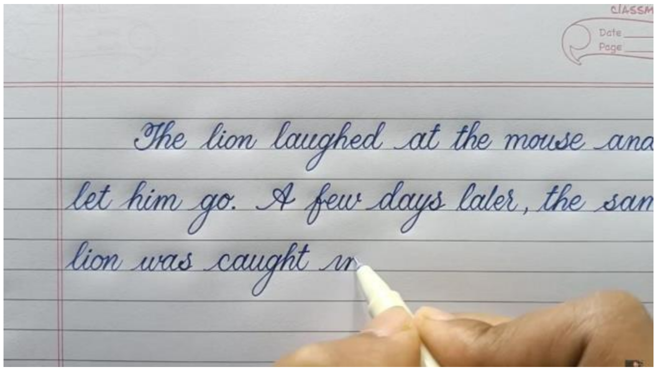
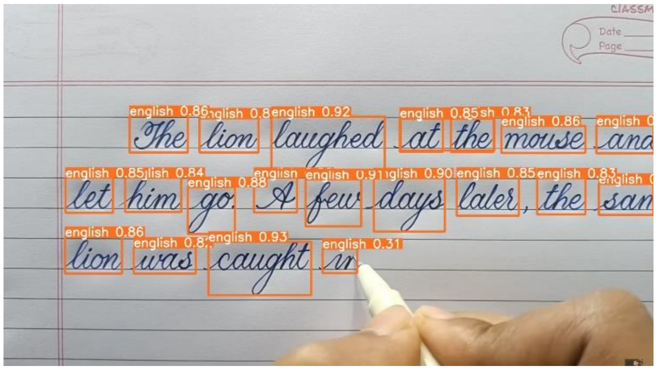
UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

# **Final Project:** Cursive Text Detection and Multi-Lingual Script Identification in Real-time Natural Scenes

*Group Members*: Apoorva Udupa (NetID: audupa3), Sagar Regmi (NetID: sagarr2).

In recent years, many researchers have been interested in detecting and identifying written text in natural scene images and videos. This is because of its potential applications in designing computer vision devices that can help visually impaired people or tourists in unfamiliar places to understand the information displayed on signs, billboards, and notice boards. The proposed method aims to detect and locate multilingual text in natural scene videos and identify its script. The method focuses on four languages: English, Hindi, Kannada, and Chinese. The most important phase of this project is to train the model for cursive data and identify them. The method uses a CNN-based YOLOv5 algorithm for image text detection and localization to achieve this. The proposed model is trained with a custom dataset of natural scene images and tested in various scenarios involving different backgrounds, fonts, orientations, resolutions, and image disturbances. Further, Hyperparameter tuning is also performed to make the model identify cursive data. The experimental results show that the proposed method is effective and robust. We further compare the model and its parameters before and after performing hyperparameter tuning to find the best weights and to find the most optimal trained model for detection.


CLASSM
Date
Page
The lion laughed at the mouse and
let him go. A few days later, the san
lion was caught in



# Introduction:

When every human has different handwriting and cursive styles, it becomes difficult for computers, and machines to predict these texts, especially cursive texts with different orientations, usually found on billboards and man-made name boards. We plan to achieve this by training a real-time object detection model with a cursive dataset. The next step would be to find out the type of language in the detected text, which can be used in applications like text translations that are used by tourists or attain voice-over to the visually impaired for assistance. In this project, we try to realize text detection of multiple languages like English, Chinese, Hindi, and Kannada by building a diverse and precisely labeled dataset and training the model. Furthermore, to achieve remarkable accuracy, we will perform hyperparameter tuning on the Yolov5 model, increasing its efficiency to detect the image with better visual reconstruction. This project has its application branched out to various fields like tourist guidance, human-robot interaction, visually impaired assistance, etc.

The main motivation of the project is to improve the object detection model to recognize and detect cursive texts and also categorize the text to its respective scripts/languages. We also focus on increasing the efficacy of the model by tuning the hyperparameter which will better recognize the cursive texts found in real-time natural scenes.

