Vol. 06, No. 1 (2024) 373-382, doi: 10.24874/PES.SI.24.02.020

Proceedings on Engineering Sciences

www.pesjournal.net

LOCALISATION OF TEXT IN A NATURAL SCENE IMAGE BASED ON DEVANAGARI SCRIPT RULE USING DEEP LEARNING

Vijay Prasad[1]
Pranab Das
Y. Jayanta Singh

Received 06.11.2023.
Received in revised form 25.12.2023.
Accepted 04.01.2024.
UDC – 004.418

Keywords:

AdaBoost, Devanagari script, ICDAR 2015, SVM, Text localization, vpdDataset, YOLO-v7

A B S T R A C T

Text localization play a crucial role in scene images to read the text content effectively. There have been built many deep learning (DL) rooted models for classification as well as localization of manifold language text in scene images in recent years. Though, after a thorough review it is explored that there is very less research conducted on text localization based on the Devanagari script rule. However, previously developed models for text classification and localization have a few setbacks namely lower performance metrics, hardship in text localization with multi-scaling shapes, intricacy in handling irregular text etc. This research presents a novel framework which is developed for the localization of text in a natural scene image based on the Devanagari script rule using deep learning. This framework initially pre-processes the input datasets for standardization. Secondly, it integrates the enhanced YOLO-v7 for candidate component detection, improvised support vector machine (SVM) and AdaBoost as a hybrid text classifier model. Lastly, text segmentation and clustering are done using balanced iterative reducing and clustering using the hierarchies (BIRCH) algorithm and transfer learning is used in the localization model. This proposed model obtains accuracy, precision, recall and F1 score on vpdDataset as well as ICDAR 2015 datasets 98.33%, 99.01%, 99.05%, 97.76%, and 98.09%, 98.02%, 99.03%, 97.46%, respectively. Therefore, the obtained findings of this proposed model are very optimal and enhanced in comparison with previous text detection and localization models.

1. INTRODUCTION

Text localization in scene images particularly for the Devanagari script is becoming highly competitive and challenging work in the present scenario as scene text attributes are extremely distinct from other kinds of scanned documents in terms of different aspects such as type, font size or style (Akallouch et al., 2022; Gao et al., 2019; Mahajan & Rani, 2021). Therefore, specific approaches are needed for text content localization as well as recognition. In fact, scene image text localization is one of the complicated yet essential works as there is not any previous information accessible on layout, position, dimension, direction colour and typeface of text content within scene pictures

Corresponding author: Vijay Prasad
Email: vpd.vijay82@gmail.com

(Francis & Sreenath, 2022; Uchida, 2014). Also, there are numerous patterns as well as textures very identical to the characters due to which localization of text is intricate (Long et al., 2021). In addition to this, natural scene text localization and recognition are competitive owing to the multifarious characters blurring effects, unequal lighting conditions, perspective as well as lower resolution conditions. The decorated characters or words of Devanagari text also make the localization task more complicated (Bagi & Dutta, 2021; Bisht & Gupta, 2023).

At present, the majority of text content is being captured with the aid of various types of cameras in comparison to scanners. Digital camera-rooted optical character recognition (OCR) is an innovation employed to identify all characters of the script clicked by cameras (Kaur et al., 2017; Zhao et al., 2022). Therefore, in this technological era, smartphones are more handy and easy to carry than scanners. The smartphone cameras are capable enough for capturing multifarious text content from natural scene images namely signboard characters, roadside hoardings etc (Rong et al., 2020). For an optimal solution to the above-described intricacy, researchers have to not only metamorphose the conventional image processing techniques besides explore and integrate the novel pattern identification methodologies for scene text content localization as well as recognition through the scene pictures (Ali & Hashim, 2016; Devi & Kumar, 2021).

Natural scene text detection as well as localization are receiving huge attention from both academia as well as industry. This is one of the vital areas for research in computer vision. Owing to the expansion of deep learning as well as mobile-based IoT (Internet of Things), text identification studies have made pragmatic progress in recent years (Udupa et al., 2022; Vaidya et al., 2020). There are reviewed extensive works on scene text localization and identification in scene images during the last decade for summarization of key threats and noteworthy development in natural scene text identification. In (Gupta & Jalal, 2022), N. Gupta et al. introduced the historical evaluation as well as a significant development of scene text identification and categorizes deep learning-based techniques and conventional techniques in detail along with key threats. Further, there is introduced the most common benchmark datasets used in previous work. Lastly, this study summarizes and forecast possible future research route for researchers. In (DASARI & Mehta, 2022), S. K. Dasari et al. developed a hybrid CNN framework for text categorization as well as identification through scene pictures. Accurate detection of the text content from the varied captured picture format is a massive threat owing to various factors namely the colour contrast, dimensions, or blurred background. Nevertheless, the proposed framework offers lower metrics in the complex background and uneven light settings. To solve issues of earlier text detection and localization models, this work presents a new text localization model for natural scene images based on the Devanagari script rule using deep learning.

