Curved Text Detection in Scenic Images via Proposal-Free Panoptic Segmentation and Deep Learning

Prachi Chhabra
Department of Computer Science
Sharda University, Greater Noida
prachichhabra@jssaten.ac.in

Ali Abidi
Department of Computer Science
Sharda University, Greater Noida
ali.abidi@sharda.ac.in

Abstract: *Curved texts pose a significant challenge in detection in 'the wild', primarily due to the inherent variabilities in text orientation and possible distortions while the images are being acquired. Standard text detection models have been observed to exhibit low accuracy in detecting texts on curved surfaces. To fill this gap, there have been a variety of deep learning-based models proposed to date which have achieved low to moderate success. This paper implements a DL-based model for the detection and recognition of text from scenic images. The proposed approach applies different image processing techniques such as Gray scale conversion, noise removal using median filter, normalization and Otsu's Binarization and a panoptic segmentation technique for achieving desired text detection performance. A synthetic dataset is created which is used to fill in the gaps of character annotation and multi-orientation. The performance of the proposed approach is determined using different evaluation metrics and the results are compared against existing techniques such as You Only Look Once version 5 (YOLOv5), HDBNet, Bidirectional Perspective Network (BiP-Net), and Res18-LVT. Results show that the proposed approach achieves better performance in terms of precision (98.7%), recall (91.7%), and F1 score (94.5%) as compared to existing classification models*.

Keywords: *Deep learning, image segmentation, natural scene images, multi-oriented text, curved text detection.*

Received June 16, 2024; accepted September 2, 2024
https://doi.org/10.34028/iajit/21/5/10

1. Introduction

Text detection from scenic images has been one of the most utilized areas of research in information retrieval system. Researchers working primarily in information retrieval domain use text detection as a potential tool in various applications such as sign recognition [22], vehicle identification [26], navigation [22], etc., Several techniques have been developed over time for extracting text from images [4, 33]. However, the demand for a more effective technique still persists. Images containing text have been historically classified into broad groups of either natural scenic images or images of document. Document images or images containing text of documentary nature are not considered as 'in the wild' or scenic images as they are mostly acquired with least interference from environmental challenges. Although, text present in scenic images in various instances might exhibit important information such as traffic signs or signboard and might be difficult to interpret. Hence, it is crucial to find legible text from such scenic images for interpretation and understanding. Unlike text extraction from images of document, it is challenging to recognize text from scenic images [33]. Whereas it is easy to recognize text information from the images but factors such as cluttered background, similarity between texts and overlapping occlusions occur more often increases the difficulty of the text recognition process [28, 29]. These factors often make it challenging to distinguish between the textual and non-textual regions in the scenic images. In addition, issues such as text alignment, poor visibility of text, multi-lingual text, improper orientation, and blurring, skewed, variations in font types and sizes in scenic images makes text detection and recognition a complex task. The emergence of artificial intelligence techniques has motivated the researchers to implement Deep Learning (DL) for text detection and recognition [35, 39]. These techniques perform text detection and text recognition simultaneously and this attribute makes them a popular candidate for this process. In the text detection process, the text instances in images are localized, and text recognition helps to obtain legible characters from cropped images containing text. Different existing techniques such as You Only Look Once version 5 (YOLOv5) [23], HDBNet [34], Bidirectional Perspective Network (BiP-Net) [33], and Res18-LVT [11] were used previously for text detection. The YoloV5 model is one of the advanced version of traditional Yolo architectures. YOLOv5 generates features from input images, which are then processed by a prediction system to draw bounding boxes around objects and classify them. During each training batch, YOLOv5 uses a data loader that

performs online data augmentation on the training data. In the HDBNet detection model, the image is first processed through the ResNet backbone network with a feature pyramid. This feature pyramid performs up-sampling from top to bottom and concatenates the up-sampled features with features of the same size to produce feature maps. These maps are then used to predict the probability map and the threshold map. The approximate binary map is calculated using these two. The probability map indicates the likelihood of a pixel being text, while the threshold map determines whether each pixel is text. Each pixel is adaptively binarized based on these maps. In the BiP-Net model, firstly, the input RAW burst is processed through the edge boosting feature alignment module to extract features, reduce noise, and correct spatial and color misalignment among the burst features. Second, a pseudo-burst is created by exchanging information so that each feature map in the pseudo-burst incorporates complementary properties from all the actual burst image features. Finally, the multi-frame pseudo-burst features are processed using the adaptive group up sampling module to generate the final high-quality image. Lastly, the Res18-LVT network basically consists of a ResNet-18 architecture with a Long Former Vision Transformer (LVT) for enhanced image processing or analysis tasks.

Despite this prevalence of related methodologies, there is a lack of availability of effective techniques which can recognize multi-oriented and curved texts from input images [15]. The work presented Yim *et al*. [39] exhibits better performance in terms of recognizing arbitrary multi-oriented texts. However, the performance of the approach deteriorates when applied for detecting curved texts due to large character spacing and limited text representation. These challenges motivate this research to implement a proposed model for recognizing curved text from scenic images.

The individuality of the planned method deceits in the implementation of the panoptic segmentation through DL for text detection. This approach combines both semantic and instance segmentation seamlessly within a single framework. Panoptic segmentation provides a holistic understanding of visual scenes by delineating both static elements and things in an image. While DL model enhance the performance of segmentation [15]. Together, they offer a comprehensive solution that performs better than traditional object detection methods by capturing richer scene semantics and individual object details simultaneously, thus attractive the accurateness and interpretability of recognition consequences.

The following is an overview of this paper's primary contributions:

- This paper proposed a DL based recognition algorithm to detect and recognize multi-oriented and curved text from scenic images.
- To address the gaps in multi-orientation and feature commenting, an artificial data set is being developed.
- The objective of this study is to improve text recognition efficiency by tackling issues including image blur, inappropriate lighting, alignment variance, and size variability.
- The recommended model's effectiveness has been verified through a comparison with the current model using various evaluation measures, including f1 score, accuracy, and recollection.

