IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 4, April, 2025, pp. 1399-1408
Journal homepage: http://iieta.org/journals/mmep

DoubleYolo: Efficient Scene Text Detection Using Double Edge Method and YOLOv8n

M. Mahesha[1*], V. N. Manjunath Aradhya[1], H. T. Basavaraju[2], Siddesha Shivarudraswamy[1]

[1] Department of Computer Applications, JSS Science and Technology University, Mysuru 570006, India
[2] Department of Computer Applications, Vidyavardhaka College of Engineering, Mysuru 570002, India

Corresponding Author Email: mahesha_m@sjce.ac.in

https://doi.org/10.18280/mmep.120430

Received: 7 January 2025
Revised: 3 March 2025
Accepted: 10 March 2025
Available online: 30 April 2025

Keywords:
double edge, Double Yolo, multi oriented text, multi-lingual scene text detection, scene text detection, text extraction

ABSTRACT

Day by day environmental and architectural changes influenced the society and world, scene texts are also changing with various styles and dimensions. There is need to detect and understanding of text in scene images like name plates, bill boards and bus routes to assist tourists and automated environments. Scene text detection poses various challenges like complex background, multilingual, multi-orientation, occlusion and poor lighting effects. Many methods have developed using machine learning and deep learning models but not achieved significant impact due to either heaviness of models and involve much training and testing of large number of images. Hence the proposed algorithm implemented the double edge method with YOLOv8n to detect scene text in images. In real world scenario, the text components exhibit double line structures with cyclic edges in nature. Using this property double edge method retains the prominent text components at primary stage. Further by employing YOLOv8n which refines the fine-grained textual components from scene images. The proposed algorithm is simple approach and yields better efficacy even with the smaller number of trained samples. The experimentation conducted on benchmark datasets like CTW1500, MSRA TD500, Total Text, MRRC, and MLe2e and serves handy for scene text detection/recognition tasks.

1. INTRODUCTION

In the context of scene text detection, day by day the outdoor environment is posing challenging issue, because of scene text appears with many constraints like multi colored, multilingual, multi orientation with various styles and fonts so on [1]. Scene text detection is the process of detecting and localizing text in the natural images, such as information boards on streets, advertisements, hoardings, transportation route boards and product labels, which are collected from constrained surroundings. The text detection process identifies the comprehend textual data present in the complex situation [2]. The significant application includes detecting labels, detecting product details, autonomous delivery systems, and navigation routes for robots, and drones. In the ever-changing environment, scene text detection poses considerable obstacles due to differences in text appearance, variety of fonts, styles, orientation, multilingual nature, occlusions, dull colored images, foggy images and images with complex background [3, 4]. Scene text detection has become independent stream, encompasses its area not just locating, detecting, and helps to extract the semantic information from images, used in many real time applications [5, 6]. Traditional Optical Character Recognition (OCR) systems perform satisfactorily on scanned or organized documents, but when used on images of scenic image context, they may have trouble with text orientation, noise, and background clutter. Scene text detection, along with OCR, helps to address these issues by first detecting and then recognizing the text [7]. Deep learning algorithms have recently emerged as the main solution to handling this challenging problem, providing significant improvements over older methods [8]. Scene text detection helps to understand the events and analyse the semantics of the respective image. Scene text detection also helps to promote multiculturalism and globalization by automatically identifying hybrid script and translate text between languages.

In the proposed design initially scene images fed into double edge process, in which it generates the prominent text components which shares the textual edges and associated properties. Then output of double edge process is given to YOLOv8n (nano) model as part of the post processing to filter the fine grain text parts, which generates candidate text regions with bounding boxes around it with confidence scores. Annotation done using LabelImg tool, annotated images and ground truth values are supplied for training and validation process, during training, model generates the best model with most efficient weight. Finally, most efficient generated model is tested against popular datasets to evaluate efficiency of the model. The model trained and validated with various ratios like 70:30, 60:40 with 100 epochs. Proposed combination of Double Yolo method with YOLOv8n model produces potential textual components with bounding boxes around it with its confidence scores. Proposed algorithm considers those bounding boxes with confidence scores greater than 50 percent as prime text components otherwise deemed to reject the selection as non-text component, so that helps in separation of falsely selected objects. Observations found that outputs produced with high accuracy with simplistic approach. In proposed Double Yolo method, the model trained for very fewer number of samples and yet yielded far better results, that implies ours is simplified approach which reduces the tedious task of training and validation process than other models which requires huge amount of training and testing and saved the time consumption. The experiments done without double edge contribution also, but it has produced very poor results as shown in section 4.8. Colab environment with Intel Xeon processor, Python-3.10.12 torch-2.5.0+cu121 is used as software platform. The proposed technique achieved good accuracy on benchmark datasets like CTW1500, MSRA TD500, Total Text, MRRC, and MLe2e. Significant feature of the proposed Double Yolo algorithm is that it has yield better efficacy even with the very smaller number of samples and serves handy for text detection and recognition applications with natural scene images and videos.

2. RELATED WORKS

Many researchers around the world involved to find out better algorithm for efficient scene text detection and recognition which include traditional as well-advanced deep learning techniques. Many machine learning approaches attempted in detection and localizing the text from scene images. Spatial relationship of adjacent pixel used to generate the co-joint textual components using MCN and Em algorithm [9, 10].

Deep learning is effectively useful for scene text detection since it does not require any pre-programmed rules or manually creating features. Because, it learns directly from the input data, by allowing the system to deal with the varied and unexpected nature of text in natural scene images. The deep learning methods adjust to changing typefaces, sizes, and directions, as well as deal with background distraction and noise information. Convolution neural network is an essential architecture for scene text identification in images, because it extracts organizational features from images. This architecture trains the system from low to high-level features, such as edges, textures, and formats, which are necessary for text detection and localization in complex contexts [8]. The popular frameworks such as VGG, ResNet, and DenseNet are commonly employed as the foundation for obtaining the features from textual images prior to identify the actual text sections [11]. Linguistic image pre-training model is employed with 2 encoders, a text encoder and an image encoder used to detect the texts on given scene images. The approach yielded better efficiency in locating and identifying textual contents by combining visual and textual data [12]. Zhong et al. [13] concentrated on centralized mask projection for text instance detection in images. A bounding box regression concept is employed for tight bounding box around the text information.

