICDAR 2013 and ICDAR 2015 is considered, where the proposed model is compared with state-of-the-art methods. Evaluation is performed based upon the recall, precision, F-score, and recognition performance. The proposed model has performed better as compared with state-of-the-art techniques. In future work, we will use our detection and recognition framework in video dataset.

ACKNOWLEDGEMENTS

We gratefully thank the Visvesvaraya Technological University, Jnana Sangama, Belagavi for financial support extended to this research work.

REFERENCES

[1] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo, "Integrating scene text and visual appearance for fine-grained image classification," *IEEE Access*, vol. 6, pp. 66322–66335, 2018, doi: 10.1109/ACCESS.2018.2878899.

[2] M. Busta, L. Neumann, and J. Matas, "Deep TextSpotter: an end-to-end trainable scene text localization and recognition framework," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2223–2231, doi: 10.1109/ICCV.2017.242.

[3] M. Liao, B. Shi, and X. Bai, "TextBoxes++: a single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018, doi: 10.1109/TIP.2018.2825107.

[4] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end textspotter with explicit alignment and attention," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5020–5029, doi: 10.1109/CVPR.2018.00527.

[5] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5248–5256, doi: 10.1109/ICCV.2017.560.

[6] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: a flexible representation for detecting text of arbitrary shapes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11206, 2018, pp. 19–35.

[7] 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 *Computer Vision ECCV 2018*, Springer International Publishing, 2018, pp. 71–88.

[8] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[9] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: recent advances and future trends," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 19–36, Feb. 2016, doi: 10.1007/s11704-015-4488-0.

[10] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 3538–3545, doi: 10.1109/CVPR.2012.6248097.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[12] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," CoRR abs/1606.09002, Jun. 2016.

[13] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 4159–4167, doi: 10.1109/CVPR.2016.451.

[14] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Computer Vision ECCV 2016*, Springer International Publishing, 2016, pp. 56–72, doi: 10.1007/978-3-319-46484-8_4.

[15] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3482–3490, doi: 10.1109/CVPR.2017.371.

[16] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7553–7563, doi: 10.1109/CVPR.2018.00788.

[17] P. Dai, H. Zhang, and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 1969–1984, Aug. 2020, doi: 10.1109/TMM.2019.2952978.

[18] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: an attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019, doi: 10.1109/TPAMI.2018.2848939.

[19] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: reading text in uncontrolled conditions," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 785–792, doi: 10.1109/ICCV.2013.102.

[20] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv.1406.222*, 2014. doi:10.48550/arXiv.1406.2227.

[21] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017, doi: 10.1109/TPAMI.2016.2646371.

[22] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2231–2239, doi: 10.1109/CVPR.2016.245.

[23] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, Apr. 2017, doi: 10.1109/TPAMI.2016.2572683.

[24] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: a fast text detector with a single deep neural network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11196.

[25] W. Feng, W. He, F. Yin, X.-Y. Zhang, and C.-L. Liu, "TextDragon: an end-to-end framework for arbitrary shaped text spotting," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 9075–9084, doi: 10.1109/ICCV.2019.00917.

[26] Y. Gao, Z. Huang, Y. Dai, K. Chen, J. Guo, and W. Qiu, "Wacnet: word segmentation guided characters aggregation net for scene text spotting with arbitrary shapes," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 3382–3386, doi: 10.1109/ICIP.2019.8803529.

[27] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: fast oriented text spotting with a unified network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5676–5685, doi: 10.1109/CVPR.2018.00595.

[28] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *arXiv:1506.07503v1*, Jun. 2015, doi: 10.48550/arXiv.1506.07503.

[29] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 1484–1493, doi: 10.1109/ICDAR.2013.221.

[30] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug. 2015, pp. 1156–1160, doi: 10.1109/ICDAR.2015.7333942.

[31] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 745–753, doi: 10.1109/ICCV.2017.87.

[32] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018, doi: 10.1109/TMM.2018.2818020.

[33] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: detecting scene text via instance segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12269.

[34] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, and W. L. Goh, "Learning Markov clustering networks for scene text detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6936–6944, doi: 10.1109/CVPR.2018.00725.

[35] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *Computer Vision ECCV 2018*, Springer International Publishing, 2018, pp. 370–387.

[36] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5909–5918, doi: 10.1109/CVPR.2018.00619.

[37] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5406–5419, Nov. 2018, doi: 10.1109/TIP.2018.2855399.

[38] J.-B. Hou *et al.*, "HAM: hidden anchor mechanism for scene text detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 7904–7916, 2020, doi: 10.1109/TIP.2020.3008863.

[39] F. Wang, L. Zhao, X. Li, X. Wang, and D. Tao, "Geometry-aware scene text detection with instance transformation network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1381–1389, doi: 10.1109/CVPR.2018.00150.

[40] X. Zhou *et al.*, "EAST: an efficient and accurate scene text detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2642–2651, doi: 10.1109/CVPR.2017.283.

[41] Y. Dai *et al.*, "Fused text segmentation networks for multi-oriented scene text detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 3604–3609, doi: 10.1109/ICPR.2018.8546066.

[42] Q. Yang, M. Cheng, W. Zhou, Y. Chen, M. Qiu, and W. Lin, "IncepText: a new inception-text module with deformable PSROI pooling for multi-oriented scene text detection," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul. 2018, pp. 1071–1077, doi: 10.24963/ijcai.2018/149.

