Text Localization and Script Identification in Natural Scene Images and Videos

Chandana Udupa
Dept. of Electrical and Electronics
KLE Technological University
Hubballi, Karnataka, India
chandanaudupaofficial@gmail.com

Anusha Upadhyaya
Dept. of Electrical and Electronics
KLE Technological University
Hubballi, Karnataka, India
anusha168786@gmail.com

Basanagouda S Patil
Dept. of Electrical and Electronics
KLE Technological University
Hubballi, Karnataka, India
basanagoudapatil275@gmail.com

Shivananda V. Seeri
Dept. of Master of Computer Applications
KLE Technological University
Hubballi, Karnataka, India
seeri@kletech.ac.in

Prakashgoud Patil
Dept. of Master of Computer Applications
KLE Technological University
Hubballi, Karnataka, India
prakashpatil@kletech.ac.in

P. S. Hiremath
Dept. of Master of Computer Applications
KLE Technological University
Hubballi, Karnataka, India
pshiremath@kletech.ac.in

2022 International Conference on Connected Systems & Intelligence (CSI) | 978-1-6654-5815-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSI54720.2022.9924044

Abstract—**Text detection and its script identification in a natural scene image/video has attracted the attention of many researchers over the recent years due to its application in the design of computer vision devices for usage by the visually impaired people, global tourists travelling in unfamiliar tourist places, etc. to facilitate them to understand the textual information displayed on sign boards, bill boards, public notice boards, etc., the objective of the proposed method is detection and localization of multilingual text in a natural scene video image and its corresponding script identification. The texts in three languages, namely, English, Hindi and Kannada, are considered. In the proposed method, CNN based YOLOv5 is used for text detection and localization in real-time videos of natural sceneand it is also trained for script identification. The YOLOv5 performance is found to yield an accuracy higher than otherobject detection algorithms. The proposed model is trained witha custom dataset containing video images of natural scenes and istested for different scenarios like texts in different backgrounds, fonts, orientations, resolutions, and disturbances in the images. The experimental results demonstrate the effectiveness and robustness of the proposed method. The performance comparison is done with other methods in the literature.**

Keywords— text localization, script identification, YOLO, natural scenes imagery, CNN.

I. INTRODUCTION

This Text is considered to be the most expressive means of communication and can be embedded into documents, images, and videos as a means of communicating information. The collection of massive amounts of street-view data is one such compelling application. The other factor that can be considered is the increasing availability of high-performance mobile devices, with both imaging and computational capability. This creates an opportunity for image acquisition and processing anytime, anywhere, and thus making it a suitable computer vision tool to recognize text in a natural scene image/video under various environments. Lastly, the advances in computer vision and pattern recognition technologies along with deep learning, make it more feasible to address such demanding application problems.

Over the years, the problem of Optical Character Recognition (OCR) for a document image is well investigated and several effective methods have been evolved. However, these methods do not work efficiently for images/documents having complex backgrounds, arbitrarily shaped texts, and different sizes of text, which are the normally observed distinguishing characteristics of natural scene images. This also requires a great deal of effort in fine-tuning the parameters and slows down the detection process. Still, there are abundant opportunities for further research to improve text detection accuracy. In countries like India, people are accustomed to multilingual communication in many regional languages like Kannada, Hindi, Gujarati, Marathi, etc. along with English. It is often seen that street-view video/image data, and some documents, are multilingual in nature. It is comparatively easier to detect and identify English texts as it consists of only 26 lowercase or uppercase characters. For Indian language scripts, text localization and identification in document images with such scripts become an arduous task with OCRs.

However, the problem of multilingual text detection and localization in natural scene videos is not addressed in the literature. In this paper, a deep learning technique is proposed for text localization and tri-lingual script identification in a natural scene image or video using YOLOv5, which is a novel object detection algorithm that uses a single CNN. It is very accurate and fast with 45 frames per second, and has better performance than RCNN and other object detection algorithms. For text localization and script identification, three languages, namely, Kannada, Hindi, and English are considered, and the training dataset is prepared by collecting still images of multilingual texts. The trained model is then tested on real time pre-recorded street-view videos to check the performance. Some sample images of the training dataset, such as street view and

978-1-6654-5815-3/22/$31.00 ©2022 IEEE

other graphic text translation images, used for training the model are shown in the Fig.1.