To extract the curved texts and multi oriented texts a frame work is developed based on dynamic convolution. Using multi features it creates the distinct text sample–aware convolutional variables this enables separation of textual components from non-text background [14]. The approach mainly focuses on encoding shape and location information based on the features of the text instance. The problem of narrow quadrilateral bounding boxes is solved by employing the angle regression model. Further to predict the bounding boxes, linear regression used by anticipating the coordinates across the quadrilateral points [15]. The pixel wise representational traits shared by kernel version and fully convolutional neural network that efficiently represent text cases, efficiently and accurately [16]. Rotational YOLO model employed to locate the arbitrary shaped texts from scene images. The bounding boxes with various orientations are drawn by using rotational anchor block. The detection accuracy is also improved from the input images of various scaling [17]. Since the regression-based models fails to locate the texts with bigger sizes in scene images, a location-based feature selection system developed to separate the texts from scene images [18]. To distinguish the text watermark and scene text from videos frames a robust framework developed by associating UNet3+ and Fourier contour embedding network to clearly categorizing textual components from input images [19].

An approach based on graphs for identifying directed scene text. The framework consists of three parts: optical identification, language, and graph-based textual reasoning. The researchers achieved a high degree of accuracy, but with a big number of variables and a sophisticated design that users may find difficult to repeat. Researchers found a trade-off among recognition accuracy and model complexity, with the quantity of layers increasing in terms of factors, reasoning time, and structural design [20]. A four-stage approach proposed to recognizing scene text in natural images. They have clearly demonstrated the trade-off among precision and effectiveness. The investigation discovered that recurring convolutional neural networks collect fewer features compared to other techniques. Variables are compared to large, complex neural network architectures such as VGG and ResNet. The most accurate model uses ResNet to extract features, which improves accuracy but reduces performance due to computational demands and memory utilization [21].

As per existing research, many text detection methods face challenges in identification of accurate textual features. Traditional techniques often struggle with variations in font style, size, orientation, and background noise. While deep learning-based models require an extensive amount of labelled training data across different scenarios. The process of annotating and training on such large datasets is resource-intensive and highly time-consuming also they demand substantial computational power for both training and inference, making them less feasible for resource-constrained environments.

Also, latest developments found that deep learning methods can detect the text only in certain pipelines. Hence, this research work proposed the Double Yolo algorithm. The Double edge method uses double stroke structure property of textual components thus eases the filtering of textual features from scene images then pipelined to YOLOv8n to further separation of texts from non-text background. Due to double edge approach, the text separation in complex background become simplistic approach and achieved good efficacy even with minimal training. Experimentation includes image scenarios with multilingual, multi orientation and challenging circumstances.

3. PROPOSED MODEL

Scene text detection is the process of determining and localizing the textual information in natural scene images. The proposed research work implemented double edge concept with Yolo V8 architecture to identify the textual information in scene text images with complex scenarios like noisy background, low illumination, different orientation, scale, and size. Initially, double edge concept is employed to generate the actual text candidates. Later, the LabelImg tool is utilized for labelling the actual text region for the outcome of doble edge process. Then annotated images is further post processed with YOLOv8n model to generate potential text candidates in scene text images, finally fine-grained textual components are generated with bounding boxes around it with high confidence scores. Proposed algorithm considers those bounding boxes with confidence scores greater than 50 percent for further fine tune process of text recognition which produces the actual words. The bounding boxes with less than 50 percent confidence scores are deemed to be rejected as non-text component. Observations found that outputs produced with high accuracy. Experimentation done for image ratios like 70:30 and 80:30 by considering few 100 samples. State of the art datasets like CTW1500, MSRA-TD500, Total Text, MRRC, and MLe2e. Finally, the proposed double edge based YOLOv8n model accurately locates the actual text regions in the complex scene text images. Flow diagram of the proposed Double Yolo method is shown in Figure 1.

3.1 Double edge for extracting possible text candidates

In the real-world scenario, normally text characters have double line structure. That means textual component exhibits two-line structure, such that starting and ending points are same and cyclic in nature. Text pixels also has continuous and uniform distribution nature. Whereas non-text components do not exhibit this property due to non-continuous pixel information. The double structure algorithm scans the uniform pixel information to generate circular outer boundary around the text region and filters the prominent textual components. If outer boundary is not circular then it is determined as non-text region and discarded as non-text components.

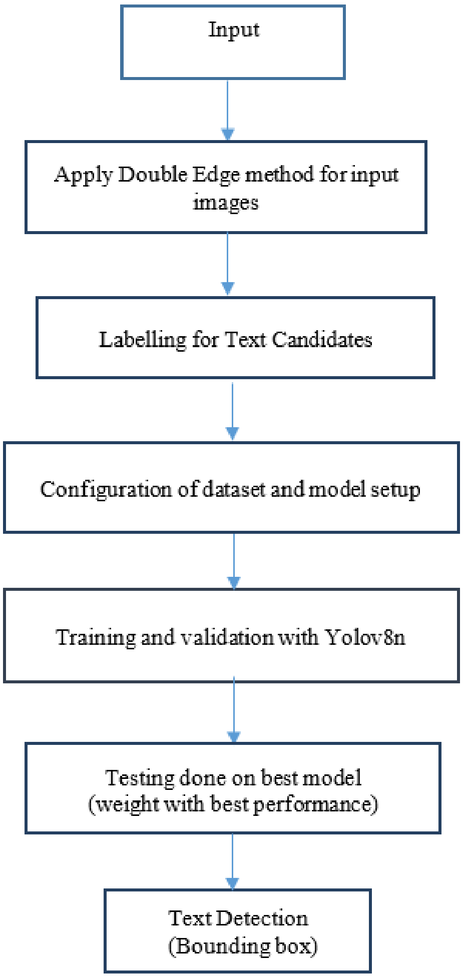

Input
Apply Double Edge method for input images
Labelling for Text Candidates
Configuration of dataset and model setup
Training and validation with Yolov8n
Testing done on best model (weight with best performance)
Text Detection (Bounding box)