The remainder of the document is organized as follows: A brief overview of current methods for text detection and recognition is given in section 2. The primary study technique that outlines the recommended approach's process is covered in section 3. It provides a quick explanation of how the recommended text recognition model operates. The simulation results and performance evaluation are covered in section 4. The paper's findings from experiments and future scope are presented in section 5.

2. Literature Review

A significant amount of research work has been published in recent times to accurately detect and recognize from large scale image databases [29, 30, 31, 32, 33, 34]. Extracting information from scenic images has been a challenging task since it requires deep feature extraction. Various research works have provided solution to this problem [10, 21, 40]. The studies presented in Gilal *et al*. [9], Lin *et al*. [20], Yang *et al*. [36], Zhang *et al*. [42], perform end-to-end text recognition using two phases namely text detection and recognition. These techniques segment and create bounding boxes around the text in scenic images. As discussed previously, identification and localization of text areas in complex backgrounds has a complex task and two main methods are adopted to overcome this problem. The first method has been based on the character region, which has been discussed in Bhatt *et al*. [3] and Manjari *et al*. [24]. These methods localize the character regions by connecting the individual components and then the localized characters are integrated to form a word. The second method uses sliding windows which is discussed in Detectron2.com [7], Yu *et al*. [37]. These methods employ a sliding window for identifying the textural areas within the image and recognize the text using machine learning models. Conventional techniques use manual feature extraction methods wherein the text has been detected from scenic images manually. These methods use features such as maximally stable extremal regions (MSERs) [5], features such as Histogram of Oriented Gradients (HOG) [24], Stroke Width Transforms (SWT) [25] for finding textual regions and generating output for text recognition. In addition to these, machine learning classifiers such as Support Vector Machines (SVM) and other Machine Learning (ML) classifiers

[27], K-Nearest Neighbors (KNN) classifiers [11]. The drawbacks of these techniques are that they require expert knowledge to achieve high performance. Majority of the previously existing techniques use manually hand-crafted features for text recognition. These techniques have been replaced with sophisticated DL techniques. Yang *et al*. [38] focused on recognizing text from objects using Deep Neural Networks (DNN). Wang *et al*. [32] implemented a DL based Rotational You Only Look Once (R-YOLO) for designing a text detection framework to detect arbitrary text from natural scenes. The obtained textural regions are given as input to the R-YOLO model for recognizing texts. The recent advancements in text recognition using DL are discussed in Yang *et al*. [35]. This review emphasizes end-to-end text recognition using different DL algorithms such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). It can be inferred from existing studies that, most of the techniques focus on recognizing texts based on single orientation (horizontal based) [8] and the demand for detecting multi-oriented and curved texts is still persistent.

3. Proposed Methodology

The preliminary aim of this research is to accurately detect and recognize multi-oriented and curved texts from scenic images. A DL based proposed model has been implemented in this research for recognizing texts. The proposed algorithm has an advanced library which is derived from Facebook AI Research (FAIR). In this research, proposed model performs instance segmentation of text obtained from scenic images. The block diagram of the proposed approach has been illustrated in Figure 1. A brief overview of the stages involved in the implementation of the proposed methodology has been discussed in the sub sections below:

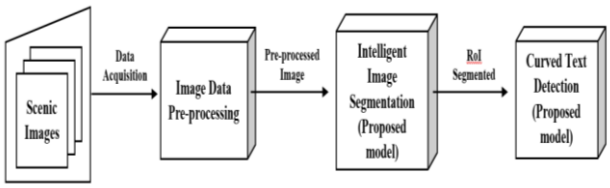

Figure 1. Block diagram of the curved text detection model.

3.1. Data Acquisition

A new synthetic dataset was created for experimental analysis. The dataset consisted of overall 1061 images of 14 different scenic backgrounds. The train-test split was such that out of 1061 images, 200 images were used for testing the model and remaining for training purposes. The collected input images had different dimensions. However, in this research the size of the images has been limited to a minimum value of 400 and maximum value of 4000. For image annotation, a 'label me' tool has been used. Image annotation refers to the process of labeling the images to highlight the characteristics of the data. In this research, the images are labeled automatically and the labeled image has been shown in Figure 2.

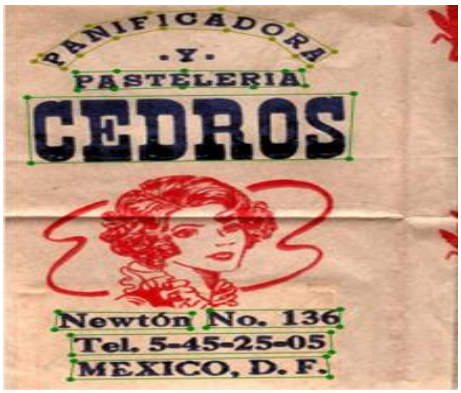

Figure 2. Annotated image.

3.2. Data Preprocessing

The scenic images fed into the model are pre-processed to increase the efficiency of the proposed model and reduce the complexity of the text detection and recognition process. Uncertainties such as external noise, missing values, redundant data are filtered out to achieve better recognition performance. In addition, the images are preprocessed to overcome possible challenges such as distortions, varying sizes, and missing image information. In this research, preprocessing has been performed using different stages such as; Gray scale conversion, noise removal using median filter, normalization and Otsu's Binarization.