The contribution of this research is highlighted as follows:

- One of the main focuses of this research is to develop a model for text localization in natural scene images based on the Devanagari script rule using the deep learning method.
- Another objective of this research is to formulate a novel publicly accessible benchmark dataset namely vpdDataset and standardization of it for performance enhancement of the proposed model.
- This novel suggested DL-based model is built using diverse algorithms namely enhanced YOLO-v7 for candidate component detection, improvised SVM and AdaBoost in text classification, BIRCH algorithm for segmentation and clustering and transfer learning in text localization model training.
- Furthermore, the proposed model is validated using two distinct datasets namely the vpdDataset and ICDAR 2015 for performance computation and validation. It is found that the suggested model obtained extremely competitive performance in comparison with previously built models.
- Moreover, this suggested model provides all the performance metrics very optimal, hence it can be implemented in modern text detection and localization systems taking into account natural scene images.

This manuscript is structured in various sections that are described herein. Section 1 narrates the introduction part along with the major contributions of this research work. Section 2 explicates the literature review on text detection, classification as well as localization methods. Section 3 presents the proposed methodology and provides thorough details of implementing work. Section 4 narrates the detailed results and discussion. The conclusion of the research is given in section 5.

2. LITERATURE REVIEW

Scene text localization and identification from scene pictures is a difficult task (Naiemi et al., 2021b; Rong et al., 2022). Text content identification from scene pictures is explored by numerous feature extraction classifiers and techniques (He et al., 2021; Naiemi et al., 2021a; Prasad & Das, 2021). Investigators utilized statistical, structural as well as topological features. There is explored some of the most recognized classifiers for scene image identification using K-Nearest Neighbours (KNN) (Khan et al., 2021), CNN (Cao et al., 2020), SVM (Lin et al., 2020), Naive Bayes (Francisca O Nwokoma et al., 2021) etc. In (Soni et al., 2019), R. Soni et al. explored an approach for natural scene image text content identification as well as localization through the text awareness method. Text identification as well as localization from scene pictures is a very effective approach to obtaining text content which may be utilized in licence plate identification, robot

navigation, as well as wearable application etc. This research proposed a novel text identification and localization scheme which is rooted on text awareness score. However, suggested text identification and localization scheme incapable to recognize the text accurately in complex background settings of scene images and varied font size scenario. In (Chaitra et al., 2022), Y. L. Chaitra et al. developed a text localization approach based on deep convolution neural network (DCNN) as well as transfer learning. Effective localization of scene images text is very significant to read text information correctly. This is one of the complicated process for localization of text content in scene pictures as scene images contains text in the scattered form. This proposed model is implemented using the VGG16 architecture. However outcome of this text localization model demonstrates that it achieves less value of F1-score in terms of model performance. Moreover, the suggested model incapable to localize text in Non-uniform illumination settings.

In (Shiravale et al., 2020), S. S. Shiravale et al. explored a scheme for Devanagari text identification using scene images. Text content available in scene pictures which are captured by digital cameras may be utilized in image analysis. Automated text identification, extraction as well as recognition is fundamental in scene image comprehension applications. In this work, for text identification from scene images, a new framework is disclosed which involves a coloured-rooted clustering approach along with edge detection. For model training, novel scene picture datasets have been formulated containing 1250 pictures. However, this suggested scheme for text identification offers less accuracy in multi-coloured as well as complex background settings. In (K. Bawa & K. Sethi, 2014), R. K. Bawa et al. explored a new binarization method for Devanagari text identification using camera-rooted pictures. This developed approach combines scene image morphological dilation as well as canny edge detection. The images used in the binarization process are performed utilizing the standard deviation and mean of the overall edge pixels value. However, this binarization method is resource-intensive as well as more changing to implement.

In (Shiravale et al., 2021), S. S. Shiravale et al. discussed SVM and stroke width transformation-based methods for the identification of text content areas within scene pictures of Indian streets. In scene pictures, the detection of text content is highly critical due to the complexity of varied script attributes. This method is developed for the identification as well as recognition of Devanagari script text content through scene pictures. However, this suggested text recognition method offers less accuracy in lighting variation settings as well as is time-consuming. In (Jangid & Srivastava, 2018), M. Jangid et al. explored a DCNN and adaptive gradient technique-based framework for handwritten character identification of the Devanagari script. In the modern digital world, handwritten character identification has gained massive attention from academicians owing to distinct application areas, for instance, human-robot communication, visually disabled individuals etc. Nevertheless, the presented DCNN and adaptive gradient technique-based framework requires post-processing for performance enhancement which makes the training process highly time taking as well as complex.

3. PROPOSED METHODOLOGY

3.1. Datasets

Text localization is a procedure to identify as well as localize text content within video frames or images. This is a very significant phase in the optical character recognition (OCR) system that is aimed to identify as well as extract text effectively from the images. This research is aimed to suggest a new framework for the localization of the text regions in scene pictures having a complicated background. The proposed text localization model is highly pragmatic in terms of the localization of text in a natural scene picture based on the Devanagari script. In this, research work two distinct datasets namely vpdDataset and ICDAR 2015 have been used for performance computation of this suggested text identification as well as localization model.

3.1.1. vpdDataset

For this research study, a new dataset namely the vpdDataset has been created. The performance standardization of entire datasets is carried out for effective training and validation of this text detection and localization framework. This proposed dataset contains a variety of natural scene images containing the Devanagari script.

3.1.2. ICDAR 2015

For cross-examination of the performance metrics of the suggested text identification as well as localization model, there have been selected another dataset namely the ICDAR 2015. This dataset contains 500 testing and 1000 training images. The ICDAR 2015 includes natural scene images and is developed for text identification and recognition tasks (Cao et al., 2021).