Figure 1. Process flow diagram for Double Yolo method

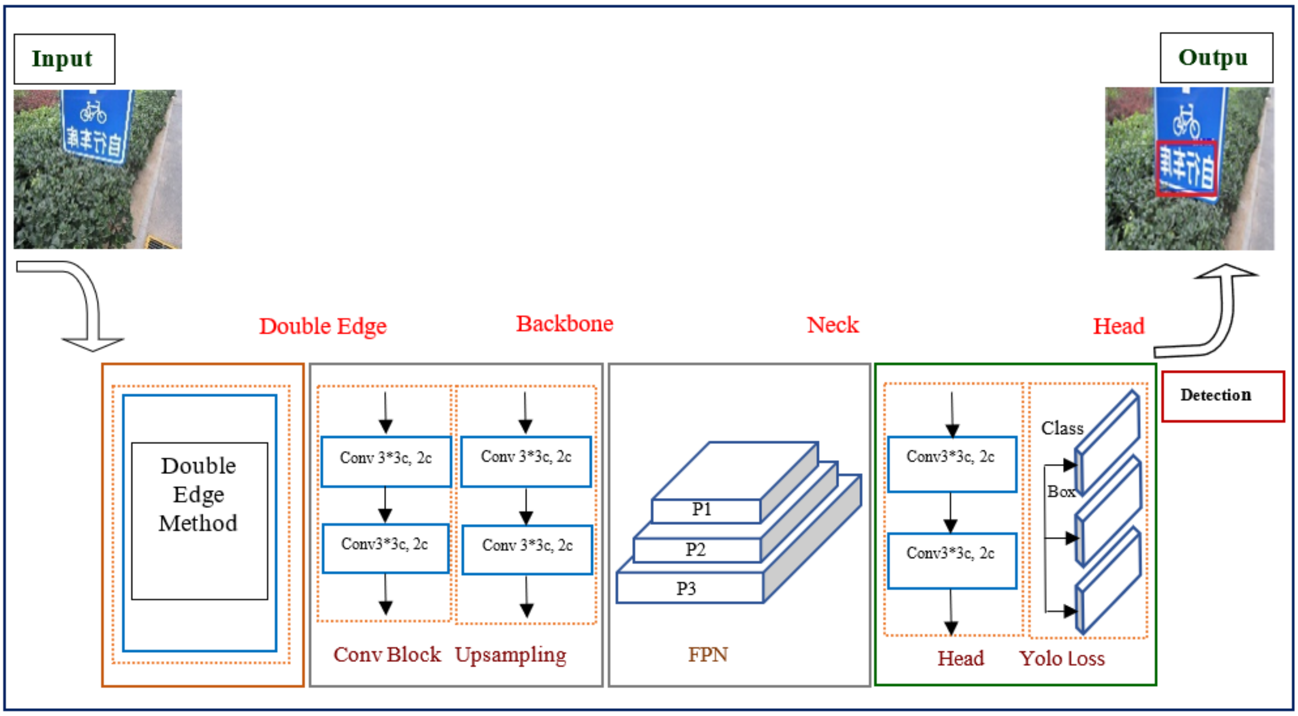
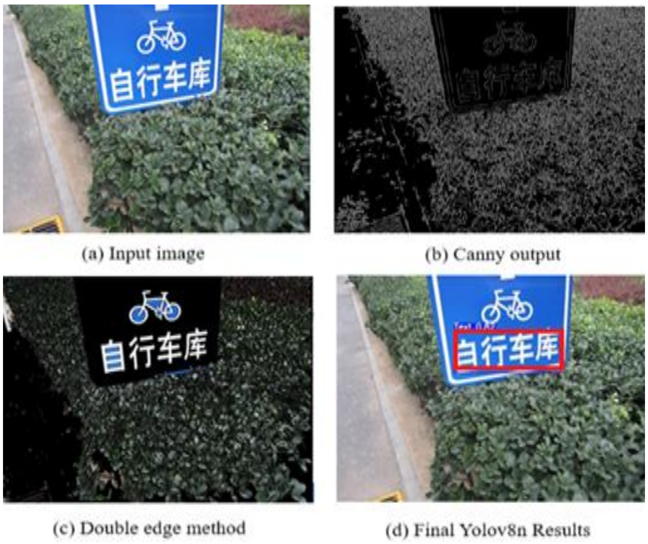


Figure 2. Block diagram of the proposed Double Yolo method



(a) Input image

(b) Canny output

(c) Double edge method

(d) Final Yolov8n Results

Figure 3. Input image, output of canny edge detection, double edge method and final Double Yolo approach

The double edge functionality block is depicted in proposed Double Yolo architecture diagram shown in Figure 2. Initially, the input samples are converted into grayscale to retain the edge information as shown in Figure 3(a). The canny edge detection algorithm is employed to generate the complete outer boundary of the text candidates and the outcome is depicted in Figure 3(b). The 2×2 kernel dilation is performed to connect the low resulted and disconnected edge information to generate double outer lines around the text regions. The double outer lines are visible as holes like structures. To extract complete region of text candidates, the hole like structures is filled with bright pixels. Finally subtract operation is performed between filled image (FI) and dilated image (DI) to obtain the potential text candidates (PTC) and mathematically represented using Eq. (1). Figure 3(c) and 3(d) show the output of double edge method and final yolo respectively.

$$\begin{gathered} PTC = DI - FI \\ PTC = (Bi \oplus Sk) - (M(\sim Bi)) \end{gathered} \tag{1}$$

where,

Bi - Binary image of input
Sk - Structured element
$\oplus$ - Dilate operation
M - Morphological reconstruction of binary image (Bi)
$\sim Bi$ - Complement of the binary image

3.2 Labelling for text candidates

LabelImg is a widely used graphical annotation tool for labelling texts in images, including text candidates for tasks like text detection and object recognition. It allows users to manually draw bounding boxes around areas of interest, such as text regions, and assign labels to these regions. The tool supports popular annotation formats like PascalVOC (XML), YOLO (TXT) and createML (JSON) formats making it compatible with many machine learning models and frameworks. Once ground truth values generated, used for training the machine learning models. With its simple interface, keyboard shortcuts, and support for normalized outputs, LabelImg makes the annotation process efficient, ensuring high-quality datasets for tasks like OCR and text detection tasks.

3.3 Double Yolov8n for text extraction

The proposed work flow of detection and extraction text using the Double edge method and YOLOv8 model, as shown Figure 1. Architecture of proposed Double Yolo8 model illustrated in Figure 2. Initial block contains double edge functionality which performs primary detection of textual components to separate from image background and non-text components. The double edge output pipelined, training and testing with YOLOv8n model with respective ground truth data. LabelImg tool used annotation. YOLOv8n part majorly consists of backbone, neck, and head components. The backbone is the core feature extraction component with a Cross Stage Partial (CSP) network and it has convolutional layers that extract spatial and semantic features from the input. Eq. (2) represents the backbone process of Double Yolo architecture. Backbone receives the output from double edge method (*X*), *H* refers the Height, *W* indicates the width with 3 to 1024 layers mathematically represented as Eq. (2).

$$\begin{gathered} x \in \mathbb{R}^{(H \times W \times 3)} \\ x \to \mathbb{R}^{\left(\frac{H}{32} \times \frac{W}{32} \times 1024\right)}, (k = 3, s = 2, p = 1) \end{gathered} \tag{2}$$

SiLU as a non-linear activation function used throughout the backbone, neck, and head to effectively learn features across the input. Further backbone output is pipelined to neck component, the neck is the feature aggregation component, that enhances the feature maps obtained by the backbone. Detection of small and large objects effectively done by Feature Pyramid Network (FPN) by combining low- semantic and high-sematic feature maps. Path Aggregation Network (PANet) is used to detect objects at various scales by creating rich feature representation using top-down and bottom-up manner, mathematical representation shown in Eq. (3).