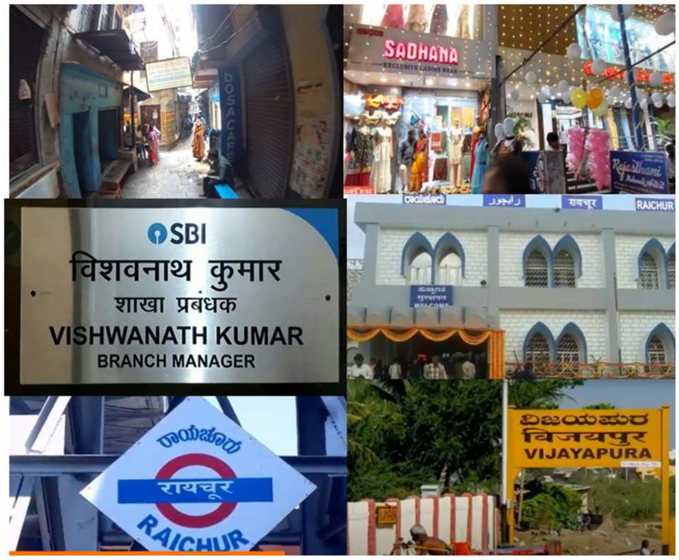

Fig. 1. Input given to trained model

II. RELATED WORKS

The natural scene image text detection and script identification is hard to achieve as they involve various challenges like text orientation, text distortion, image resolutions, multi-lingual scripts involved in an image, complex backgrounds, various fonts and styles of the text. Over the years, researchers have focused on achieving the text detection and text translation in natural scene images and videos by introducing various architectures and models. While many researchers have aligned towards OCR as the possible best solution for text localization with script identification and text extraction. It is also witnessed that OCR does not work well with multi-lingual scripts and different text orientations. There is still scope for a better model that is able to localize the text and extract it from the image/video.

As the first step, the text localization in natural scene images, Seeri et.al proposed a novel method based on Wavelet features and fuzzy classification [1]. The deep learning based model was proposed for text detection and script identification in [2]. The proposed uses Haar wavelet edge features, K-means clustering, fuzzy classification and threshold concepts for text localization in the natural scene images. Later, a blended approach was proposed for extraction and recognition of text present in natural scene images. In the first step, test regions are located using GLCM features extracted from Contourlet transformed image and SVM (Support Vector Machine) classifier [3]. In second step, characters are segmented using horizontal projections and vertical projections. Further, segmented characters are recognized for English script using Contourlet transform, moment invariants and Probabilistic Neural Network (PNN) classifier [4].

While some researchers saw as an opportunity to work on multi-lingual scripts, other worked on single script to achieve greater accuracy and better refinement in text detection and subsequent text extraction. Khanaghavalle. G. R and N. Rajeswari [5] dealt with random shaped Hindi text detection in scene images by proposing a novel Hindi text detector using ResNet as the backbone network. The proposed method is able to locate arbitrary shaped Hindi text in complex background images and has been experimented with the benchmark dataset IC19- MLT(Hindi). This method is able to achieve accuracy in the range of 74% -78.5%.

Huibai Wang and Hongqing Shi [6] proposed method based on UDSP-Yolov3 for a text detection, In addition, it uses a CLAHE image enhancement method to remove the effect of lighting changes in images. The method uses S3 Pooling technique for feature extraction.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman [7], tested synthetic data for text localization in natural images by adopting fully convolutional regression network (FCRN). The proposed method is compared with other end-to-end object detection systems. FCRN achieves an F-measure of 84.2% on the standard ICDAR 2013 dataset.

Tianxiang Zhou, Ke Wang, Jun Wu, and Ruifeng Li [8] proposed a video text processing based on image stitching. The YOLO algorithm is used for Text detection and ECO algorithm is used for tracking text in video. The panorama is realized is using global stitching method.

Ankan Kumar et.al [9] proposed a novel attention method using Convolutional-LTSM network for script identification of text present in natural scene images. In the proposed method, CNN-LSTM framework is used to extracts local, global features for script identification. The model is experimented on four datasets: SIW-13, CVSI2015, ICDAR17 and MLe2e. The suggested model yields an accuracy of 97.75% for script identification. There are two vital parts in achieving text localization and script identification successful one is the training dataset while other one is the machine learning model used for the purpose.

The dataset helps the model to effectively localize and classify the text that are found in natural scene images or videos. Manish Verma et.al [10] proposed a custom dataset constructed that has images of railway sign-boards written in 5 different Indic scripts. In the proposed model involves local textual features that are used for feature extraction along with SVM classifier. The proposed model yields an accuracy range 90%-97.5% and is used for multi-lingual script detection.

As there are many deep learning methods for text detection and its script identification, Ashwaq Khalil et.al [11] proposed is fully convolutional networks for model enhancement and classification. In this paper, two end-to-end deep learning methods are proposed, namely, multi-channel mask (MCM) and multi-channel segmentation (MCS). MCS outperforms existing methods on the ICDAR MLT 2017 and MLe2e datasets, respectively. With increase in text detection architectures for natural scene images and videos was proposed by various researchers, some focused on re- viewing those methods to find the most suitable algorithm for the purpose. Shilpa Mahajan

[12] discusses comparison of recent machine learning and deep learning based approaches for the text detection and localization of natural scene images having text in different languages. The models are tested on various benchmarked datasets, and comparative analysis was carried out on benchmarked datasets. The related challenges and future scope of the different models on the various fields are given in the survey paper.