Researchers have proposed various approaches to achieving text localization and script identification in natural scene images. Seeri et al. [2] utilized wavelet-based edge features and fuzzy classification as an initial step. Wang and Shi introduced a new method using Haar wavelet, edge features, K-means clustering, fuzzy classification, and threshold concepts to localize text and eliminate non-textual regions from images with complex backgrounds. Gupta, Vedaldi, and Zisserman [3] utilized synthetic data to introduce a fully convolutional regression network (FCRN) for text recognition and bounding box regression, achieving an F-measure of 84.2% on the ICDAR2013 dataset. Kumar et al. [5] proposed an attention-based Convolutional-LTSM network for script identification, which extracted local and global features using the CNN-LSTM framework and weighted them for script identification, achieving an accuracy of 97.75% on four script identification datasets. Successful text localization and script identification depend on both the training dataset and the machine learning model used.

We accomplished text localization and related script identification (Hindi, English, Kannada, and Chinese) in this project by training the YOLOv5 model with a strong custom dataset of over 2000 images. We also expanded the dataset and added cursive text detection. We additionally performed hyperparameter tuning to improve outcomes, particularly when recognizing cursive data. In addition, to test the cursive data detection, we manually constructed a cursive dataset and trained the YOLOv5 on it.

# Methodology:

The first phase of the project was creating the custom dataset to train the YOLOv5 Model. We created a custom dataset that includes the following:

- Texts of scripts/languages: English, Chinese, Hindi, Kannada.
- Cursive Style texts.
- Texts found on billboards, Name Boards, and on Streets.

- Captions and texts found in Pre-recorded videos.
- Texts of different orientations.
- Images captured in daylight and at night.

The texts found in an image are then labeled using the makesense.ai tool which creates a labeled text file for each image included in the dataset. This custom dataset is given to the model to train and validate.

For text localization and script recognition in a natural scene image/video, the suggested technique includes a deep learning neural network, YOLOv5 based on DarkNet53. It is divided into two stages: training and testing. In a video, the YOLO detects and tracks several targets (objects). The YOLOv5 is taught utilizing a training dataset of natural scene images/videos containing multi-lingual text items throughout the training stage. During the testing step, a natural scene picture or video is fed into the trained YOLO, which produces an image or video with a bounding box for the detected text region and labels the box with the recognized script. During the training process, the YOLO employs binary cross-entropy loss and logistic regression to accomplish category prediction. This enables YOLO to classify a target (object) as having several labels.

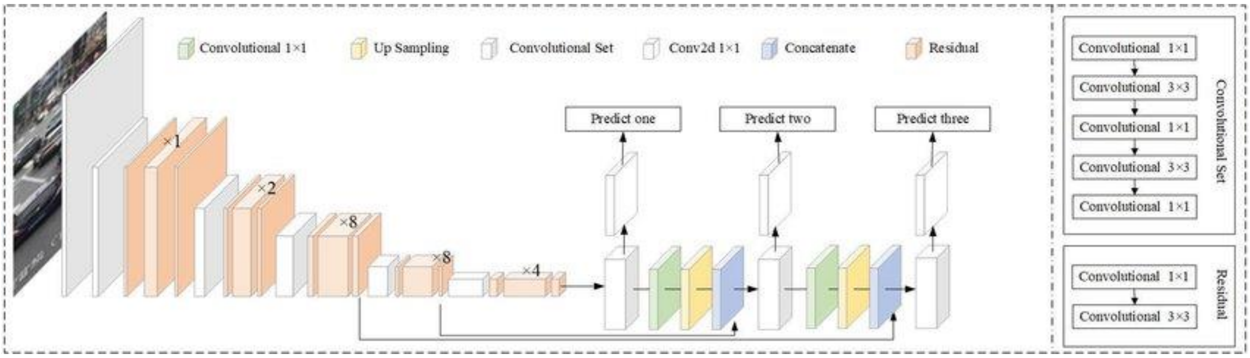


*Figure 1. Overview of YOLOv5*

***Case 1****:* Custom Dataset without Hyperparameter Tuning.

Initially, we trained the model with a custom training dataset that comprises 1959 images and its corresponding labels but without performing any hyperparameter training.

The values of the parameters:

1. No. of GPU(s) = 1
2. Name of the classes = ’Kannada’, ’Hindi’, ’English’, ‘Chinese’
3. Batch size = 16
4. Number of Epochs = 3
5. Image grid size = 640x640
6. Confidence score = 0.25

This yielded a training accuracy of 55% and a validation accuracy of 49%. To further enhance this model we performed Hyperparameter tuning, to fine-tune the results and increase the efficiency of the model in both training and testing.