$$x \to \mathbb{R}^{\left(\frac{H}{32} \times \frac{W}{32} \times 64\right)}, (k = 3, s = 1, p = 1) \tag{3}$$

The YOLOv8n has an anchor-free detection that directly predicts the center, width, and height of the bounding boxes. Finally, the head component receives the text proposal by the neck, it produces the final predictions in terms of bounding boxes, class probabilities, and confidence scores for detected prominent text blocks as shown in Figure 2. Fine-grained textual components are generated with bounding boxes around it with its confidence scores, mathematical representation of head component is shown in Eq. (4).

$$x \to \mathbb{R}^{\left(\frac{H}{32} \times \frac{W}{32} \times 255\right)}, (k = 1, s = 1, p = 0) \tag{4}$$

Proposed algorithm considers generated bounding boxes with confidence scores greater than 50 percent for further fine tune process of text recognition which produces the actual words else low confidence level bounding boxes are rejected as non-text component. Proposed algorithm is experimented on all benchmark datasets like CTW1500, MSRDA TD500, Total text, MRRC and MLe2e datasets and yielded far better results. Experimentation is done without applying the double edge for scene images also, but it's not produced very good results, detailed analysis shown in experimental section.

4. EXPERIMENTAL SETUP AND DISCUSSION

To evaluate the performance of the designed model Double Yolov8n (nano), Google Colab notebook platform with Nvidia T4 GPU and Python 3.10.12 torch 2.5.0+cu121 used for experimental setup. Paths for training and validation image dataset, paths for training and validation labels, model architecture details, class names and number of classes (text class in our case) are configured in Yaml file. The model has been trained for 100 images with different training and validation ratios with learning rate of 0.01 with 100 epochs. Initially input images applied with double edge method to detect the probable textual regions. Annotated using LabelImg tool to create the ground truth values. Then annotated images is further post processed with the YOLOv8n model to generate potential text regions, finally fine-grained textual components were generated with bounding boxes around it with confidence scores. Proposed algorithm considers those bounding boxes with the confidence scores greater than 50 percent as successful text selection else reject the bounding boxes as non-text component. Proposed technique produced better accuracy with very minimal number of samples trained.

4.1 Experimental analysis

Experimental setup for the proposed Double Yolo model is detailed in the above section. Input images with various background and complex scenarios are trained and validated candidly. The benchmark datasets like CTW1500, MSRA-TD500, Total text, MRRC and MLe2e were used for the experiment. Dataset images include multiple scripts, multiple orientation, text embossing, lighting effects, blurred, night vision, low contrast and text overlapping. Our aim is to reduce training and testing time and develop slimmer approach, hence designed this optimized model. We considered only 100 images for training and testing purpose. In primary stage,

double edge method separated the candidate text regions from scene image, later part YOLOv8n model refined the potential text components from non-text background by drawing bounding boxes with confidence scores around it. Proposed Double Yolo technique obtained significantly better results even with few samples trained and it is optimized approach to detect the text components from scene text images. With the proposed double edge technique, it has reduced the training time of large number of images which typically many deep learning models suffer. Time taken for the training and validation of the model is approximately 15 minutes. Time taken for testing unseen images about $2^{\pm}1$ msecs. The proposed model stands out with various parameters like light weight, simpler and faster approach proved far better model than other huge network models.

As part of comparison analysis, the images also tested with only YOLOv8n model without using double edge technique, it yielded poor results with lots of false negatives. The proposed Double Yolo method flourishes in a variety of ways, since it is simpler, lightweight and successful text detection strategy than heavy deep learning models such as EAST, Pixel link, ResNet Textsnake, and PAN. Proposed method efficiently detects true text regions and extracts text sequences from natural scene images, the performance is evaluated using precision (P) (Eq. (5)), recall (R) (Eq. (6)) and F- measure (F) (Eq. (7)). Three characteristics are utilized to determine the effectiveness of the text detection approach, True Detected Text Regions (TDR), Actually Text Regions Present (APTR) in the image, and False Identified Text Regions (FIR), which incorrectly identify non-text regions as text in the image. Following section contains the results of Precision (P) and Recall (R) for the datasets mentioned.

$$Precision(P) = \frac{TDR}{TDR + FIR} \tag{5}$$

$$Recall(R) = \frac{TDR}{APTR} \tag{6}$$

$$F - Measure(F) = \frac{2 * p * r}{P + R} \tag{7}$$

4.2 Experimentation with CTW1500 dataset

CTW1500 is bench mark publicly available to evaluate the computer vision tasks. This dataset contains 1500 images comprising of street views, shopping mall boards, including indoor or outdoor scenes. It is designed to evaluate the performance of scene text algorithms. Most preferably contain English and Chinese characters. Images are in JPEG. format with the resolution 720×1280 pixels. Data set contains images with different orientations, Color, style, complexity with variety of background. Experimentation results with the proposed Double Yolo algorithm is shown in Figure 4. It has achieved the accuracy of 93.10 percent and stands out to be simple technique than peer methods available with far better results. Other network models compared comes with heavy models and needed large number training and testing process. The proposed model thrives in terms simplicity yet effective model with very less training and testing samples and yielded higher efficiency and comparison of results shown in Table 1.

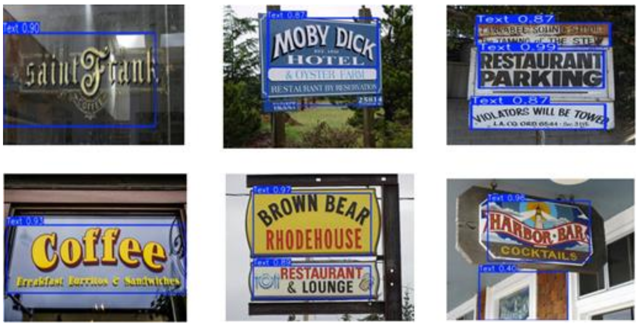


Figure 4. Results of Double Yolo on CTW1500 dataset

Table 1. Performance analysis of Double Yolo method on CTW1500 dataset

Methods	Recall	Precision	F-Measure
PAN [22]	81.2	86.4	83.7
Textsnake [23]	85.3	67.9	75.6
RFRN [24]	86.5	76.4	81.0
Diff Binarization [25]	82.8	87.9	85.3
Double Yolo	**94.04**	**92.26**	**93.14**

4.3 Experimentation with MSRA-TD500 dataset

MSRA-TD500 publicly available dataset is very distinguishable benchmark dataset for the evaluation of scene text detection algorithms. Dataset contains both indoor as well as outdoor images comprising of mostly with signs, door number plates, street views, shopping malls, ad boards and large hoardings taken from pocket friendly camera. Image resolutions may vary from 1296×864 to 1920×1280.