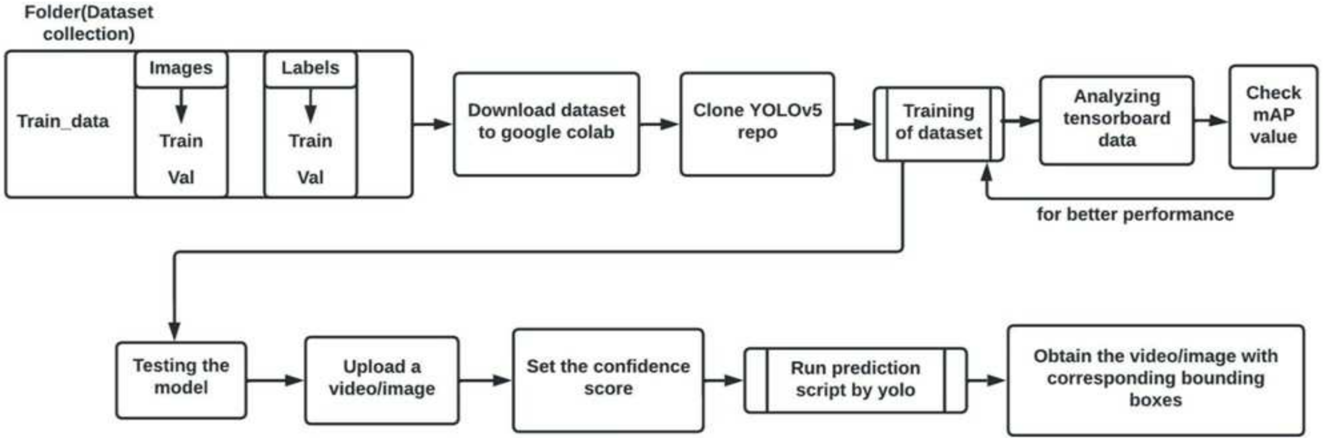

Fig. 2. Overview of the proposed model

III. PROPOSED METHODOLOGY

The proposed methodology comprises deep learning neural network, namely, YOLOv5 based on DarkNet53, for text localization and script identification in a natural scene image/video. It consists of two phases: training and testing. The YOLO performs multiple target (object) detection and tracking in a video. During the training stage, the YOLO is trained using the training dataset of natural scene images/videos, which contain multi-lingual text objects. During the testing phase, a natural scene image/video is given as input to the trained YOLO and obtain image/video with bounding box for the detected text region and label the box with the identified script as the output.

The YOLO adopts binary cross-entropy loss and logistic regression to predict multiple objects in an image/video. The block diagram of the proposed methodology is shown in the Fig. 2.

The parameters set to train the model are:

No. of GPU(s) = 1

No. of classes/objects = 3

Name of the classes = ’Kannada’, ’Hindi’,’English’

Batch size = 16

Number of Epochs = 250

Weights = yolov5s.pt

The model is trained by running the YOLOv5 with DarkNet53 as the CNN. The training dataset consists of images with text object annotated with class labels. The testing of the model is done by using the test dataset comprising graphic text videos and real time street videos. The implementation is done using Python in Google COLAB environment.

Many videos of natural scenes are available that have the tri-lingual texts, i.e., text in English, Hindi, and Kannada scripts. But a standard dataset having the images of these tri-lingual scripts is scarcely found. Hence, an own dataset is prepared, which comprises two sets, one being used for training the model while the other one is for validation. The images have been collected from several sources. Most of the images with horizontally oriented scripts are acquired from commercial translation videos. The natural scene images are obtained from the videos having street-view of few places located in Karnataka, Varanasi, Delhi, and Mumbai.

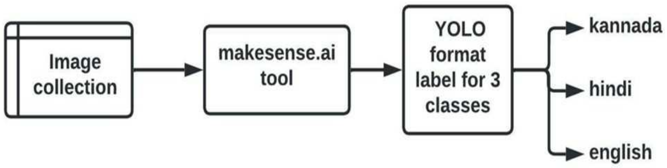

Fig. 3. Overview of dataset collection

The natural scene images have scripts of different orientation, colors, backgrounds, and image resolutions. The images with noise and disturbances are also used to train the model to make it robust and efficient.