- *Gray scale conversion*: in this step, the colored input images are converted into the grayscale image. The conversion has been performed to extract descriptors from small image areas instead of operating entirely on the Gray scale image. This process simplifies the recognition process, reduces the dimension and complexity.
- *Normalization*: this research performs normalization to logically group the image information within the same range (usually the range is between 0 and 1). In other words, to normalize the values of all images used for the recognition process.
- *Skew Correction*: when the images are scanned there is a possibility that the scanned or captured image might be slightly skewed. Skew correction has been performed in this research to determine the skewness in the image and to correct it [41] This is done to achieve better performance in terms of text recognition.
- *Image Scaling*: for effective text recognition, the size of the image should be greater than 300 Pixel Per Inch (PPI). The size of the images lesser than 300 PPI are rescaled to increase the size. As mentioned before, the size of the images are limited to a

minimum value of 400 and maximum value of 4000 scale invariant.

- *OTSU's Binarization*: in this research, binarization has been performed using OTSU's thresholding method. The OTSUs threshold method employs a linear discriminant criterion, wherein a single image has been considered with text visible in both foreground and background. Here, the variability and stratification of the background has not considered for the analysis and a suitable threshold value has been set for reducing the overlapping conditions of the class distributions. In this research, the threshold value has been set to 0.7 and 0.5.
- *Noise removal using median filter*: a median filter has been used in this research since it performs better in comparison to the mean filter without losing any important data [3]. In this research, the median filter reduces the magnitude of intensity variation between the pixels in the input image.

3.3. Image Segmentation and Curved Text Detection Using Proposed Model

In order to obtain effective text recognition, the pixels in the images are clustered wherein the pixels are grouped into different regions of coherent colors. This process is known as image segmentation. In this process, the input images are divided into multiple segments to simplify the analysis. In this case, segmentation has been performed to cluster images based on text and non-text pixels. However, for text recognition, the annotations will differ and each pixel will be segmented as one of the available characters. In this research, the image segmentation has been performed using the proposed object detection model. The proposed model can overcome the limitations associated with other object detection models such as high computational complexity. This model belongs to the FAIR library [4, 6] which has been extensively used in object detection and image segmentation tasks. The proposed model has been advanced version of the Detectron and the MaskRCNN model [19] which provides better results for both tasks. This model incorporates a Faster R-CNN [18], Mask R-CNN [21], RetinaNet, and other widely used detection methods are all incorporated into this model. Related to semantic instance, and panoptic segmented are the three primary segmentation techniques that it carries out.This research employs a panoptic segmentation model which creates a unifying path between semantic and instance segmentation [14]. For implementing proposed model in a more effective manner, a JavaScript Object Notation (JSON) based image file system has been used wherein the images to be labeled are considered. The file consists of three parts namely:

1. Images.
2. Categories.
3. Annotations.

The information has been extracted from the binary masks which represent the actual image on black color and marked images on white color as shown in Figure 3.

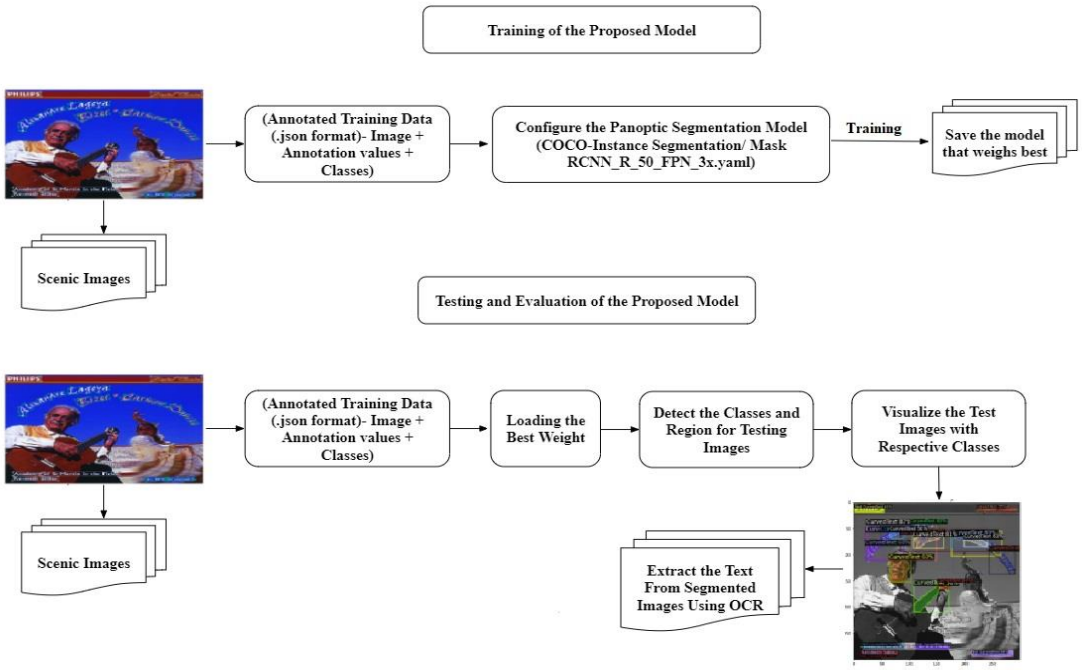

Figure 3. Training and testing of the proposed model for panoptic segmentation.

In the first part, each image file has been connected with an identifier and the dimensions of each image has been determined. The categories section in the file system allows registration of each image to the classes that are grouped into a specific category. Lastly, annotations used an object identifier to identify whether

the image or group of images belongs to the correct category or not. Further, it identifies the sequence of the coordinate pairs which restrict the region it occupies and also identifies the coordinates of the bounding boxes.