Many images come with trees and grasses as background and windows and brick patterns, which increases the complexity of the scene text images with more distortions. Dataset appears with multilingual and multi orientation with various Colors, styles and sizes, mostly in English and Chinese languages. Figure 5 shows the experimental results with the proposed Double Yolo method, it's evident that even though images have dark background still it is able to locate and detect the textual contents from the image background which is very promising, comparison of other methods shown in Table 2.

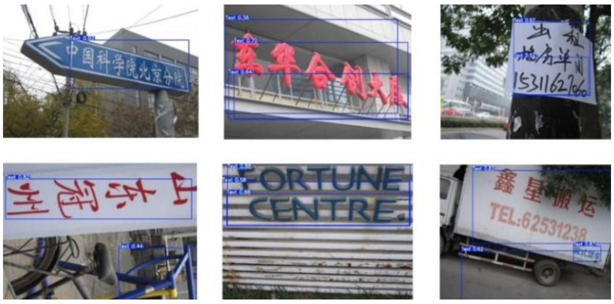


Figure 5. Results of Double Yolo on MSRA-TD500 dataset

Table 2. Performance analysis of Double Yolo method on MSRA-TD500 dataset

Methods	Recall	Precision	F-Measure
PAN [22]	83.8	84.4	84.1
Textsnake [23]	73.9	83.2	78.3
HOCC [4]	72	62	66
EAST [4]	87	67	76
Pixel Link [4]	83.0	73.2	77.8
CNN-Bitplane-DEdge [4]	71.23	49.87	58.67
Double Yolo	**92.46**	**88.98**	**90.69**

4.4 Experimentation with total text dataset

Total text dataset consists of total 1555 images, in which the training split is 1255 images and for the testing split 300 images. The images include scene text appears with horizontal, multi oriented and curved in urban environments with complex background. Experimental results for the proposed Double Yolo method and accuracy detecting text in on dataset images is shown in Figure 6. Proposed double edge and Yolo model has outsmarted with very good accuracy compare to other methods with least number training and testing, where as other models compared in Table 3 needed very high amount training and testing time and large number of samples.

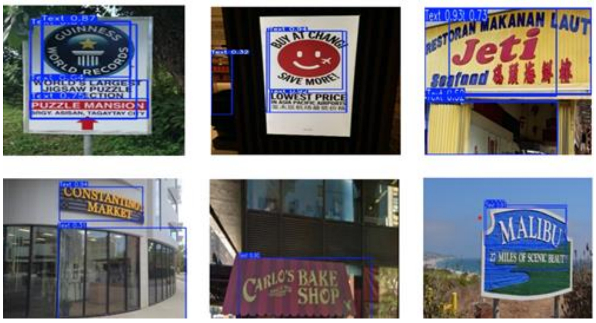


Figure 6. Results of Double Yolo on total text dataset

Table 3. Performance analysis of Double Yolo method on total text dataset

Methods	Recall	Precision	F-Measure
PAN-640 [22]	81	89.3	85
Textsnake [23]	74.5	82.7	78.4
Baseline CNN [4]	80.06	85.45	82.67
SDM Resnet-50 [4]	86.03	90.85	88.37
CNN-Bitplane-DEdge [4]	87.7	73.06	79.69
Double Yolo	**93.33**	**86.66**	**89.87**

4.5 Experimentation with MRRC dataset

This dataset created by IISC researchers for ICDAR 2013 competition, to find the best algorithm with highest accuracy. Dataset contains 4000 images which are captured in surroundings of Bangalore streets with various south Indian languages and variety of styles. The experimentation results of proposed Double Yolo method on MRRC dataset are shown in Figure 7. Other models compared and respective results also listed in Table 4. Tested images contain Urdu, Kannada and English with dull images, still it has yielded far better results. It proves that efficient detection and extraction done even with image variations and constraints and results have outperformed with south Indian textual images other models compared.

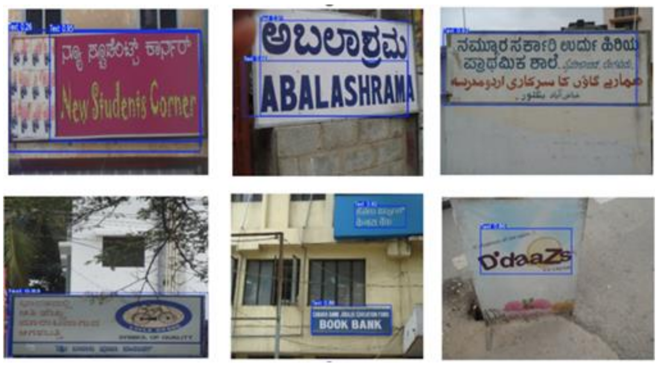


Figure 7. Results of Double Yolo on MRRC dataset

Table 4. Performance analysis of Double Yolo method on MRRC dataset

Methods	Recall	Precision	F-Measure
MRF [26]	71.37	64.92	67.99
Gomez method [4]	64	58	61
Yin method [4]	71	67	69
Basavaraju et al. [3]	81.81	69.52	75.16
CNN-Bitplane-DEdge [4]	87.27	73.62	79.87
Double Yolo	**96.85**	**84.38**	**90.19**

4.6 Experimentation with multilingual end-to-end dataset

The MLe2e dataset contains mixture of pictures compiled from existing scene text datasets, with the images and ground truth updated to ensure homogeneity. And it is used for evaluation tasks like text detection, text recognition and script identification.

The dataset consisting of 711 images with Latin, Chinese, Kannada and Hangal scripts with varieties of complex background. Figure 8 shows the results of experiment conducted on MLe2e dataset, and Table 5 shows results comparison with other approaches. Proposed Double Yolo method is significantly yielded far better results, thrived in terms of simplicity since minimal set of images used for training and testing, which is most distinguishable characteristics of the proposed methods. Double Yolo results have outperformed with in detecting south Indians multilingual and multi-oriented textual images.