The training dataset has images having characters of Hindi, English, and Kannada as well as words in these scripts in order to enrich the model training. The annotation of text objects in the images is done using the labeling tool, namely, ‘makesense.ai’. The text object labels are Hindi, English and Kannada that indicate the script of the text. The overview of data acquisition is displayed in Fig.3.

In the YOLO format, a label is 5-tuple (x,y,w,h,c), where parameters (x,y) denote the upper-left corner coordinates of the bounding box, (w, h) denote its width and height, and c corresponds to the object class (script). The prepared dataset has

1308 images with suitable annotations. The 1290 images are used for training the model and remaining 18 images are used for testing. The Fig.4 shows some sample images of text in natural scenes captured in street view videos.

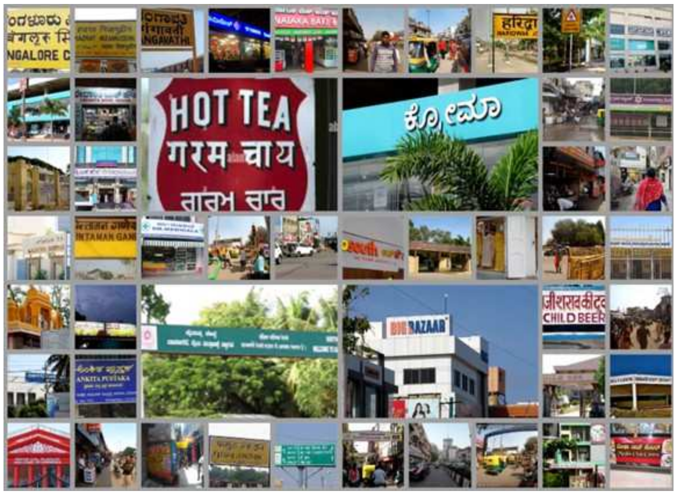

Fig. 4. Images from recorded street videos

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To test the efficacy of the model, the trained model is tested on several graphic text translation videos and real-time recorded street view videos. The model has a speed of 45 frames per second and hence it is able to detect texts in the video very fast and also identifies the script of the localized text. The YOLO detects the text by drawing the bounding box around it and identifies the script as one among 3 classes and displays the confidence score for each detection. Fig. 5 shows the images from real-time recorded street view videos where only the texts in English, Hindi and Kannada scripts are detected.

Fig. 6 shows the images from videos where only the texts of languages English, Hindi and Kannada are detected with different text size and style.

To check the robustness of the trained YOLOv5 model, it is tested on different scenarios, i.e., color backgrounds, text fonts, orientations of the text, lighting conditions in the images and image resolutions, etc. The Fig. 7 shows a sample result, wherein Hindi text with varying font style is detected by the trained model. In Fig.8, text on a glass surface is detected with the light being reflected on the image. When the text in the image is vertically oriented, the text detection is done as shown in the Fig. 9. This model is also able to detect text in a blurred image. The trained model treats any untrained objects present in an image as background and detects only the texts.