For Annotation, after gathering enough images, the images are labelled using a "labelme" tool. For panoptic segmentation: the algorithm begins by taking an image (X) and the number of classes in the dataset as input. It then initializes the model components: loading a pre-trained ResNet50 backbone network, initializing Semantic and Instance Segmentation Heads, and defining the Panoptic Segmentation Fusion Layer. Additionally, it sets up the loss function (Cross Entropy Loss) and optimizer (Adam) for training. During the forward pass, the input image goes through the backbone network to extract features. These features are used to generate semantic and instance segmentation predictions through their respective heads. The predictions are fused together using the Panoptic Segmentation Fusion Layer to produce the final panoptic segmentation predictions. The model is then trained over multiple epochs in a loop. Each epoch involves iterating through training data batches, computing the loss, and updating the model parameters using gradient descent. Throughout training, the forward pass is performed to obtain panoptic segmentation predictions, and the loss is computed to measure the model's performance against ground truth targets. Finally, the gradients of the loss with respect to the model parameters are computed during the backward pass, and the model weights are updated using the optimizer's update rule. This process repeats until the model converges to optimal performance. The architecture of the proposed object detection model and the workflow used in this implementation has been illustrated in Figure 4.

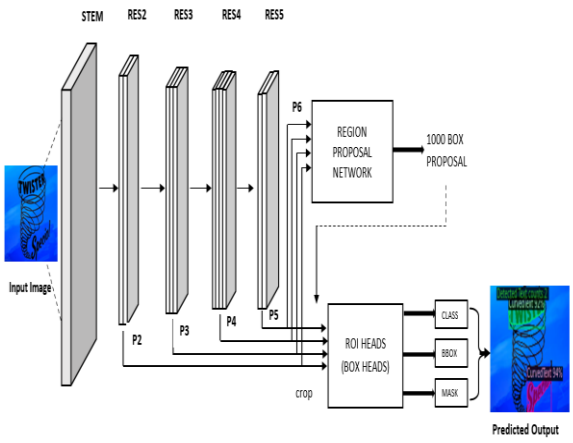

Figure 4. Architecture of the proposed model.

In this proposed model incorporates a Faster Regional Convolutional Neural Network (R-CNN) with a Feature Pyramid Network (FPN) [18] for detecting the bounding boxes put around the text. With base layer as R-CNN and FPN, the model has been extended to a mask R-CNN model which generates the segmentation mask in proposed model. As a result, as illustrated in Figures 3 And 4, the suggested architecture features a two-layered system with three key components: a backbone of the internet, a Region Proposal Network (RPN), and a Region Of Interest (ROI) head (box head). The RPN uses the numerous scales map of characteristics to identify text sections, and by standard, it produces 1000 recommended boxes along with an accuracy score. Additionally, the box head uses an entirely interconnected layer of the ResNet model to classify the identified images, recognize the precise positioning of the box, and crop each feature map into several features of different sizes. ResNet layers are used to extract map features from the submitted scenic images, with FPN serving as the network's backbone. A base block containing four bottleneck blocks (res 2, res 3, res 4, and res 5) is incorporated into the ResNet model. The input image has been down-sampled using a max pooling layer with the same duration as the 77 convolution layers with a stride of 2 in the stem block. The dimensions of the stem block's output mapping of features are 64×H/4xW/4, where H and W stand for the input image's dimensions in pixels, correspondingly. The FPN network is made up of four production characteristic maps from the ResNet model's bottle Netblocks: lateral, output convolutional, and output. Each later and convolution layer takes the output features from the bottleneck blocks and converts them to 256-channel feature maps using different channel numbers (256, 512, 1024, and 2048). The output of the res 4 initiates the forward operation of the FPN, and the same channel numbers are employed in a 3x3 output convolution layer. The obtained output feature map has been termed as P4 and the output of res 4 is given as input to the up sampler and has been added with the output of previous block (res 3) using a lateral convolution. The resultant feature map has been considered as the output and being termed as P3. The process was repeated twice and the obtained feature maps are termed as P2 and P1. A down-sampled feature representation of the res 4 final was used to create the final P5 result map of characteristics. In line with this, the ROI head block is made up of two distinct heads: the face head and the box head. The ROI pooling procedure is used to feed the images of the box into the box head. Bounding box detection and score prediction have been achieved by the utilization of the box head's final output. Conversely, the mask head receives the outcomes from the box head along with the four output feature maps from the FPN layer as input. Images in the box head with a few prediction maps make up the final output image of the proposed model: one for the class (image-level classification), one for the bounding box (localization), and one for the segmented mask (pixel-level classification). The resulting prediction maps are used as the segmentation mask of the output image. The curved texts are detected based on three criteria's:

1. If each pixel in the input scenic image corresponds to the text.
2. If the pixels are located within the text area.
3. If the input image pixel specifies the orientation of the text pixel. The performance of the proposed approach has been discussed in the below section.

4. Results and Discussion

This segment discusses the performance evaluation of the future object detection model used for the recognition and appreciation of curved text from scenic images. The model has been trained using a similar training process used in the object detection processes. The only difference is that this research uses a panoptic segmentation process instead of a traditional object detection process. All parameters used as user-defined. The parameters used for training the future model are tabulated in Table 1.

Table 1. Parameters used for training proposed model.

Parameters	Value
Learning rate	0.00025
Batch size	128
Image per batch	2
Number of iterations	700
Number of classes	3

4.1. Performance Evaluation of the Proposed Model

The performance of the proposed model has been evaluated using different performance metrics such as precision, recall, and F1 score. These metrics are represented using a confusion matrix whose elements are defined using four elements namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as shown in Figure 5.

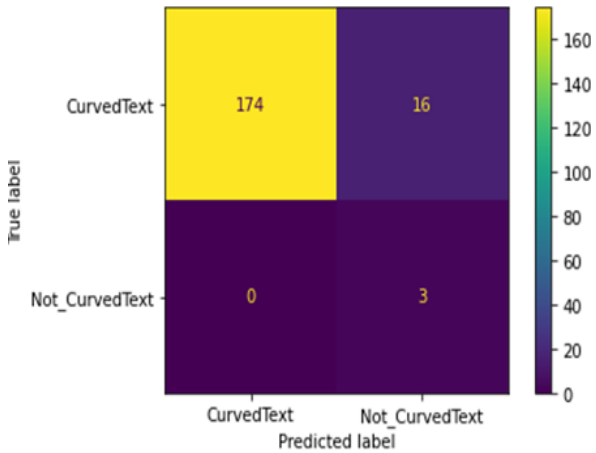

Figure 5. Confusion matrix.