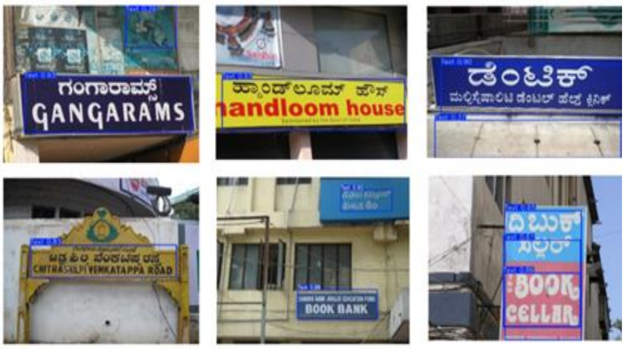


Figure 8. Results of Double Yolo on MLe2e dataset

Table 5. Performance analysis of Double Yolo method on MLe2e dataset

Methods	Recall	Precision	F-Measure
ECN [4]	51	62	56
MCM [27]	81	84	82
Enhanced Receptive Field [4]	80.0	86.0	83
Gradient Morphology [28]	88	90	89
CNN-Bitplane-DEdge [4]	94.57	85.33	89.71
Double Yolo	**99.12**	**87.96**	**93.21**

4.7 Experimentation without double edge structure method

Experimentation also conducted using only YOLOv8n model without using double edge method for same training and testing samples (same consideration 100 images) and for 100 epochs. Results found were very poor with lots of false negatives and it has even failed to locate the text regions in some cases, respective results shown in Figure 9. Observations found that using double edge technique along with YOLOv8n model, it is able locate the texts in scenic image with better efficacy otherwise it may not produce better results. In following section, we further analyse the impact with and without using double edge technique for all the benchmark datasets and corresponding results with evaluation matrices also discussed.

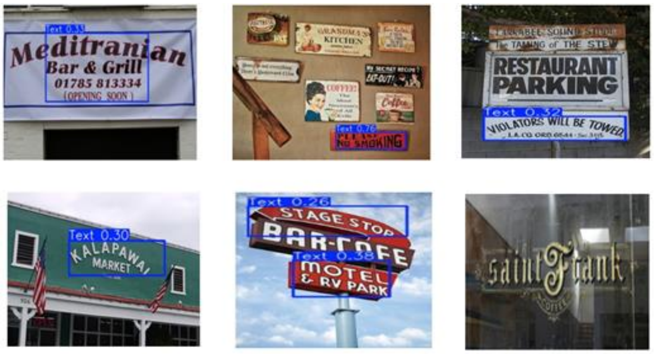


Figure 9. Results without double edge technique

4.8 Ablation study on significance of double edge technique in Double Yolo model

In this section, we explore major impact of the double stroke technique in the proposed Double Yolo model. The experimentation done on all the benchmark datasets by applying Double Yolo and without double edge technique and analysed the results to draw inference on the impact of the double edge method.

Fundamentally all text components are associated with complete and unform edge information. The double edge method looks for the text components have double line structure and has the same starting and ending points. If any component doesn't adhere this principle will be rejected as non-text component else it retains the component as text. Alone yolo could not able to segregate the edge-based components and non-edge-based components. Hence, the double edge technique is combined with yolo technique to accurately identify the actual text regions in the given scene images. Using Double Yolo method, time taken for the training and validation of the model is approximate 15 minutes. Time taken for testing unseen images about $2^{\pm}1$ msecs, which indeed good performance.

In the following section experiments conducted for all the datasets with 100 images, considering Double Yolo and without using double edge component and corresponding performance results are listed.

4.8.1 Experimentation with CTW1500 dataset

In the following Figure 10 and Table 6, we have shown sample outputs experimented on CTW1500 dataset. First row shows performance without DE component of model and second row shown results with double edge component. It's evident from the results that our proposed Double Yolo with double edge technique made the significant impact in identifying the textual components and we can observe recall rate significantly reduced.

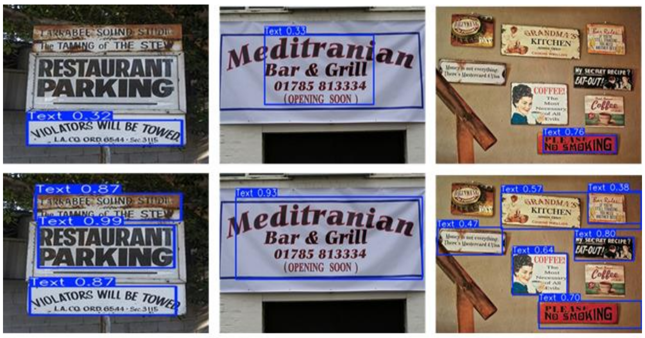


Figure 10. Sample outputs of Double Yolo on CTW1500 dataset, without DE and with DE component

Table 6. Performance analysis for CTW1500 dataset

Methods	Recall	Precision	F-Measure
Without DE	73.75	80.50	79.19
Double Yolo	**94.04**	**92.26**	**93.14**

4.8.2 Experimentation with MSRA-TD500 dataset

Figure 11 and Table 7 show the results of images of MSRA-TD500 dataset, first row contains results of without double edge component and second row shows results with double edge component (Double Yolo) model with 100 samples. It's found that with degraded images resulted in more false negatives and f-score is much reduced in the absence of DE component.

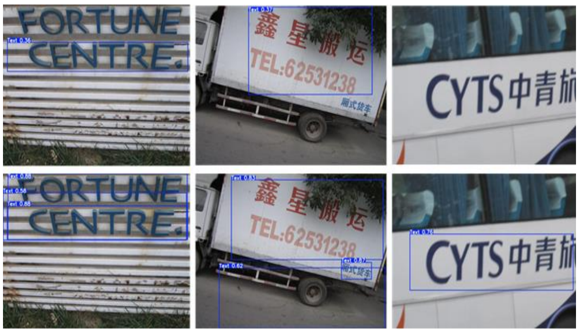


Figure 11. Sample outputs of MSRA-TD500 dataset for without DE and with DE component

Table 7. Performance analysis for MSRA-TD500 dataset

Methods	Recall	Precision	F-Measure
Without DE	70.33	81.48	77.19
Double Yolo	**92.46**	**88.98**	**90.69**

4.8.3 Experimentation with total text dataset

Figure 12 and Table 8, shown with results of sample outputs and performance comparison for without double edge component and Double Yolo method for the Total text dataset. We can observe that recall has reduced drastically, without double edge contribution, alone YOLOv8n model struggling the fetch text components from the scene images. But with Double Yolo (DE based) method has achieved good accuracy compare to without DE method.