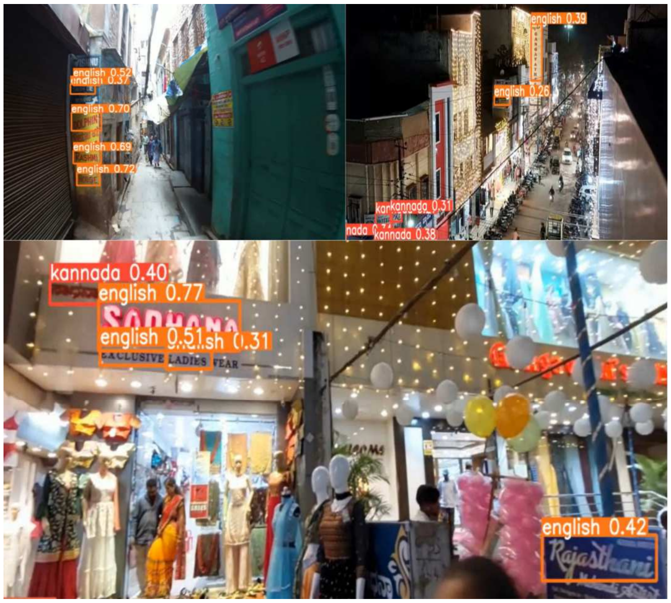

Fig. 5. Text detection in videos with different text size and style

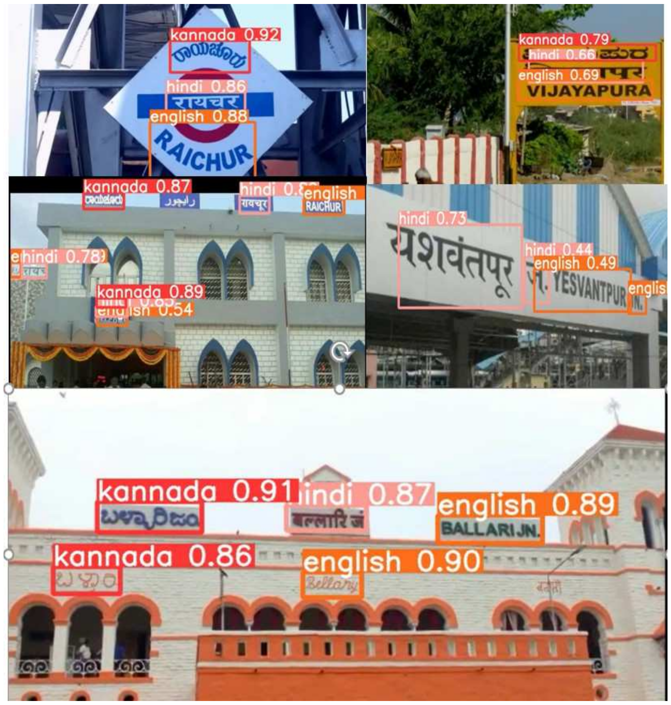

Fig. 6. Text detection in recorded real-time videos

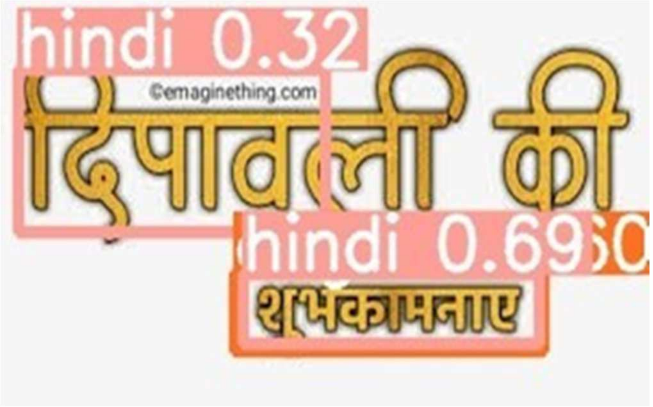

Fig. 7. Different font style.

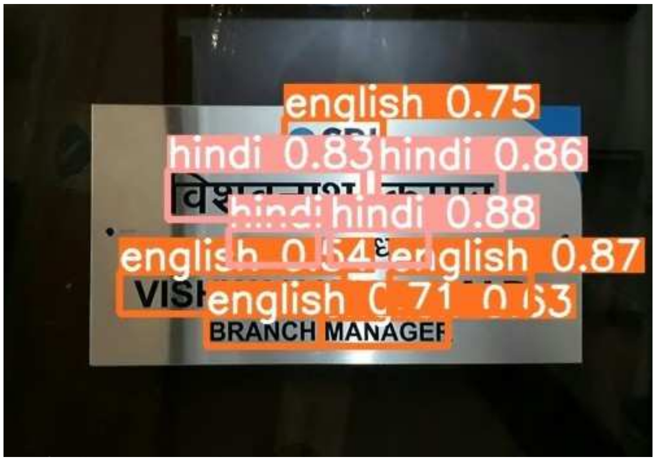

Fig. 8. Image with light reflection

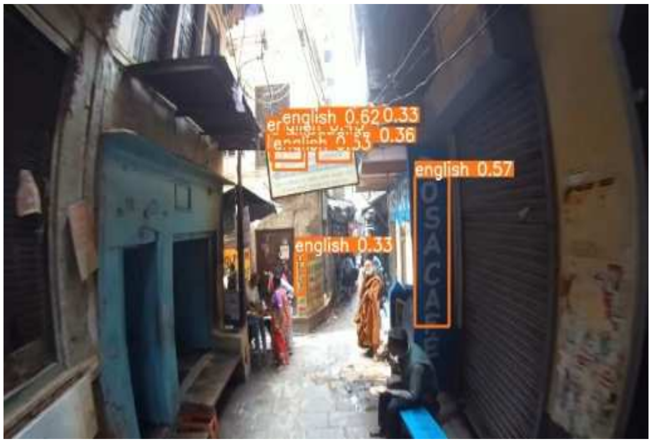

Fig. 9. Image with different text orientations

- The Fig.10 represents the precision-confidence curve for Kannada, Hindi, English scripts and all classes. It is seen that as the confidence score increases from 0 to 1, the precision also increases, i.e., true positives increases.