***Case 2****:* Custom Dataset with Hyperparameter Tuning I.

To further increase efficiency and improve text detection and script identification. We performed hyperparameter tuning on parameters such as the number of epochs, and confidence score.

The values of the parameters:

1. No. of GPU(s) = 1
2. Name of the classes = ’Kannada’, ’Hindi’, ’English’, ‘Chinese’
3. Batch size = 16
4. Number of Epochs = 105
5. Image grid size = 728X728
6. Confidence score = 0.50

For these parameter values, we received a training accuracy of 99.1%. However, the test accuracy was not up to the mark as the model was overfitting. Hence, we performed Hyperparameter tuning again to obtain optimal results for text detection.

***Case 3****:* Custom Dataset with Hyperparameter Tuning II.

To further increase efficiency and improve text detection and script identification. We performed hyperparameter tuning on parameters such as the number of epochs, and confidence score.

The values of the parameters:

1. No. of GPU(s) = 1
2. Name of the classes = ’Kannada’, ’Hindi’, ’English’, ‘Chinese’
3. Batch size = 16
4. Number of Epochs = 100
5. Image grid size = 640X640
6. Confidence score = 0.50

The results obtained from this model were much better than that of the previous model (without hyperparameter tuning). We attained a training efficiency of 95.1% while the validation efficiency was 89.5%.

***Case 4:*** Cursive Dataset

As our agenda was to train the model to detect cursive data. We prepared a separate dataset that consisted of only cursive text images (500 images) and trained the model.

1. No. of GPU(s) = 1
2. Name of the classes = ’Kannada’, ’Hindi’, ’English’, ‘Chinese’
3. Batch size = 16
4. Number of Epochs = 100
5. Image grid size = 640X640
6. Confidence score = 0.25

The model was able to achieve a training accuracy of 99.5% and a testing accuracy of 99.1%. Hence, we have considered the results obtained from the test runs from cases 3 and 4 and discarded the results obtained from case 1, and case 2. we initially planned on implementing Data augmentation and data transformation to increase the efficacy of the model but we achieved higher precision on the validation set so we went ahead with the results that we achieved from the already trained model. The experimental results are displayed below.

## Results and Discussions:

The trained model is evaluated on multiple graphic text translation pictures and real-time recorded street view images to determine its efficacy. With a frame rate of 45 frames per second, the model can recognize texts in images quickly and identify the script of the localized text. The YOLO recognizes the text by drawing a bounding box around it, classifies the script, and shows the confidence score for each detection. The results obtained from the model are displayed below. The test images obtained from the *mixed data* with Hyperparameter tuning are as follows:

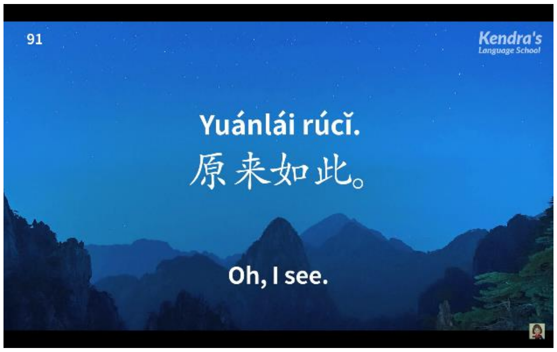


Figure 2. Sample image from mixed data

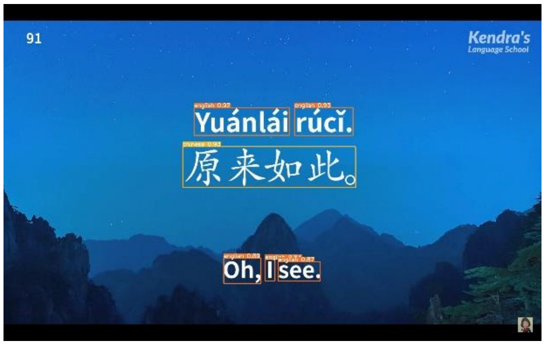


Figure 3. Result image from trained YOLOv5

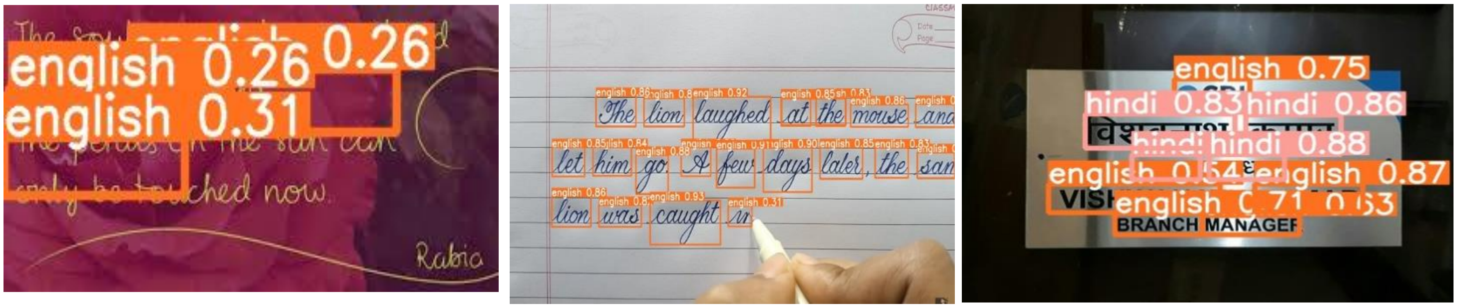


Figure 4. Result images from the Trained YOLOv5 model

The images obtained from the *cursive data* with Hyperparameter tuning are displayed below.

Figure 5. Sample image

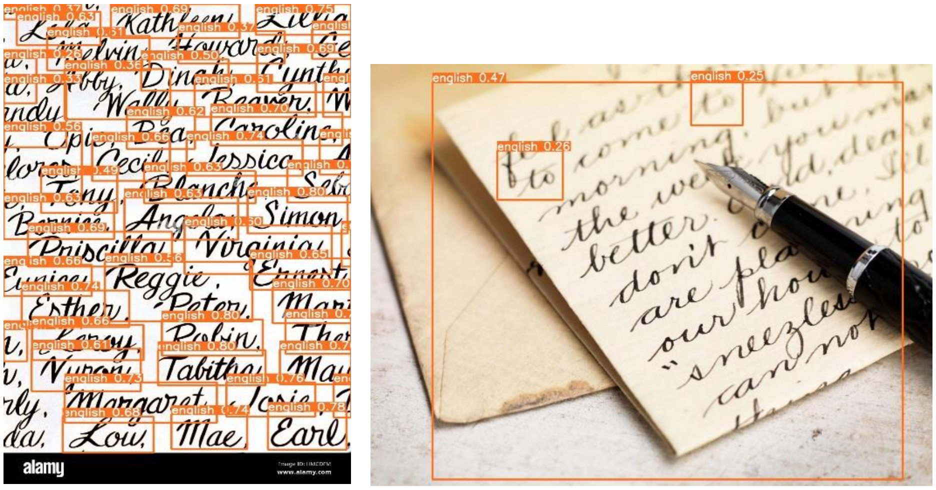


Figure 5. Result images from the Trained YOLOv5 model

The Fig below represents the precision-confidence curve for Kannada, Hindi, English, and Chinese 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 increase for both mixed and cursive data.

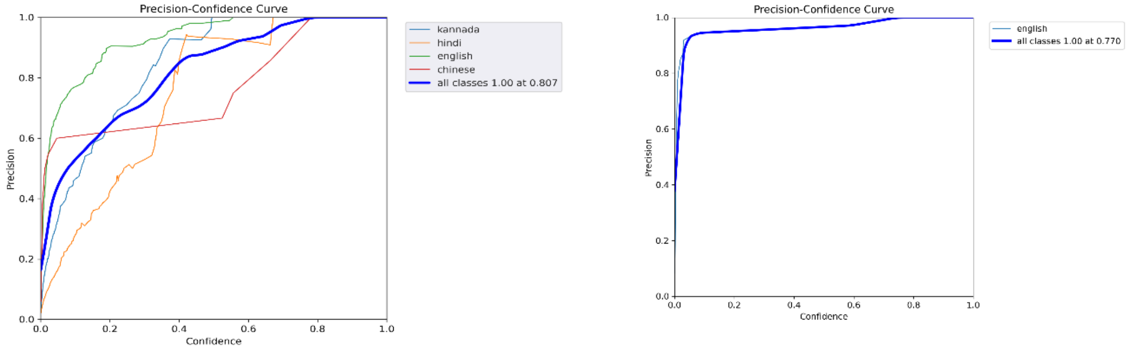


*Figure 6. P curve from mixed data and English cursive data*

The diagram below depicts the recall-confidence curve for 100 epochs; it can be observed that as the confidence score climbs from 0 to 1, the recall declines.