3.2. Sample Size

Table 1 shows the selected vpdDataset and ICDAR 2015 datasets summary. For this proposed text detection and localization model training and testing utilizing the vpdDataset overall 528 natural scene images have been selected. Furthermore, these scene images contain an average number of words of 15.3, a total number of words of 3028 and several letters containing 14639, respectively. However, from the ICDAR 2015 dataset, there are selected 455 natural scene images for training and testing of this model. The selected scene images of

the ICDAR 2015 dataset contain an average number of words of 11.6, a total number of words of 2252 and several letters containing 11333, respectively.

Table 1. Selected vpdDataset and ICDAR 2015 datasets summary

Datasets Details	Overall Images	The average number of words	Total number of words	Number of letters
vpdDataset	528	15.3	3028	14639
ICDAR 2015	455	11.6	2252	11333

3.3. System configuration

This text detection and localization model is implemented with a personal computer which has the under-mentioned arrangement: 12th Generation Intel Core i7 processor (12MB cache, 10 cores and 12 threads), 16GB RAM, Windows 11 and NVIDIA RTX 3050 GPU unit. The experiments on both selected datasets i.e., vpdDataset and ICDAR 2015 have been performed using the Tensor Flow module which is based on Keras 2.0, OpenCV2 as well as Python 3.11. The backbone has been fine-tuned through the vpdDataset pre-trained model along with dataset augmentation.

3.4. System architecture

This research, there is proposed and implemented a new system for text identification as well as localization within scene pictures based on the Devanagari script rule using deep learning techniques for enhanced performance assessment. Figure 1 illustrates the architecture of the suggested text localization hybrid model.

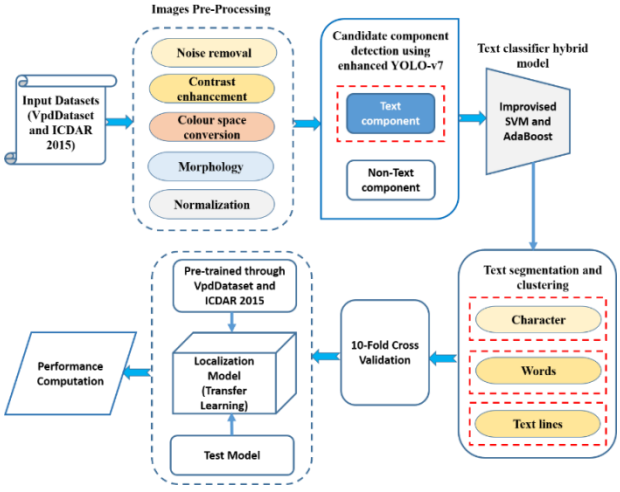

Figure 1. Architecture of proposed hybrid model

In this research, the proposed model is implemented using two distinct datasets namely vpdDataset and ICDAR 2015 for cross-validation of the proposed model performance. Initially, the proposed vpdDataset is standardized using enhanced image pre-processing. In this step, multiple internal pre-processing steps are performed for making the dataset suitable and standardized for improved outcomes in training and validation. The image pre-processing involves multiple operations on the image datasets namely noise removal, contrast enhancement, colour space conversion, morphology and normalization.

Removal of noise is very essential for smoothening the image through the elimination of minor patches or dots from the pictures. For noise removal adaptive filtration approach is utilized. The contrast enhancement process is used to improve image quality by enhancing the contrast level between the diverse parts of an image. Colour space conversion is one of the vital phases in image pre-processing which involves the conversion of pictures from one colour space to another. The morphology operation is used to process chosen images following varied shapes of images to make the images appropriate for the training phase. The image normalization process is employed for adjusting the image pixel values to a common scale. The chosen pre-processing steps make the datasets standardized effectively. Further, the candidate component detection process is performed using the enhanced YOLO-v7. It is a DL-based algorithm which utilizes the convolutional neural network (CNN) for real-time identification of the Devanagari script through scene pictures. Devanagari text identification in the YOLO-v7 algorithm is done as a regression problem and offers the class likelihood of detected scene images. This YOLO-v7 algorithm is opted for because it needs only a single forward propagation via a CNN to detect the Devanagari script from scene images, thus minimising the computational complexity in implementation. This enhanced YOLO-v7 algorithm performs the candidate component identification in terms of text components as well as non-text components.

In the next phase, a text classifier hybrid model based on the improvised SVM and AdaBoost has been implemented for effective classification. This classifier hybrid model improves the Devanagari script classification and boosts the overall performance as well as minimizes the time in model training and testing. After classification, the text segmentation and clustering operation is performed. In this model, text segmentation and clustering are done using balanced iterative reducing and clustering using the hierarchies (BIRCH) algorithm. It has been selected for clustering the incoming, varied-size metric dataset within an incremental order as well as dynamically generating the optimum clustering for scene images. It is one of the finest algorithms in terms of memory utilization as well as time constraints. The clustering is done according to classified scene images involving the character, words as well as text lines. In the next phase, the cross-validation of the dataset is done using the K-fold cross-validation in which the value for K is

taken 10. This cross-validation approach split the dataset into the train and test data in a ratio of 80:20. In the later stage localization model is trained using transfer learning along with the pre-trained through the vpdDataset and ICDAR 2015. The transfer learning algorithm is integrated because of the higher learning rate, minimal need for datasets and saving the resources in the training and testing process for performance evaluation of this model.