[43] Z. Tian *et al.*, "Learning shape-aware embedding for scene text detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 4229–4238, doi: 10.1109/CVPR.2019.00436.

[44] C. Zhang *et al.*, "Look more than once: an accurate detector for text of arbitrary shapes," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 10544–10553, doi: 10.1109/CVPR.2019.01080.

[45] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: semantics enhanced encoder-decoder framework for scene text recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 13525–13534, doi: 10.1109/CVPR42600.2020.01354.

[46] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep structured output learning for unconstrained text recognition," *Computer Vision and Pattern Recognition*, 2014, doi: 10.48550/arXiv.1412.5903.

[47] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 4168–4176, doi: 10.1109/CVPR.2016.452.

[48] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, Jan. 2016, doi: 10.1007/s11263-015-0823-z.

[49] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: towards accurate text recognition in natural images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5086–5094, doi: 10.1109/ICCV.2017.543.

[50] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu, "Scene text recognition with sliding convolutional character models," *Computer Vision and Pattern Recognition*, Sep. 2017.

[51] W. Liu, C. Chen, and K.-Y. Wong, "Char-Net: a character-aware neural network for distorted scene text recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12246.

[52] Z. Liu, Y. Li, F. Ren, W. L. Goh, and H. Yu, "SqueezedText: a real-time scene text recognition by binary convolutional encoder-decoder network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12252.

[53] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: towards arbitrarily-oriented text recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5571–5579, doi: 10.1109/CVPR.2018.00584.

[54] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *Computer Vision ECCV 2018*, Springer International Publishing, 2018, pp. 449–465.

[55] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit probability for scene text recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1508–1516, doi: 10.1109/CVPR.2018.00163.

[56] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8714–8721, Jul. 2019, doi: 10.1609/aaai.v33i01.33018714.

[57] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognition*,

vol. 90, pp. 109–118, Jun. 2019, doi: 10.1016/j.patcog.2019.01.020.

[58] Z. Xie, Y. Huang, Y. Zhu, L. Jin, Y. Liu, and L. Xie, “Aggregation cross-entropy for sequence recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 6531–6540, doi: 10.1109/CVPR.2019.00670.

[59] H. Li, P. Wang, C. Shen, and G. Zhang, “Show, attend and read: a simple and strong baseline for irregular text recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8610–8617, Jul. 2019, doi: 10.1609/aaai.v33i01.33018610.

[60] F. Zhan and S. Lu, “ESIR: end-to-end scene text recognition via iterative image rectification,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 2054–2063, doi: 10.1109/CVPR.2019.00216.

[61] M. Yang *et al.*, “Symmetry-constrained rectification network for scene text recognition,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 9146–9155, doi: 10.1109/ICCV.2019.00924.

[62] C. Wang and C.-L. Liu, “Scene text recognition by attention network with gated embedding,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206802.

[63] X. Chen, T. Wang, Y. Zhu, L. Jin, and C. Luo, “Adaptive embedding gate for attention-based scene text recognition,” *Neurocomputing*, vol. 381, pp. 261–271, Mar. 2020, doi: 10.1016/j.neucom.2019.11.049.

[64] C. Luo, Y. Zhu, L. Jin, and Y. Wang, “Learn to augment: joint data augmentation and network optimization for text recognition,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 13743–13752, doi: 10.1109/CVPR42600.2020.01376.

BIOGRAPHIES OF AUTHORS

Ravikumar Hodikehosahally Channegowda is a research scholar in the Department of Electronics and Communication Engineering at KSSEM, Bengaluru, affiliated to VTU, Belagavi and currently working as asst. professor in Dept. of ECE at Ghousia College of Engineering, Ramanagaram. He has done his masters in VLSI design and embedded systems from VTU Extension Centre, PESCE, Mandya. His areas of interest are image processing, machine learning, pattern recognition and multimedia concepts. He can be contacted at raviec40@gmail.com.

Palani Karthik received a doctoral degree from Dr MGR University, Master from Sathyabama University, during the year 2013 and 2006. Currently he is working as a professor in the Dept. of Electronics and Communication Engineering at K S School of Engineering, Bengaluru. His current research interests are acoustics sensors, image processing, machine learning, smart grids and mobile communication. He is actively involved in various professional bodies like IEEE senior member, member in IEI and Member in ISTE and IAENG and ACM. He can be contacted at karthik.p@kssem.edu.in.

Raghavendra Srinivasaiah is currently working as Associate Professor in the Department of Computer Science and Engineering at Christ Deemed to be University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has 17 years of teaching experience. His interests include data mining, artificial intelligence, and big data. He can be contacted at raghav.trg@gmail.com.

Mahadev Shivaraj received the M.Tech. Degree in VLSI Design and Embedded System from R.V.C.E, Bengaluru, under VTU, Belagavi, Karnataka, India in 2007. He received the B.E. degree in Electronics and Communication Engineering from VTU, Belagavi, Karnataka, India in 2003. He is currently working as an assistant professor in Ghousia College of Engineering, Ramanagara. He has published two patents. His area of research interest includes image processing, hardware description language machine learning and Python. He can be contacted at mahadevkhanderao@gmail.com.