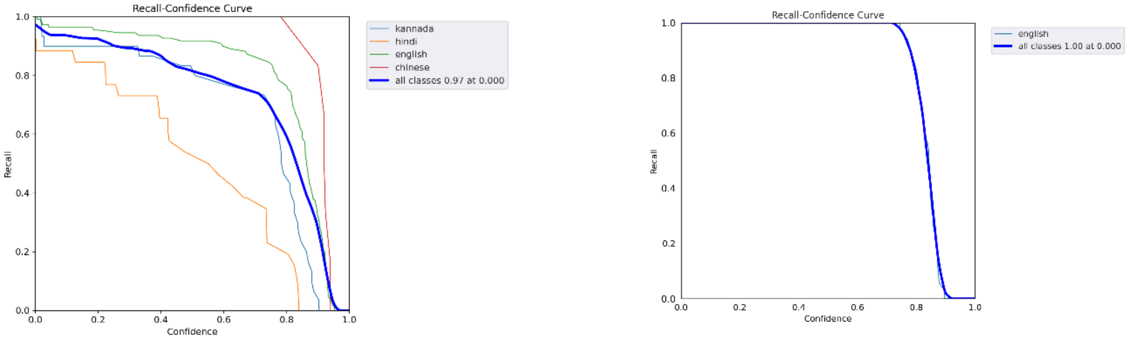


*Figure 6. R curve from mixed data and English cursive data*

A better Precision-recall curve, commonly known as a PR curve, has a larger AUC (area under the curve). The graph depicts the PR curve for Kannada, Hindi, English, and Chinese scripts, as well as all classes, at mAP 0.5. AUC that is larger, i.e., close to 1, indicates better detection outcomes.

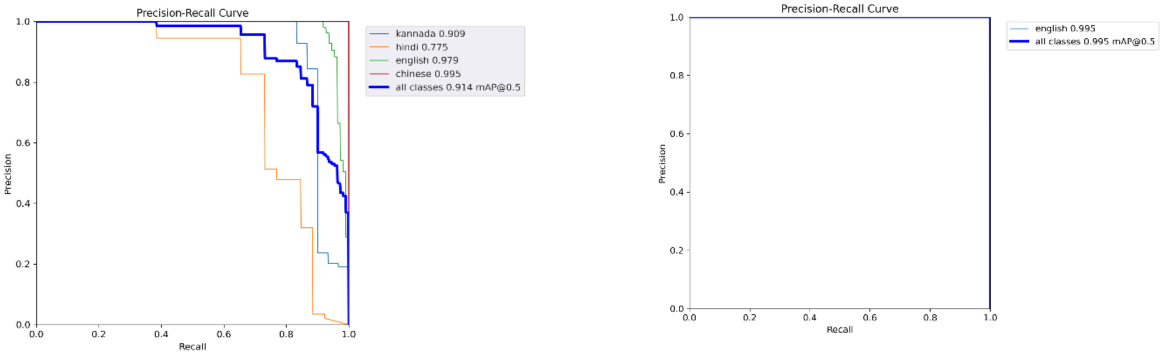


*Figure 6. PR curve from mixed data and English cursive data*

The F1 score is an average of Precision and Recall, it gives Precision and Recall equal weight: If both Precision and Recall are high, a model will have a high F1 score. In most cases, a higher confidence value and F1 score are desirable.