Pseudo Code: This pseudo-code is utilized for expressing the design of the program in a pragmatic way and offers a detailed program template. The steps of this pseudo-code are presented as follows:

Input: vpdDataset and ICDAR 2015

Output: $A_M, P_M, R_M, and\ F1_M$.

Step 1: To initialize the model for aggregation of the vpdDataset and ICDAR 2015 datasets.

Step 2: To upload the vpdDataset and ICDAR 2015 for standardization and to obtain $VpD_1\ and\ ICD_2$.

Step 3: Encore the previous step until $VpD_1\ and\ ICD_2$ are in specific-goals $SG_1 and\ SG_2$.

Step 4: To initiate the candidate component CD_C detection by YOLO-v7 for segregation of text component and non-text component i.e., $Tx_C\ and\ NTx_C$.

While,

Text components are detected proceed further,

Otherwise,

Terminates and go to the previous step.

Step 5: To apply all Tx_C in text classifier hybrid model and for obtaining the HS_{Ada}.

Step 6: To input the HS_{Ada} in text segmentation and clustering module for getting the segregated data clusters represented by $\{Ca_1, Ca_2, Ca_3, Ca_4, \dots\dots\dots\dots Ca_n\}$.

While,

Clusters $\{Ca_1, Ca_2, Ca_3, Ca_4, \dots\dots\dots\dots Ca_n\}$ are structured to proceed further

Otherwise

Terminates and go to the previous step.

Step 7: To split the aggregated $\{Ca_1, Ca_2, Ca_3, Ca_4, \dots\dots\dots\dots Ca_n\}$ in train and test groups $G_{TR}, and\ G_{TT}$.

Step 8: System initialized for training on G_{TR} using transfer learning.

Step 9: System initialized for training on G_{TT} using transfer learning.

While,

The system is trained and tested in defined goals $SG_1 and\ SG_2$, and iterations proceed further.

Otherwise,

Terminates and repeat the previous step 8 and step 9.

Step 10: To compute the $A_M, P_M, R_M, and\ F1_M$.

3.5. Performance metrics:

In the training as well as the testing approach of this text localization scheme network weights are updated in each epoch and the end objective of the training procedure is to obtain optimal weight which provides minimal error for improved performance. During model training, the optimal learning rate has been considered a noteworthy parameter that chooses alteration in network weights. The correct choice of learning rate is a crucial task in the model training procedure because if the learning rate parameter is selected lower the case overall optimization process may become slower and the network can consume much time in terms of finding the minimal error. While the learning rate value chosen is greater in that case the overall optimization process would diverge, thereby an optimal selection of the learning rate becomes very essential for the performance computation of this proposed text localization model.

The loss function value in YOLO-v7 may be accessed through equation 1. It is an aggregation of bounding-box regression losses, confidence-losses as well as object identification-losses.

$$Loss_T = Loss_B + Loss_C + Loss_{CL} \quad (1)$$

The classification as well as confidence loss may be described as follows:

$$Loss_C = -\sum_{i=0}^{P\times P}\sum_{j=0}^{C} Ia_{ij}^{Objc}\,[Ca_i \log(Ca_i) + (1-Ca_i)\log(1-Ca_i)] - \lambda_{no-objc}\sum_{i=0}^{P\times P}\sum_{j=0}^{C} Ia_{ij}^{no-objc}[Ca_i \log(Ca_i) + (1-Ca_i)\log(1-Ca_i)] \quad (2)$$

$$Loss_{CL} = -\sum_{i=0}^{P\times P} Ia_{ij}^{objc}\sum_{C\in CL}[Pa_i \log(Pa_i) + (1-Pa_i)\log(1-Pa_i)] \quad (3)$$

Where, $Loss_T = text\ loss$, $Loss_B = box\ loss$, and $Loss_{CL} = class\ loss$

This text localization model performance is computed through different metrics that are defined herein.

Accuracy performance metrics are represented in Equation 4. It is the ratio of accurate prediction of the model and overall prediction. In equation 4, TP = True

Positive, and TN =True Negative, as well as FN = False Negative and FP = False Positive.

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \quad (4)$$

Precision metrics are represented in Equation 5. It is the ratio of the TP prediction of the model and overall positive predictions. Herein, TP = True Positive, and FP =False Positive.

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

The recall metric is represented in Equation 6. It is the ratio of TP predictions of the model as well as total real samples. In this equation 6, TP = True Positive and FN = False Negative.

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

F1-Score is represented in equation 7. It integrates recall as well as precision in one united metric.

$$F1 - Score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (7)$$

In this proposed text localization model, certain enhancements based on YOLO-v7 have been made to allow correct text localization in scene pictures wherein text content is randomized and of altered scale etc. For improving this text localization model accuracy along with other metrics, multiple algorithms namely SVM, AdaBoost, BIRCH, and transfer learning have been integrated in various phases.

4. RESULTS AND DISCUSSION

In the modern digital world, effective text identification as well as localization schemes are becoming famous in the arena of image text analysis models owing to their applicability within numerous applications namely assistive systems for visually disabled individuals, smartphone transliteration technologies and many more. The text localization approaches are very significant and useful in finding the exact position of specific text components area within the scene images. Digital cameras are gaining popularity as they can be mounted easily within a variety of devices namely smartphones, tablets as well as other handheld devices and are capable to capture images of distinct objects in natural scenes while a person travels.