It can be inferred from the confusion matrix that the number of TP, TN, FP, and FN are 174, 3, 0, and 16 respectively. Results show that there are no false positives detected by the proposed model. The model has been trained using both training and testing images with appropriate batch size. The training loss and validation loss models that we created for our research algorithm as graphs. After training this model, the total loss of proposed model with respect to the number of iterations has been computed as illustrated in Figure 6.

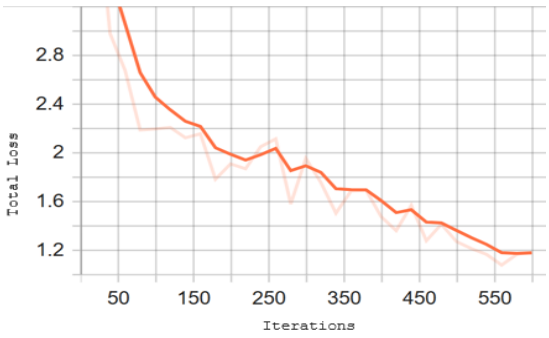

Figure 6. Total loss of the proposed model.

The loss of the proposed model has been determined for more than 500 iterations. It can be inferred from Figure 7 that the loss decreases with the increase in the number of iterations. This shows that the loss attains stability after certain iterations. The performance of the proposed model has been determined for two threshold values namely 0.7 and 0.5. The results of the same are discussed in below Figures 8.

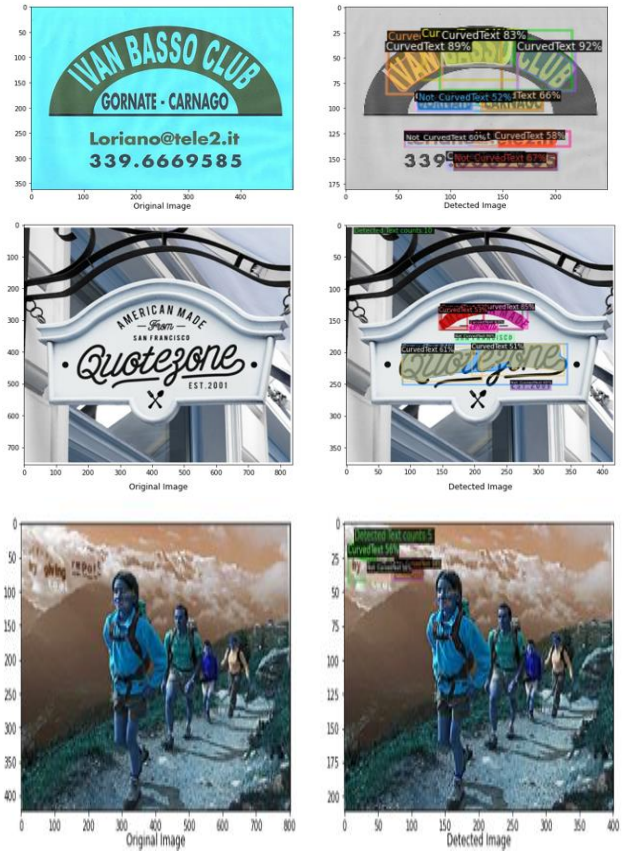

Figure 7. Results of proposed model for a threshold value of 0.7.

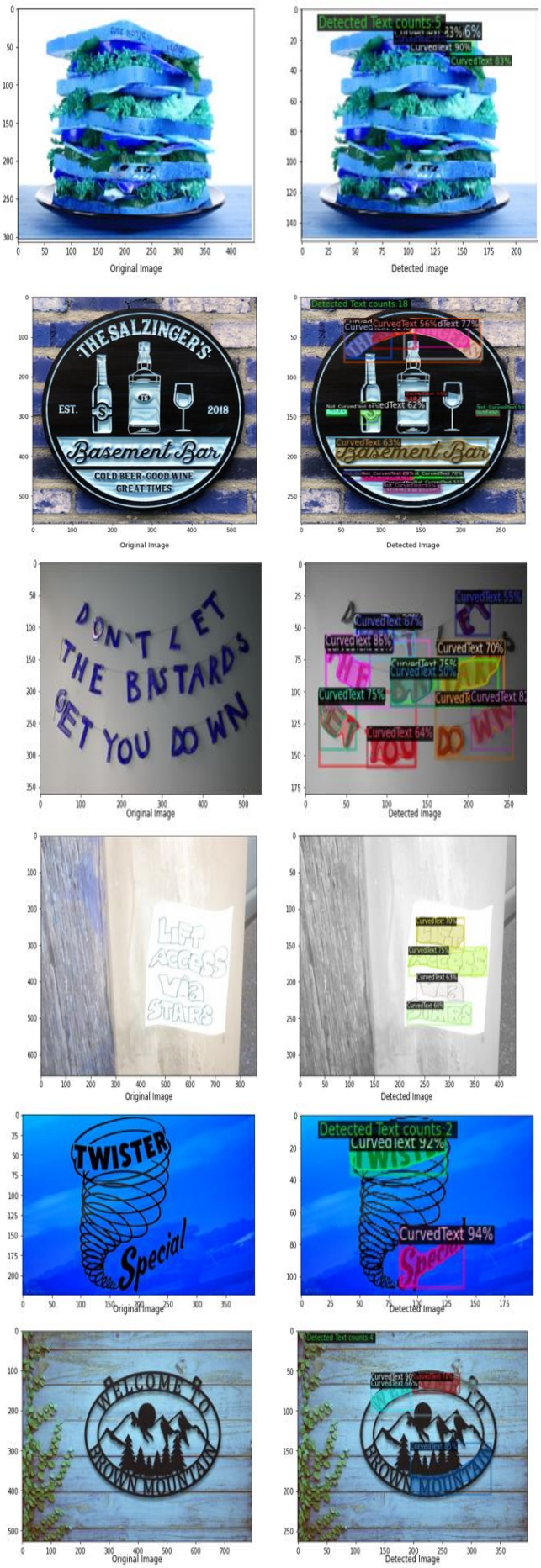

Figure 8. Results of proposed model for a threshold value of 0.5.