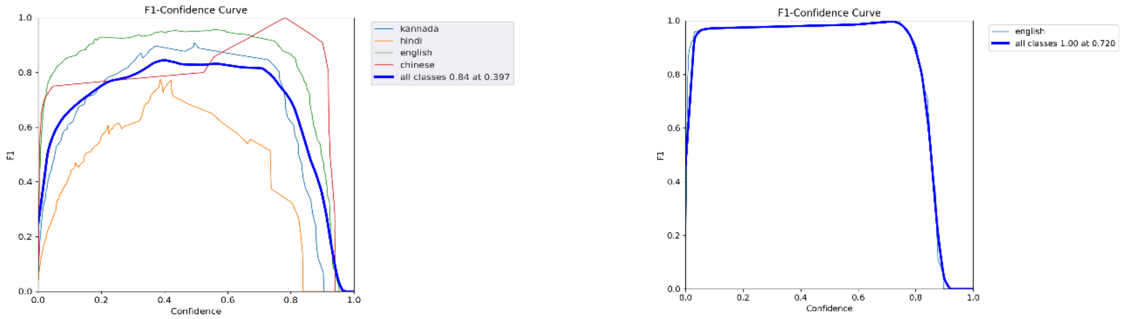


*Figure 7. F1 curve from mixed data and English cursive data*

The confusion matrix for the trained YOLOv5 model on mixed data and cursive data is shown below.

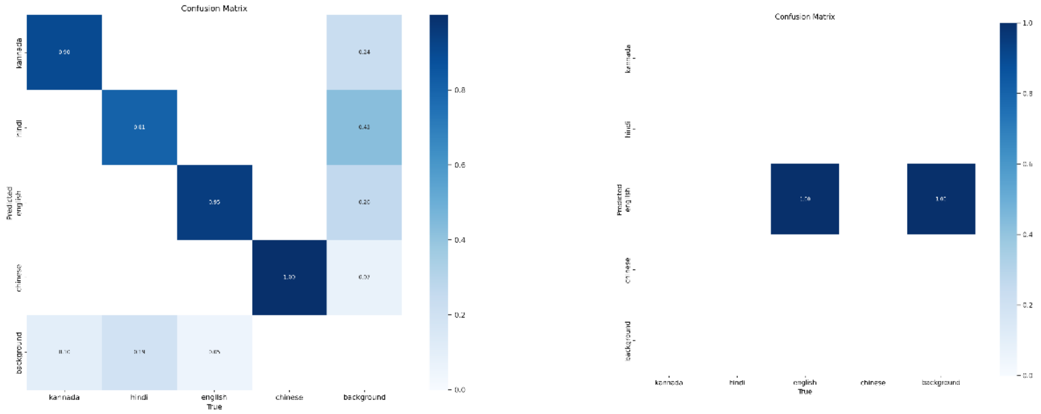


*Figure 7. F1 curve from mixed data and English cursive data*

# Conclusion:

We employed the most efficient object detection algorithm YOLOV5 for text localization and script recognition in natural scene photos using the DarkNet53 Architecture in this project. Despite the complexity in the photos and the nuances in the curves of Kannada, Hindi, English, and Chinese characters, particularly the cursive data contained in the dataset, the suggested model correctly detects the text and identifies the script. The model has been evaluated on some of the most difficult scenarios in text identification in natural scene photographs, such as varied image backgrounds, text orientations, font styles, resolutions, and lighting conditions, which make text detection in natural scene images more difficult. The model is further fine-tuned using Hyperparameter tuning to increase accuracy.

Based on the increased dataset and the tuning of a few parameters, the suggested model yields overall accuracy of 95%-99%. However, there is still room for improvement in text detection and script recognition in natural scene images, which is found in many important applications, such as text extraction and subsequent translation to user-choice languages, which would help tourists and the visually impaired have a better understanding of the world around them.

## Member Contribution:

Apoorva Udupa played a crucial role as the main contributor in our project. Her contributions encompassed various essential tasks, including data set generation, data labeling, hyperparameter tuning, and report writing. Apoorva's meticulousness in collecting and curating diverse data sets, accurate data labeling, optimizing model performance through hyperparameter tuning, and effectively communicating our findings through well-written reports greatly elevated the overall success and impact of our project.

Sagar Regmi contributed in dataset generation and report writing

## References:

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

[2] 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.

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

[4] 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.

[5] 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.

[6] Chandana Udupa, Anusha Upadhyaya et. al, “Text Localization and Script Identification in Natural Scene Images and Videos”, 2022 International Conference on Connected Systems & Intelligence (CSI), 1 December 2022.

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