Due to the rising prominence of automated digital gadgets like a smartphone as well as huge image expansion, element-based image analysis is gained massive attention from academicians. It has become a vital element for computer vision rooted application namely element-rooted retrieval of images, assistive technology as well as reading aids and many more. There is built various text detection and localization models using deep learning methods in recent years. Nevertheless, these models employed in real-time applications for text localization are still ineffective in terms of some performance metrics like accuracy, F1 score etc. Furthermore, another key issue related to the performance of previous models is due to a lack of standardization of the accessible benchmark datasets. Therefore, this research is focused to build a model for the localization of text within the scene pictures based on the Devanagari script rule using deep learning. This model initially standardized the proposed vpdDataset and performs the text localization task.

Table 2 represents details of the system arrangement. This proposed text localization model was implemented using a personal computer that involves the under-mentioned arrangement: 12th Generation Intel Core i7 processor (12MB cache, 10 cores and 12 threads), 16GB RAM, Windows 11 and NVIDIA RTX 3050 GPU unit. The experiments on both selected datasets i.e., vpdDataset and ICDAR 2015 have been performed using the Tensor Flow module which is based on Keras 2.0, OpenCV2 and Python programing language. Table 3 represents chosen hyperparameter details in model training and testing. For this text localization model training, a number of epochs were taken at 26 and RMSProp optimizer has been utilized. Furthermore, the number of filters, batch size, dropout value and activation function are 20, 15, 0.3, and Sigmoid [0, 1], respectively.

Table 2. Represents system arrangement details.

S. No.	System Arrangement	Details
1	Processor	12th Generation Intel Core i7 processor (12MB cache, 10 cores and 12 threads)
2	RAM	16 GB, DDR4 having a clock-speed of 3200MHZ
3	GPU (Graphic Processing Unit)	NVIDIA RTX 3050
4	Frameworks	Tensor Flow (Keras 2.0), OpenCV2
5	Programming	Python 3.11.3

Table 3. Illustrates the chosen hyperparameter details in model training and testing.

S. No.	Hyperparameters	Details
1	Number of Epochs	26
2	Optimizer	RMSProp
3	Filters	20
4	Batch-Size	15
5	Dropout value	0.3
6	Activation function	Sigmoid [0, 1]

Figure 2 depicts text localization accuracy (a) shows a standardized image of vpdDataset before training (b) shows scene image localized text by the proposed

model (c) depicts another standardized image of vpdDataset before training and (d) presents scene image localized by the proposed model. It is noticeable from Figure 2 that the proposed text localization model accurately captures the words completely and highlights them through a bounding box.

Figure 2. Text localization accuracy (a) shows a standardized image of vpdDataset before training (b) shows scene image localized text by the proposed model (c) depicts another standardized image of vpdDataset before training and (d) presents scene image localized by the proposed model

Figure 3 presents the computed training and testing accuracy on vpdDataset for the proposed model. This text localization model training accuracy on epochs 2, 5, 8, 11, 14, 17, 20, 23 and 26 is measured at 87.22%, 88.11%, 89%, 89.89%, 90.78%, 91.67%, 93.56%, 95.45%, and 97.34%, respectively. Furthermore, the text localization proposed model testing accuracy on epochs 2, 5, 8, 11, 14, 17, 20, 23 and 26 is measured at 88.34%, 89.77%, 91.2%, 92.63%, 94.06%, 95.49%, 96.92%, 98.35%, and 98.33%, respectively. Both the training and test accuracy of this text localization model is very competitive and improved on the proposed vpdDataset.

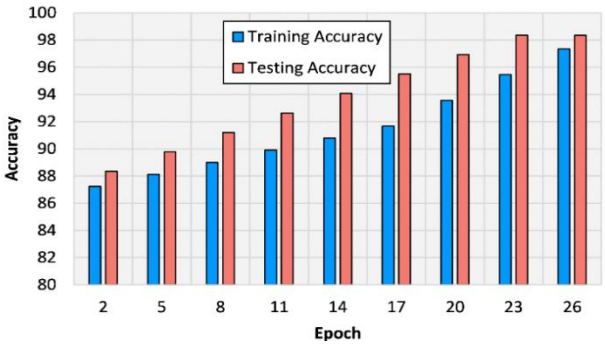

Figure 3. Computed training and testing accuracy on vpdDataset

Figure 4 presents an average accuracy comparison of the suggested model on vpdDataset along with previous research. The accuracy of M. P. Kantipudi et al. (Kantipudi et al., 2021), S. V. Seeri et al. (Seeri et al., 2016), and A. A. Chandio et al. (Chandio et al., 2020), models was 93.83%, 89.90 and 98.13%, respectively. This proposed text detection and localization model provides 98.33%, accuracy on vpdDataset which is more competitive in comparison with previous work. Moreover, this proposed text localization model provided improved accuracy and was highly robust in terms in terms of handling computational costs.