4.2. Comparative Analysis

The performance of the proposed approach has been validated with respect to different performance metrics and has been compared with existing models such as YOLOv5 [23], HDBNet [31], BiP-Net [33], and Res18-LVT [11]. Different metrics such as precision, recall, and F1 score, are used to evaluate the performance. The mathematical expressions for determining the performance metrics are defined as [38]:

The percentage of TFs found among all pertinent ground facts is known as recall. The definition of the recall is:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

The amount of accurate positive forecasts is known as precision. The level of precision is established.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

The results of the comparative analysis are tabulated in Table 2.

Table 2. Results of the comparative analysis.

Metrics	Proposed model	YOLOv5	HDB Net	BiP-Net	Res18-LVT
Precision	0.99	0.90	0.94	0.87	0.90
Recall	0.92	0.85	0.79	0.82	0.84
F1 score	0.95	0.88	0.86	0.85	0.87

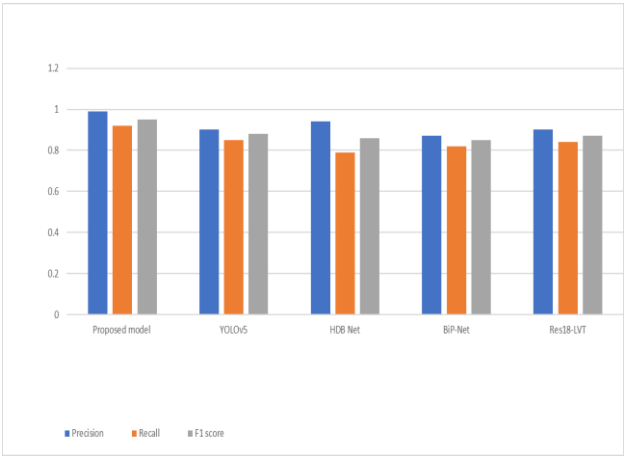

Figure 9. Comparative analysis.

The results indicate that, based on various evaluation measures, the recommended approach performs better than the current classification models. The recommended model's performance is entirely dependent on its validation and training results. With a precision of 98.7%, recall of 91.7%, and an F1 score of 94.5%, the recommended approach performed admirably. As compared to other methods, the BiP-Net model achieved lowest performance with a precision of 87%. Results have been validated that the proposed model can be used for real-time detection and recognition of curved texts from scenic images.

5. Conclusions and Future Scope

An efficient curved text detection and recognition approach using DL-based proposed model has been presented in this paper. A new dataset has been created that consists of different scenic images which are used to trained the proposed model for text recognition. The proposed approach employed image preprocessing and a panoptic segmentation technique for accurate text recognition. The proposed model has been an advanced version of the traditional Detectron model which was trained using the training dataset and the effectiveness of the model was validated using the testing images. Results shows that the proposed model achieved superior performance as compared to existing YOLOv5, HDBNet, BiP-Net, and Res18-LVT models in terms of achieving better precision and F1 score. The advantages of the proposed approach are better holistic scene understanding**,** improved accuracy, and efficient implementation. While the model also suffers from certain limitations such as high complexity and computational cost**,** data requirements**,** model interpretability, fine-tuning and optimization and deployment challenges**.** For future research, this work can be extended to the implementation of advanced ensemble learning techniques for the detection of curved text from scenic images.

References

[1] Atitallah A., Said Y., Atitallah M., Albekairi M., Kaaniche K., Alanazi T., Boubaker S., and Atri M., "Embedded Implementation of an Obstacle Detection System for Blind and Visually Impaired Persons' Assistance Navigation," *Computers and Electrical Engineering*, vol. 108, pp. 108714, 2023. https://doi.org/10.1016/j.compeleceng.2023.108714

[2] Alshanqiti A., Bajnaid A., Gilal A., Aljasir S., Alsughayyir A., and Albouq S., "Intelligent Parallel Mixed Method Approach for Characterising Viral Youtube Videos in Saudi Arabia," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 661-671, 2020. DOI:10.14569/IJACSA.2020.0110382

[3] Bhatt M., Arya D., Mishra A., Singh M., Singh P., and Gautam M., "A New Wavelet-Based Multifocus Image Fusion Technique Using Method Noise-Median Filtering," *in Proceedings of the 4th International Conference on Internet of Things: Smart Innovation and Usages*, Ghaziabad, pp. 1-6, 2019. DOI:10.1109/IoT-SIU.2019.8777615

[4] Baek Y., Lee B., Han D., Yun S., and Lee H., "Character Region Awareness for Text Detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, pp. 9365-9374, 2019. DOI:10.1109/CVPR.2019.00959

[5] De Carvalho O., de Carvalho Júnior O., Albuquerque A., Bem P., Silva C., Ferreira P., Santos de Moura R., Trancoso Gomes R., Guimarães R., and Borges, D., "Instance Segmentation for Large, Multi-Channel Remote Sensing Imagery Using Mask-RCNN and a Mosaicking Approach," *Remote Sensing*, vol. 13, no. 1, pp. 39, 2021. https://doi.org/10.3390/rs13010039

[6] Choudhary S., Singh N., and Chichadwani S., "Text Detection and Recognition from Scene Images Using MSER and CNN," *in Proceedings of the 2nd International Conference on Advances in Electronics, Computers and Communications*, Bangalore, pp. 1-4, 2018. DOI:10.1109/ICAECC.2018.8479419

[7] Detectron2. detectron2 documentation, https://detectron2.readthedocs.io/en/latest/ Last Visited: 2024.