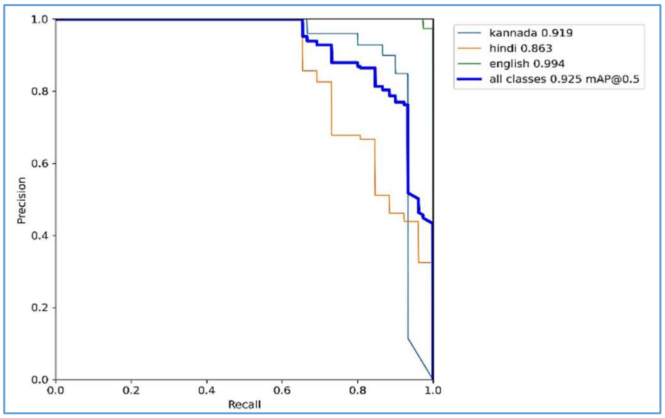

Fig. 10. Precision-Confidence curve

- The Fig. 11 represents the recall-confidence curve for 250 epochs respectively the recall-confidence curve, it is seen that as the confidence score increases between 0 and 1, the recall decreases.

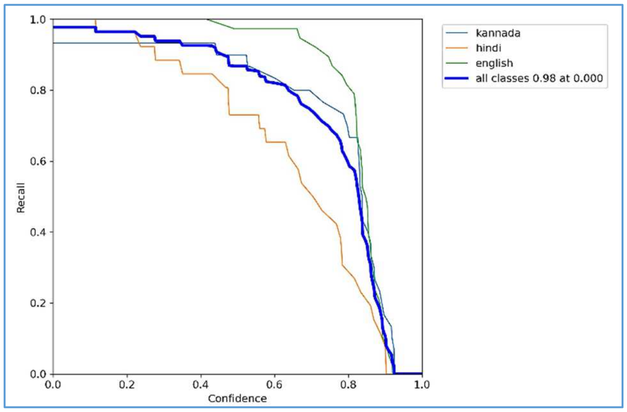

Fig. 11. Recall-Confidence curve

- A good Precision-recall curve also called as PR curve has a greater AUC (area under the curve). The Fig.12 represents the PR curve for Kannada, Hindi, English scripts and all classes at mAP 0.5. A greater AUC, i.e., with a value nearer to 1, implies better detection results.

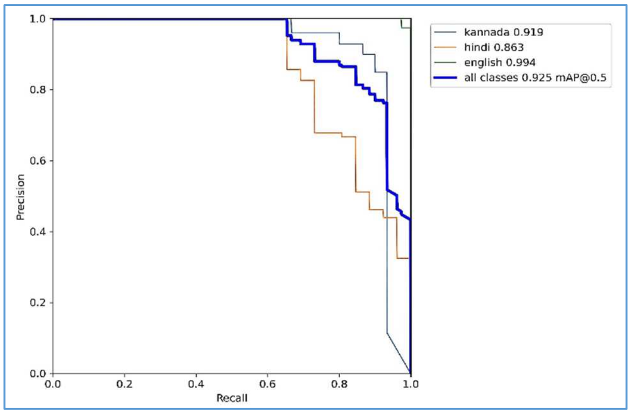

Fig. 12. Precision-Recall curve

- F1 curve: Fig. 13 represents the F1curve.The confidence value that optimizes the precision and recall is 0.557, corresponding to the maximum F1 value (0.88). In most cases, a higher confidence value and F1 score are desirable.

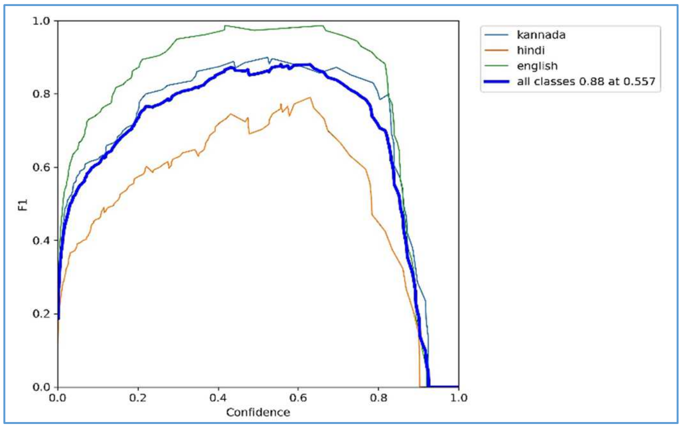

Fig. 13. F1 Curve

mAP value: Fig. 14 represents the curve with mAP value. The mAP is used to evaluate the performance of both classification and localization using bounding boxes with the help of the concept of IoU. Here the IoU is set as 0.5(threshold value). The Fig. 15 represents mAp curve with IoU threshold value ranging from 0.5 to 0.95 with increase in step of 0.05.