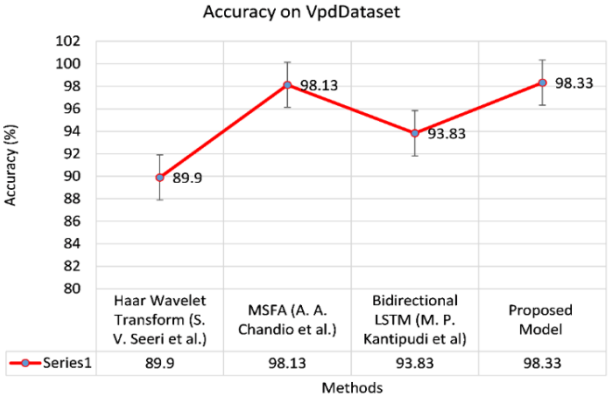

Figure 4. Average accuracy comparison on vpdDataset (Chandio et al., 2020; Kantipudi et al., 2021; Seeri et al., 2016)

Figure 5 depicts the measured error rate in the training and testing process of the suggested text localization model. A measured training error rate of the proposed text localization model on epochs 2, 5, 8, 11, 14, 17, 20, 23, and 26 is 2.9%, 2.4%, 2.2%, 1.8%, 1.4%, 1.1%, 0.9%, 0.5% and 0.3%, respectively. Furthermore, the testing error rate of the proposed text localization model on epochs 2, 5, 8, 11, 14, 17, 20, 23, and 26 are measured at 2.6%, 2.2%, 2%, 1.7%, 1.3%, 1%, 0.8%, 0.3%, and 0.1%, respectively. Thus, this proposed text localization model offers enhanced metrics and a minimal error rate in training and testing.

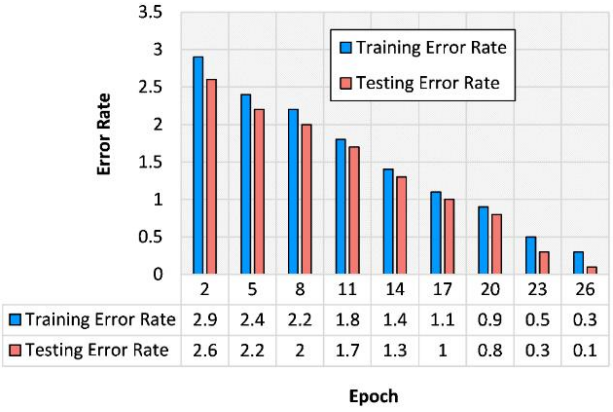

Figure 5. Measured error rate in training and testing

Figure 6 presents the performance analysis of proposed and existing methods in terms of various evaluation metrics. To compute the performance of the deep learning framework,

evaluation metrics are required highly optimized. The assessed value of precision, recall and F1-score of M. P. Kantipudi et al. method (Kantipudi et al., 2021), is 90.18%, 98.19%, and 97.07%, respectively. Further, the assessed value of precision, recall and F1-score of S. V. Seeri et al. (Seeri et al., 2016) method is 98.85%, 90.85%, and 94.68, respectively. The assessed values of precision, recall and F1-score for another model explored by A. A. Chandio et al. (Chandio et al., 2020) are 90%, 91%, and 91%, respectively. However, the suggested text localization model achieves the precision, recall and F1-score values on vpdDataset, 99.1%, 99.05%, as well as 97.76%, correspondingly. Hence, this proposed text localization model obtained all metrics highly optimized and competitive in comparison to previous work.

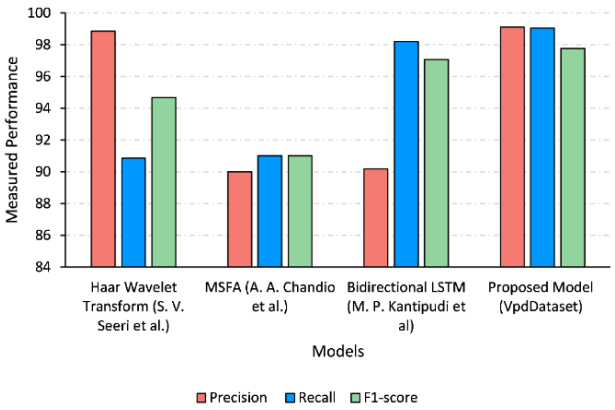

Figure 6. Performance analysis of proposed and existing methods considering precision, recall and F1-score (Chandio et al., 2020; Kantipudi et al., 2021; Seeri et al., 2016)

Table 4. Performance metrics on ICDAR 2015 benchmark Dataset

S. No.	Performance Metrics	Measured Values
1	F1-score	97.46
2	Recall	99.03
3	Precision	98.02
4	Accuracy	98.09

Table 4 illustrates the performance metrics of the proposed text localization model on the ICDAR 2015 benchmark Dataset. The assessed value of this text localization model on the ICDAR 2015 benchmark dataset considering the F1-score, recall, precision and accuracy is obtained at 97.46%, 99.03%, 98.02%, and 98.09%, respectively. All the performance metrics on ICDAR 2015 dataset were also found very competitive in comparison to the existing works.

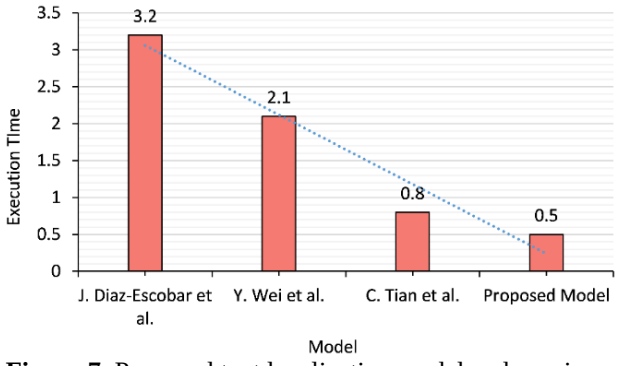

Figure 7. Proposed text localization model and previous methods of time consumption assessment (Diaz-Escobar & Kober, 2020; Tian et al., 2017; Wei et al., 2018)

Figure 7 presents the proposed text localization model and previous methods of time consumption assessment. Previous work of J. Diaz-Escobar et al. (Diaz-Escobar & Kober, 2020), Y. Wei et al. (Wei et al., 2018), and C. Tian et al. (Tian et al., 2017), takes time during model execution 3.2s, 2.1s, and 0.8s. While this scene text localization model takes time in model training and testing only 0.5s. Therefore, it is apparent from Figure 7, that this suggested text localization model is highly competitive and requires minimal time.