[8] Dey R., Balabantaray R., and Mohanty S. "Sliding Window Based Off-Line Handwritten Text Recognition Using Edit Distance," *Multimedia Tools and Applications*, vol. 81, pp. 22761-22788, 2022. https://doi.org/10.1007/s11042-021-10988-9

[9] Gilal A., Jaafar J., Capretz L., Omar M., Basri S., and Aziz I., "Finding an Effective Classification Technique to Develop a Software Team Composition Model," *Journal of Software: Evolution and Process*, vol. 30, no. 1, pp. e1920, 2018. DOI:10.1002/smr.1920

[10] Geetha M., Pooja R., Swetha J., Nivedha N., and Daniya T., "Implementation of Text Recognition and Text Extraction on Formatted Bills Using Deep Learning," *Int J Contrl Automat*, vol. 13, no. 2, pp. 646-651, 2020.

[11] He K., Zhang X., Ren S., and Sun J., "Deep Residual Learning for Image Recognition," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 770-778, 2016. DOI:10.1109/CVPR.2016.90

[12] Huang L., Tseng H., Hsieh, C., and Yang C., "Deep Learning Based Text Detection Using Resnet for Feature Extraction," *Multimedia Tools and Applications*, vol. 82, pp. 46871-46903, 2023. https://doi.org/10.1007/s11042-023-15449-z

[13] Huang J., Haq I., Dai C., Khan S., Nazir S., and Imtiaz M., "Isolated Handwritten Pashto Character Recognition Using a K-NN Classification Tool Based on Zoning and HOG Feature Extraction Techniques," *Complexity*, vol. 2021, no. 558373, pp. 1-8, 2021. https://doi.org/10.1155/2021/5558373

[14] Islam M., Monda C., Azam M., and Islam A., "Text Detection and Recognition Using Enhanced MSER Detection and a Novel OCR Technique," *in Proceedings of the 5th International Conference*

on Informatics, Electronics and Vision, Dhaka, pp. 15-20, 2016. DOI:10.1109/ICIEV.2016.7760054

[15] Jamieson L., Moreno-Garcia C., and Elyan E., "Deep Learning for Text Detection and Recognition in Complex Engineering Diagrams," *in Proceedings of the International Joint Conference on Neural Networks*, Glasgow, pp. 1-7, 2020. DOI:10.1109/IJCNN48605.2020.9207127

[16] Kirillov A., He K., Girshick R., Rother C., and Dollár P., "Panoptic segmentation," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, pp. 9404-9413, 2019. DOI:10.1109/CVPR.2019.00963

[17] Lin Z., Chen Y., Chen P., Chen H., Chen F., and Ling N., "JMNET: Arbitrary-Shaped Scene Text Detection Using Multi-Space Perception," *Neurocomputing*, vol. 513, pp. 261-272, 2022. https://doi.org/10.1016/j.neucom.2022.09.095

[18] Liao M., Zhang J., Wan Z., Xie F., Liang J., Lyu P., Yao C., and Bai X. "Scene Text Recognition from Two-Dimensional Perspective," *in Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, pp. 8714-8721, 2019. https://doi.org/10.1609/aaai.v33i01.33018714

[19] Liu X., Meng G., and Pan C., "Scene Text Detection and Recognition with Advances in Deep Learning: A Survey," *International Journal on Document Analysis and Recognition*, vol. 22, pp. 143-162, 2019. https://doi.org/10.1007/s10032-019-00320-5

[20] Lin T., Dollár P., Girshick R., He K., Hariharan B., and Belongie S. "Feature Pyramid Networks for Object Detection," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 2117-2125, 2017. DOI:10.1109/CVPR.2017.106.

[21] Maskrcnn Benchmark. https://github.com/facebookresearch/maskrcnn-benchmark, Last Visited: 2024.

[22] Ma Y. and Wang Y., "Feature Refinement with Multi-Level Context for Object Detection," *Machine Vision and Applications*, vol. 34, no. 49, 2023. https://doi.org/10.1007/s00138-023-01402-5

[23] Manjari K., Verma M., Singal G., and Namasudra S., "QEST: Quantized and Efficient Scene Text Detector Using Deep Learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1-18. 15, 2023. https://doi.org/10.1145/3526217

[24] Obi Y., Claudio K., Budiman V., Achmad S., and Kurniawan A., "Sign Language Recognition System for Communicating to People with Disabilities," *Procedia Computer Science*, vol. 216, pp. 13-20. 2023. https://doi.org/10.1016/j.procs.2022.12.106

[25] Qiao Z., Zhou Y., Yang D., Zhou Y., and Wang W., "Seed: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition," *in Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition*, Seattle, pp. 13528-13537, 2020. DOI:10.1109/cvpr42600.2020.01354

[26] Sah A., Bhowmik S., Malakar S., Sarkar R., Kavallieratou E., and Vasilopoulos N., "Text and Non-Text Recognition Using Modified HOG Descriptor," *in Proceedings of the IEEE Calcutta Conference*, Kolkata, pp. 64-68, 2017. DOI:10.1109/CALCON.2017.8280697

[27] Titijaroonroj T., "Modified Stroke Width Transform for Thai Text Detection," *in Proceedings of the International Conference on Information Technology*, Khon Kaen, pp. 1-5, 2018. DOI**:**10.23919/INCIT.2018.8584869

[28] Tan S., Chuah J., Chow C., Kanesan J., and Leong H., "Artificial Intelligent Systems for Vehicle Classification: A Survey," *Engineering Applications of Artificial Intelligence*, vol. 129, pp. 107497, 2024. https://doi.org/10.1016/j.engappai.2023.107497