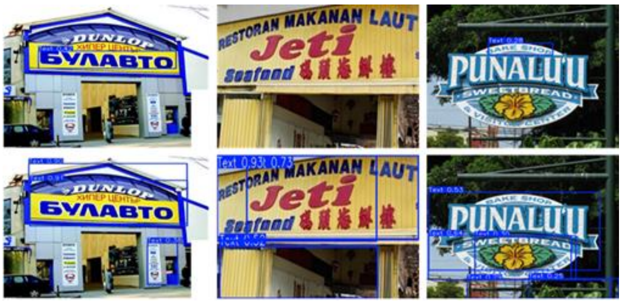


Figure 12. Sample outputs of total text dataset, without DE and with DE component

Table 8. Performance analysis for total text dataset

Methods	Recall	Precision	F-Measure
Without DE	71.42	86.95	78.43
Double Yolo	**93.33**	**86.66**	**89.87**

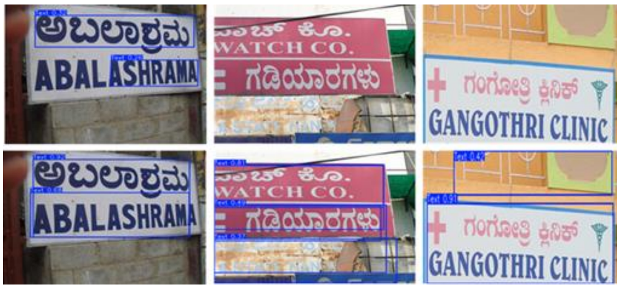


Figure 13. Sample outputs of MRRC dataset, Double Yolo-without DE and with DE component

Table 9. Performance analysis for MRRC dataset

Methods	Recall	Precision	F-Measure
Without DE	71.42	83.33	76.92
Double Yolo	**96.85**	**84.38**	**90.19**

4.8.4 Experimentation with MRRC dataset

MRRC dataset mostly contain images with south Indian scripts. Experimentation with MRRC dataset and sample results is shown in Table 9 and Figure 13. Here also, with more reduction in the recall can be seen. It ensures the effect of double edge component and f-score also seen low.

4.8.5 Experimentation with MLe2e dataset

MLe2e is derived dataset from existing scene text datasets. It contains scripts like English, Latin, Kannada and Hangal. Here we have used to evaluate the influence of double edge component in proposed model, the performance and sample out shown in Table 10 and Figure 14, respectively. More false negative proving that double edge method has impact in selection of textual candidate in scene images. The double edge technique and implementation is detailed in Section 3.2.

4.8.6 Performance comparison with Double Yolo Vs without DE

The comparison study of Double Yolo Vs Without DE component on all bench mark datasets done in previous section and same demonstrated with the curve graph as shown in Figure 15.

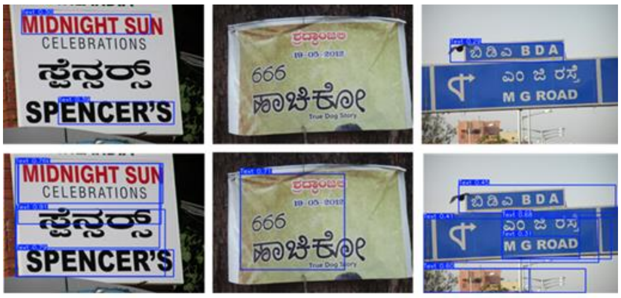


Figure 14. Sample outputs of on MLe2e dataset, without DE and With DE (Double Yolo) component

Table 10. Performance analysis for MLe2e dataset

Methods	Recall	Precision	F-Measure
Without DE	75	85.71	80
Double Yolo	**99.12**	**87.96**	**93.21**

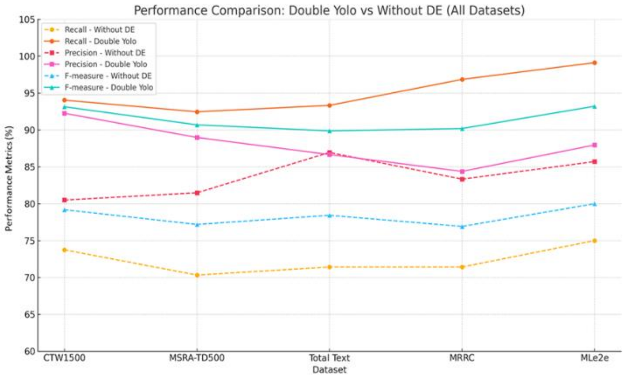


Figure 15. Performance comparison without DE component and Double Yolo on all benchmark datasets

Observations found that recall has reduced drastically in all cases, its proved alone YOLOv8n model not able to fetch text components successfully without double edge contribution.

4.9 Scope and limitations of Double Yolo method

The performance of Double Yolo method may limit when texts appear with backgrounds blended with text-like structures and overlapped object regions. This challenge may lead to false positives in text detection. Additionally, the model's performance may degrade when processing poor-quality images or images with dark shade, resulting in suboptimal results as shown in Figure 16.

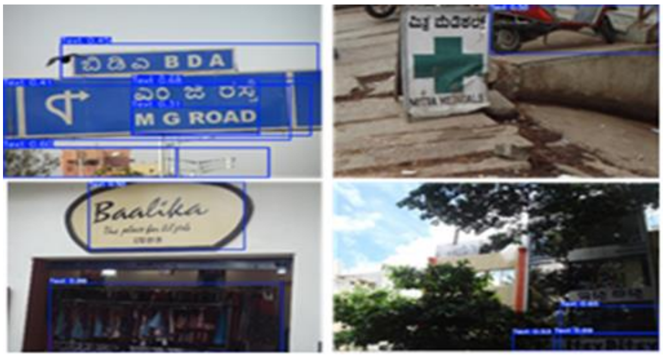


Figure 16. Poor performance of Double Yolo model with poor quality images or complex background

5. CONCLUSIONS

Scene text appearances in today's environment is very challenging due to complexities in the nature and text presence in assorted environment. Many approaches include machine learning and deep learning models were developed but have own drawbacks in detecting texts in complex scene images. Our intent is developing lighter and efficient method, hence proposed Double Yolo algorithm containing Double edge method and YOLOv8n (nano) model to detect the text from scenic images. Primarily Double edge method extracts the most probable text components in the given image which exhibits textual properties. In double edge concept, components which have edge points having starting and ending points are same, components treated as text else it is considered as non-text background. Further the output of double edge method fed in to YOLOv8n model to fine tune for fine grained text components. Finally, the model generates bounding boxes around the text components with confidence scores around it. Proposed model trained and tested with very few samples, still yielded very effective results due to optimized double edge technique. The approach found simpler, light weight and effective than deep learning models which need training and testing of large number of images. Proposed method may not perform well, when images with occluded with textual like structures and poor-quality images, hence this limitation can be considered for future works to achieve the almost accuracy. Proposed Double Yolo method with its simplistic approach can be considered as suitable model and good choice for scene text detection for mobile applications and other resource constraint environments.