5. CONCLUSION

The efficient and correct text localization in scene images especially the Devanagari script is vital yet intricate in computer vision. In the last decade, several text localization as well as recognition schemes is presented. However, the developed text localization methods do not satisfy the real-world need for text localization with scene pictures or video frames particularly for the Devanagari script as well as not competitive and accurate. So, this research explores a new text localization model for natural scene pictures based on the Devanagari script rule using deep learning. In this work, a new dataset namely vpdDataset has been created and standardized. This text localization model involves enhanced YOLO-v7 in candidate component detection of the Devanagari script. Classification is carried out utilizing the improvised SVM and AdaBoost, segmentation and clustering are done by the BIRCH algorithm and transfer learning is used in localization model training. This text localization model achieves optimized and competitive performance metrics involving precision, recall, F1-score and accuracy on proposed vpdDataset and ICDAR 2015 Benchmark datasets, 99.1%, 99.05%, 97.76%, 98.33% and 98.02%, 99.03, 97.46%, 98.09%, respectively. Therefore, it is explicit from the outcomes of this proposed text localization model that all metrics are received very optimal. Moreover, this text localization model attains less time in execution in comparison to previous work.

Acknowledgement: Authors acknowledge the immense help received from the scholars whose articles are cited and included in references to this manuscript. The authors are also grateful to authors/ editors / publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

References:

Akallouch, M., Boujemaa, K. S., Bouhoute, A., Fardousse, K., & Berrada, I. (2022). ASAYAR: A Dataset for Arabic-Latin Scene Text Localization in Highway Traffic Panels. *IEEE Transactions on Intelligent Transportation Systems*. DOI: https://doi.org/10.1109/TITS.2020.3029451

Ali, S. A., & Hashim, A. T. (2016). Wavelet transform based technique for text image localization. *Karbala International Journal of Modern Science*. DOI: https://doi.org/10.1016/j.kijoms.2016.03.004

Bagi, R., & Dutta, T. (2021). Cost-Effective and Smart Text Sensing and Spotting in Blurry Scene Images Using Deep Networks. *IEEE Sensors Journal*. DOI: https://doi.org/10.1109/JSEN.2020.3024257

Bisht, M., & Gupta, R. (2023). Handwritten Devanagari Word Detection and Localization using Morphological Image Processing. *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, 126–130. DOI: https://doi.org/10.1109/SPIN57001.2023.10116577

Cao, D., Dang, J., & Zhong, Y. (2021). Towards accurate scene text detection with bidirectional feature pyramid network. *Symmetry*. DOI: https://doi.org/10.3390/sym13030486

Cao, D., Zhong, Y., Wang, L., He, Y., & Dang, J. (2020). Scene Text Detection in Natural Images: A Review. *Symmetry*, *12*(12), 1956. DOI: https://doi.org/10.3390/sym12121956

Chaitra, Y. L., Dinesh, R., Gopalakrishna, M. T., & Prakash, B. V. A. (2022). Deep-CNNTL: Text Localization from Natural Scene Images Using Deep Convolution Neural Network with Transfer Learning. *Arabian Journal for Science and Engineering*. DOI: https://doi.org/10.1007/s13369-021-06309-9

Chandio, A. A., Asikuzzaman, M., & Pickering, M. R. (2020). Cursive Character Recognition in Natural Scene Images Using a Multilevel Convolutional Neural Network Fusion. *IEEE Access*, *8*, 109054–109070. DOI: https://doi.org/10.1109/ACCESS.2020.3001605

DASARI, S. K., & Mehta, S. (2022). Scene Based Text Recognition and Classification Based on Hybrid Cnn Models with Performance Evaluation. *SSRN Electronic Journal*, 293–300. DOI: https://doi.org/10.2139/ssrn.4174796

Devi, R., & Kumar, B. (2021). Recent Trends in Text Region Identification and Localization in Native Surrounding Images. *Proceedings of the IEEE International Conference Image Information Processing*. DOI: https://doi.org/10.1109/ICIIP53038.2021.9702660

Diaz-Escobar, J., & Kober, V. (2020). Natural Scene Text Detection and Segmentation Using Phase-Based Regions and Character Retrieval. *Mathematical Problems in Engineering*. DOI: https://doi.org/10.1155/2020/7067251

Francis, L. M., & Sreenath, N. (2022). Robust scene text recognition: Using manifold regularized Twin-Support Vector Machine. *Journal of King Saud University - Computer and Information Sciences*. DOI: https://doi.org/10.1016/j.jksuci.2019.01.013

Francisca O Nwokoma, Juliet N Odii, Ikechukwu I Ayogu, & James C Ogbonna. (2021). Camera-based OCR scene text detection issues: A review. *World Journal of Advanced Research and Reviews*. DOI: https://doi.org/10.30574/wjarr.2021.12.3.0705