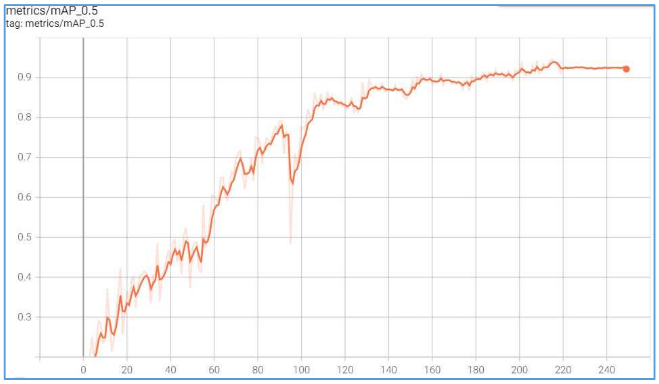

Fig. 14. 0.5 mAP curve

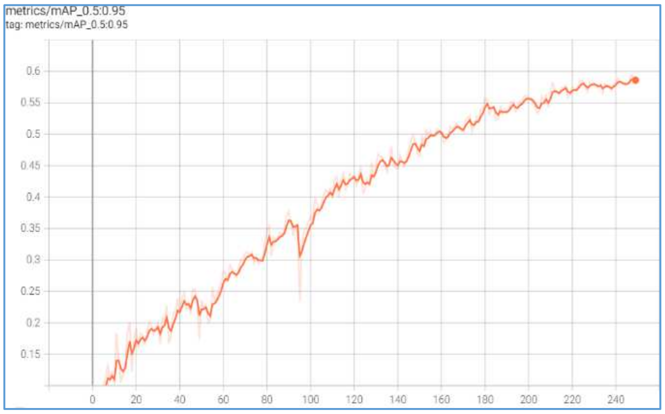

Fig. 15. 0.5:0.95 mAP curve

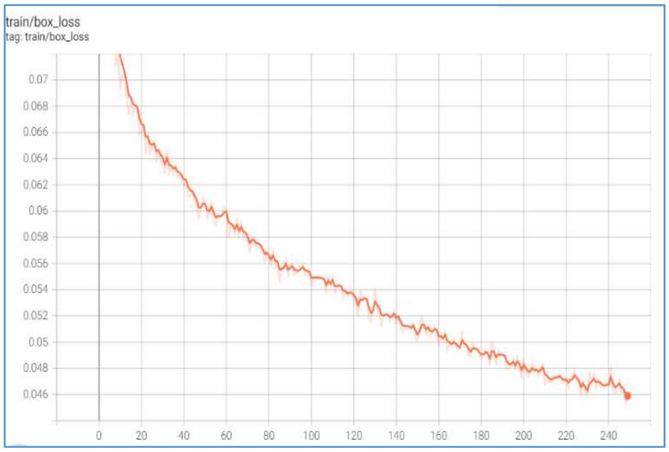

Fig. 16. Box Loss curve

Train losses: There are three different types of train losses, namely, Box loss, Objectness loss, and Classification loss, which are shown in the Fig.16, Fig. 17 and Fig. 18, respectively.

TABLE I. PROPOSED MODEL ACCURACY FOR EACH CLASS

Methods	Accuracy			
	Kannada	Hindi	English	Overall
Proposed Yolov5 model	97.30%	88.50%	98.60%	**94.80%**