[29] Turki H., Elleuch M., Othman K., and Kherallah M., "Arabic Text Detection on Traffic Panels in Natural Scenes, Arabic Text Detection on Traffic Panels in Natural Scenes," *The International Arab Journal of Information Technology*, vol. 21, no. 4, pp. 571-588, 2024 https://doi.org/10.34028/iajit/21/4/3

[30] Verma M., Sood N., Roy P., and Raman B., "Script Identification in Natural Scene Images: A Dataset and Texture-Feature Based Performance Evaluation," *in Proceedings of International Conference on Computer Vision and Image Processing*, Venice, pp. 309-319, 2017. DOI:10.1007/978-981-10-2107-7_28

[31] Wang X., He Z., Wang K., Wang Y., Zou L., and Wu Z., "A Survey of Text Detection and Recognition Algorithms Based on Deep Learning Technology," *Neurocomputing*, vol. 556, 2023. https://doi.org/10.1016/j.neucom.2023.126702

[32] Wang L., Yao X., and Song C., "Text Detection Method Based on HDBNet in Natural Scenes," *The Journal of Engineering*, vol. 2023, no. 1, pp.1-10, 2023. https://doi.org/10.1049/tje2.12212

[33] Wu F., Zhu C., Xu J., Bhatt M., and Sharma A., "Research on Image Text Recognition Based on Canny Edge Detection Algorithm and K-Means Algorithm," *International Journal of System Assurance Engineering and Management*, vol. 13, pp. 72-80, 2021. https://doi.org/10.1007/s13198-021-01262-0

[34] Wang X., Zheng S., Zhang C., Li R., and Gui L., "R-YOLO: A Real-Time Text Detector for

Natural Scenes with Arbitrary Rotation," *Sensors*, vol. 21, no. 3, 2021. https://doi.org/10.3390/s21030888

[35] Yang C., Chen M., Yuan Y., and Wang Q., "Bip-Net: Bidirectional Perspective Strategy Based Arbitrary-Shaped Text Detection Network," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, pp. 2255-2259, 2022. DOI:10.1109/ICASSP43922.2022.9747331

[36] Yang K., Yi J., Chen A., and Jin Z., "Buffer-Text: Detecting Arbitrary Shaped Text in Natural Scene Image," *Engineering Applications of Artificial Intelligence*, vol. 130, pp. 107774, 2024. https://doi.org/10.1016/j.engappai.2023.107774

[37] Yu D., Li X., Zhang C., Liu T., Han J., Liu J., and Ding E., "Towards Accurate Scene Text Recognition with Semantic Reasoning Networks," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, pp. 12113-12122, 2020. https://doi.org/10.48550/arXiv.2003.12294

[38] Yang L., Ergu D., Cai Y., Liu F., and Ma B., "A Review of Natural Scene Text Detection Methods," *Procedia Computer Science*, vol. 199, pp. 1458-1465, 2022. https://doi.org/10.1016/j.procs.2022.01.185

[39] Yu R., Jin F., Qiao Z., Yuan Y., and Wang G., "Multi-Scale Image-Text Matching Network for Scene and Spatio-Temporal Images," *Future Generation Computer Systems*, vol. 142, pp. 292-300, 2023. https://doi.org/10.1016/j.future.2023.01.004

[40] Yim M., Kim Y., Cho H., and Park S., "Synth TIGER: Synthetic Text Image Generator Towards Better Text Recognition Models," *in Proceedings of the Document Analysis and Recognition-16th International Conference*, Lausanne, pp. 109-124, 2021. https://doi.org/10.1007/978-3-030-86337-1_8

[41] Yin F., Wu Y., Zhang X., and Liu C., "Scene Text Recognition with Sliding Convolutional Character Models," *arXiv Preprint*, arXiv:1709.01727.24, 2017.

[42] Zhang F., Luan J., Xu Z., and Chen W., "DetReco: Object-Text Detection and Recognition Based on Deep Neural Network," *Mathematical Problems in Engineering*, vol. 2020, no. 2365076, pp. 1-15, 2020. https://doi.org/10.1155/2020/2365076

[43] Zhang C., Ding W., Peng G., Fu F., and Wang W., "Street View Text Recognition with Deep Learning for Urban Scene Understanding in Intelligent Transportation Systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no.7, pp. 4727-4743, 2020. DOI:10.1109/TITS.2020.3017632

[44] Zhong D., Lyu S., Shivakumara P., Pal U., and Lu Y., "Text Proposals with Location-Awareness-Attention Network for Arbitrarily Shaped Scene Text Detection and Recognition," *Expert Systems with Applications*, vol. 205, 2022. https://doi.org/10.1016/j.eswa.2022.117564

[45] Zacharias E., Teuchler M., and Bernier B., "Image Processing Based Scene-Text Detection and Recognition with Tesseract," *ArXiv Preprint*, vol. arXiv:2004.08079.20, pp. 1-6, 2020. DOI:10.48550/arXiv.2004.08079

Prachi Chhabra is an Assistant Professor at JSS Academy of Technical Education, Noida. She holds an M.Tech from Kurukshetra University and has contributed to research in AI, particularly in neural networks and text detection. She is currently pursuing PhD from Sharda university, Greater Noida. Her research interest includes Deep learning, Text Detection, Object Detection Models, she has 15 years of teaching experience. She is also active in academic administration and other related activities.

Ali Abidi is an academic expert in Computer Vision, Visual Data Analytics, and Deformable Image Registration, with a Ph.D. from IIT (BHU) Varanasi. His initial research focused on thoracic CT image registration, addressing breathing-induced deformations. Since then, he has widened his research avenues to a broader range of computer vision applications. Currently an Associate Professor at Sharda University, he has 9 years of experience in academia and industry, including roles at Infosys and the National Institute of Design. Dr. Abidi has published widely in reputable journals, contributed to conferences, and holds patents in image processing and AI-driven technologies. He is also active in academic administration and Ph.D. supervision, emphasizing hands-on learning.