REFERENCES

[1] Basavaraju, H.T., Manjunath Aradhya, V.N., Guru, D.S. (2020). Neighborhood pixel-based approach for arbitrary-oriented multilingual text localization. In Intelligent Systems, Technologies and Applications: Proceedings of Fifth ISTA 2019, India pp. 1-12. https://doi.org/10.1007/978-981-15-3914-5_1

[2] Manjunath Aradhya, V.N., Basavaraju, H.T., Guru, D.S. (2021). Decade research on text detection in images/videos: A review. Evolutionary Intelligence, 14(2): 405-431. https://doi.org/10.1007/s12065-019-00248-z

[3] Basavaraju, H.T., VN, M.A., Guru, D.S., Harish, H.B.S. (2018). LoG and structural based arbitrary oriented multilingual text detection in images/video. International Journal of Natural Computing Research, 7(3): 1-16. https://doi.org/10.4018/IJNCR.2018070101

[4] Mahadevappa, M., Aradhya, V.N., Basavaraju, H.T., Shivarudraswamy, S. (2024). CNN-DEdge: Multilingual scene text detection and extraction. Mathematical Modelling of Engineering Problems, 11(11): 3152-3160. https://doi.org/10.18280/mmep.111125

[5] Liu, F., Chen, C., Gu, D., Zheng, J. (2019). FTPN: Scene text detection with feature pyramid based text proposal network. IEEE Access, 7: 44219-44228. https://doi.org/10.1109/ACCESS.2019.2908933

[6] Wang, Q., Huang, Y., Jia, W., He, X., Blumenstein, M., Lyu, S., Lu, Y. (2020). FACLSTM: ConvLSTM with focused attention for scene text recognition. Science China Information Sciences, 63: 120103. https://doi.org/10.1007/s11432-019-2713-1

[7] Soora, N.R., Kotte, V.K., Dorthi, K., Vodithala, S., Kumar, N.C. (2024). A comprehensive literature review of vehicle license plate detection methods. Traitement du Signal, 41(3): 1129-1141. https://doi.org/10.18280/ts.410304

[8] Aluri, M., Tatavarthi, U.D. (2024). Geometric deep learning for enhancing irregular scene text detection. Revue d'Intelligence Artificielle, 38(1): 115-125. https://doi.org/10.18280/ria.380112

[9] Liu, Z., Lin, G., Goh, W.L. (2020). Bottom-up scene text detection with Markov clustering networks. International Journal of Computer Vision, 128(6): 1786-1809. https://doi.org/10.1007/s11263-020-01298-y

[10] Basavaraju, H.T., Aradhya, V.M., Guru, D.S. (2019). Text detection through hidden Markov random field and EM-algorithm. In Information Systems Design and Intelligent Applications, pp. 19-29. https://doi.org/10.1007/978-981-13-3329-3_3

[11] Pal, U., Halder, A., Shivakumara, P., Blumenstein, M. (2024). A comprehensive review on text detection and recognition in scene images. Artificial Intelligence and Applications, 2(4), 229-249. https://doi.org/10.47852/bonviewAIA42022755

[12] Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X. (2023). Turning a clip model into a scene text detector. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 6978-6988. https://doi.org/10.1109/CVPR52729.2023.00674

[13] Zhong, D., Lyu, S., Shivakumara, P., Pal, U., Lu, Y. (2022). Text proposals with location-awareness-attention network for arbitrarily shaped scene text detection and

recognition. Expert Systems with Applications, 205: 117564. https://doi.org/10.1016/j.eswa.2022.117564

[14] Cai, Y., Liu, Y., Shen, C., Jin, L., Li, Y., Ergu, D. (2022). Arbitrarily shaped scene text detection with dynamic convolution. Pattern Recognition, 127: 108608. https://doi.org/10.1016/j.patcog.2022.108608

[15] Keserwani, P., Dhankhar, A., Saini, R., Roy, P.P. (2021). Quadbox: Quadrilateral bounding box based scene text detection using vector regression. IEEE Access, 9: 36802-36818. https://doi.org/10.1109/ACCESS.2021.3063030

[16] Wang, W., Xie, E., Li, X., Liu, X., et al. (2021). Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 5349-5367. https://doi.org/10.1109/TPAMI.2021.3077555

[17] Wang, X., Zheng, S., Zhang, C., Li, R., Gui, L. (2021). R-YOLO: A real-time text detector for natural scenes with arbitrary rotation. Sensors, 21(3): 888. https://doi.org/10.3390/s21030888

[18] Guo, Z., Fang, P., Li, H., Wang, Z., Gao, W. (2022). Location-aware feature selection network for multi-oriented scene text detection. In 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, pp. 1-6. https://doi.org/10.1109/ICME52920.2022.9860011

[19] Banerjee, A., Shivakumara, P., Acharya, P., Pal, U., Canet, J.L. (2022). TWD: A new deep E2E model for text watermark/caption and scene text detection in video. In 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, pp. 1492-1498. https://doi.org/10.1109/ICPR56361.2022.9956279

[20] He, Y., Chen, C., Zhang, J., Liu, J., He, F., Wang, C., Du, B. (2022). Visual semantics allow for textual reasoning better in scene text recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1): 888-896. https://doi.org/10.1609/aaai.v36i1.19971

[21] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 4714-4722. https://doi.org/10.1109/ICCV.2019.00481

[22] Wang, W., Xie, E., Song, X., Zang, Y., et al. (2019). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 8439-8448. https://doi.org/10.1109/ICCV.2019.00853

[23] Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In ECCV 2018: 15th European Conference, Munich, Germany, pp. 19-35. https://doi.org/10.1007/978-3-030-01216-8_2

[24] Deng, G., Ming, Y., Xue, J.H. (2021). RFRN: A recurrent feature refinement network for accurate and efficient scene text detection. Neurocomputing, 453: 465-481. https://doi.org/10.1016/j.neucom.2020.10.099

[25] Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X. (2020). Real-time scene text detection with differentiable binarization. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 11474-11481. https://doi.org/10.1609/aaai.v34i07.6812

[26] Mahesha, M., Manjunath Aradhya, V.N., Basavaraju, H.T., Siddesha, S. (2023). MRFScene: Multi-lingual multi-oriented scene text detection using Markov random fields. In International Conference on Computational Intelligence, Surat, India, pp. 439-449. https://doi.org/10.1007/978-981-97-3526-6_34

[27] Khalil, A., Jarrah, M., Al-Ayyoub, M., Jararweh, Y. (2021). Text detection and script identification in natural scene images using deep learning. Computers & Electrical Engineering, 91: 107043. https://doi.org/10.1016/j.compeleceng.2021.107043

[28] Xiao, S., Peng, L., Yan, R., An, K., Yao, G., Min, J. (2020). Sequential deformation for accurate scene text detection. In European Conference on Computer Vision, 16th European Conference, Glasgow, UK, pp. 108-124. https://doi.org/10.1007/978-3-030-58526- 6_7