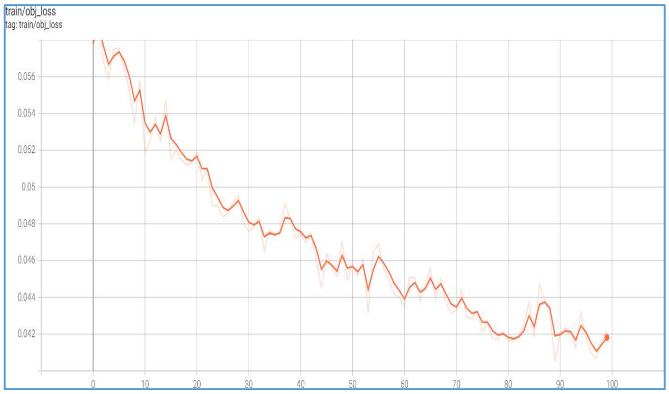

Fig. 17. Objectness Loss curve

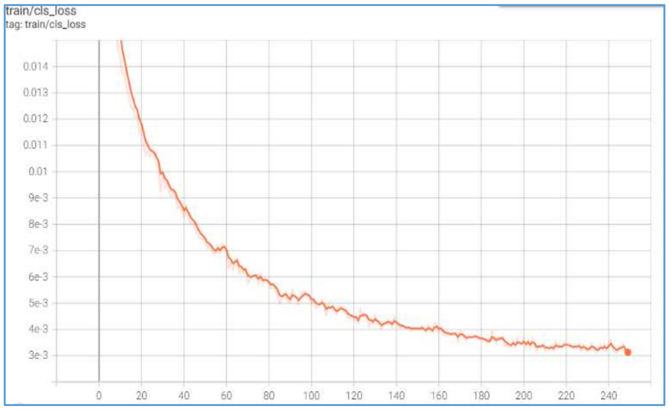

Fig. 18. Classification Loss curve

TABLE II. COMPARISON WITH OTHER METHODS

Methods	**Precision**	**Recall**	**F-measure**	**Accuracy**
CT-SVM[2]	98.85%	90.85%	94.68%	89.90%
CT-PNN[3]	98.85%	90.85%	95.00%	89.90%
UDSP-YOLO[5]	65.3%	45.1%	53.4%	95%
Proposed Model	92.6%	88.7%	90.60%	94.80%

The accuracy obtained for each class and the overall accuracy of the proposed model are displayed in TABLE I. In the TABLE II, the proposed method is compared with other text detection methods in the literature in terms of accuracy, precision, recall and F-measure. The proposed method based on YOLOv5 with DarkNet53 is found to perform better in comparison with other methods.

V. CONCLUSION

The problem of multilingual text detection and localization in natural scene videos is addressed in this paper. A deep learning method is proposed herein for multilingual detection, localization and script identification of text in natural scene images/videos using YOLOv5 with DarkNet53. Despite the complexities in the images and intricacies in the curves of Kannada, Hindi, and English characters, the proposed model is able to detect the text and identify the script successfully. Some of the most challenging issues in text detection of natural scene images have been addressed, which include varying backgrounds of the image, text orientations, font styles, resolutions, and lighting conditions that makes the text detection in the natural scene image more complex.

The proposed model provides overall accuracy of 94.80%-97% based on increase in dataset and tweaking few parameters. However, there is still scope for further improvement in the text detection and script recognition in natural scene images/videos, which finds in many important applications, such as text extraction and subsequent translation to user- choice languages which would help tourists and the visually impaired to have a better understanding of the surroundings.

REFERENCES

[1] Shivananda V. Seeri, J. D. Pujari and P. S. Hiremath. “Multilingual Text Localization in Natural Scene Images using Wavelet based Edge Features and Fuzzy Classification”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 4, Issue 1, January- February 2015.

[2] Ashwaq Khalil, Moath Jarrah, Mahmoud Al-Ayyoub, Yaser Jararweh, “Text detection and script identification in natural scene images using deep learning”, Elsevier journal of Computers and Electrical Engineering, 22 February 2021.

[3] Shivananda V. Seeri, J. D. Pujari, P. S. Hiremath, “Text Localization and Character Extraction in Natural Scene Images using Contourlet Transform and SVM Classifier”, I.J. Image, Graphics and Signal Processing, 36-42. May 2016.

[4] Shivananda V. Seeri, P. S. Hiremath, J. D. Pujari, Prakashgoud Patil. “Text Extraction and Recognition in Natural Scene Images using Contourlet Transform and PNN”, International Journal of Innovative Technology and Exploring Engineering (IJITEE),Volume-9 Issue-2S, December 2019.

[5] Khanaghavalle. G. R, Dr. N. Rajeswari, “Arbitrary Shape Hindi Text Detection for Scene Images”, International Research Journal of Engineering and Technology (IRJET), Volume: 07, Issue: 05, May 2020.

[6] Hubai Wang, Hongqing Shi, “Research on text detection method based on improved yolov3”, IEEE 5TH Advanced Information Technology, Electronic and Automation Control Conference, 2021.

[7] Ankush Gupta, Andrea Vedaldi, Andrew Zisserman, “Synthetic Data for Text Localisation in Natural Images”, IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[8] Tianxiang Zhou, Ke Wang, Jun Wu, Ruifeng Li, “Video Text Processing Method Based on Image Stitching”, IEEE 4th International Conference on Image, Vision and Computing, 2019.

[9] Ankan Kumar Bhunia, Aishik Konwer, Ayan Kumar Bhunia, Abir Bhowmick, Partha P. Roy, Umapada Pal. “Script Identification in Natural Scene Image and Video Frame using Attention based Convolutional-LSTM Network, 1 January 2018.

[10] Manisha Verma, Nitakshi Sood, Partha Pratim Roy and Balasubramanian Raman, “Script Identification in Natural Scene Images: A Dataset and Texture-Feature Based Performance Evaluation”, Proceedings of International Conference on Computer Vision and Image Processing, December 2017.

[11] Shilpa Mahajan, Rajneesh Rani , “Text detection and localization in scene images: a broad review”, Springer Nature B.V. on Artifcial Intelli- gence Review (2021), 16 April 2021.

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR), 2016.