Gao, X., Han, S., & Luo, C. (2019). A Detection and Verification Model Based on SSD and Encoder-Decoder Network for Scene Text Detection. *IEEE Access*. DOI: https://doi.org/10.1109/ACCESS.2019.2919994

Gupta, N., & Jalal, A. S. (2022). Traditional to transfer learning progression on scene text detection and recognition: a survey. *Artificial Intelligence Review*. DOI: https://doi.org/10.1007/s10462-021-10091-3

He, M., Liao, M., Yang, Z., Zhong, H., Tang, J., Cheng, W., Yao, C., Wang, Y., & Bai, X. (2021). MOST: A Multi-Oriented Scene Text Detector with Localization Refinement. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. DOI: https://doi.org/10.1109/CVPR46437.2021.00870

Jangid, M., & Srivastava, S. (2018). Handwritten Devanagari character recognition using layer-wise training of deep convolutional neural networks and adaptive gradient methods. *Journal of Imaging*. DOI: https://doi.org/10.3390/jimaging4020041

K. Bawa, R., & K. Sethi, G. (2014). A Binarization Technique for Extraction of Devanagari Text from Camera Based Images. *Signal & Image Processing : An International Journal*, *5*(2), 29–37. DOI: https://doi.org/10.5121/sipij.2014.5203

Kantipudi, M. P., Kumar, S., & Jha, A. K. (2021). Scene text recognition based on bidirectional lstm and deep neural network. *Computational Intelligence and Neuroscience*. DOI: https://doi.org/10.1155/2021/2676780

Kaur, A., Dhir, R., & Lehal, G. S. (2017). A survey on camera-captured scene text detection and extraction: towards Gurmukhi script. In *International Journal of Multimedia Information Retrieval*. DOI: https://doi.org/10.1007/s13735-016-0116-5

Khan, T., Sarkar, R., & Mollah, A. F. (2021). Deep learning approaches to scene text detection: a comprehensive review. *Artificial Intelligence Review*. DOI: https://doi.org/10.1007/s10462-020-09930-6

Lin, H., Yang, P., & Zhang, F. (2020). Review of Scene Text Detection and Recognition. *Archives of Computational Methods in Engineering*, *27*(2), 433–454. DOI: https://doi.org/10.1007/s11831-019-09315-1

Long, S., He, X., & Yao, C. (2021). Scene Text Detection and Recognition: The Deep Learning Era. *International Journal of Computer Vision*. DOI: https://doi.org/10.1007/s11263-020-01369-0

Mahajan, S., & Rani, R. (2021). Text detection and localization in scene images: a broad review. *Artificial Intelligence Review*, *54*(6), 4317–4377. DOI: https://doi.org/10.1007/s10462-021-10000-8

Naiemi, F., Ghods, V., & Khalesi, H. (2021a). A novel pipeline framework for multi oriented scene text image detection and recognition. *Expert Systems with Applications*. DOI: https://doi.org/10.1016/j.eswa.2020.114549

Naiemi, F., Ghods, V., & Khalesi, H. (2021b). MOSTL: An Accurate Multi-Oriented Scene Text Localization. *Circuits, Systems, and Signal Processing*. DOI: https://doi.org/10.1007/s00034-021-01674-0

Prasad, V., & Das, P. (2021). Recent Trends and Techniques in Text Detection and Text Localization in a Natural Scene: A Survey. *ADBU Journal of Engineering Technology*.

Rong, X., Yi, C., & Tian, Y. (2020). Unambiguous Scene Text Segmentation with Referring Expression Comprehension. *IEEE Transactions on Image Processing*. DOI: https://doi.org/10.1109/TIP.2019.2930176

Rong, X., Yi, C., & Tian, Y. (2022). Unambiguous Text Localization, Retrieval, and Recognition for Cluttered Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: https://doi.org/10.1109/TPAMI.2020.3018491

Seeri, S. V., Pujari, J. D., & Hiremath, P. S. (2016). Text Localization and Character Extraction in Natural Scene Images using Contourlet Transform and SVM Classifier. *International Journal of Image, Graphics and Signal Processing*. DOI: https://doi.org/10.5815/ijigsp.2016.05.02

Shiravale, S. S., Jayadevan, R., & Sannakki, S. S. (2020). Devanagari Text Detection From Natural Scene Images. *International Journal of Computer Vision and Image Processing*, *10*(3), 44–59. DOI: https://doi.org/10.4018/IJCVIP.2020070104

Shiravale, S. S., Sannakki, S. S., & Jayadevan, R. (2021). Text Region Identification in Indian Street Scene Images Using Stroke Width Transform and Support Vector Machine. *SN Computer Science*. DOI: https://doi.org/10.1007/s42979-021-00745-y

Soni, R., Kumar, B., & Chand, S. (2019). Text detection and localization in natural scene images based on text awareness score. *Applied Intelligence*. DOI: https://doi.org/10.1007/s10489-018-1338-4

Tian, C., Xia, Y., Zhang, X., & Gao, X. (2017). Natural scene text detection with MC–MR candidate extraction and coarse-to-fine filtering. *Neurocomputing*. DOI: https://doi.org/10.1016/j.neucom.2017.03.078

Uchida, S. (2014). Text Localization and Recognition in Images and Video. In *Handbook of Document Image Processing and Recognition* (pp. 843–883). Springer London. DOI: https://doi.org/10.1007/978-0-85729-859-1_28

Udupa, C., Upadhyaya, A., Patil, B. S., Seeri, S. V., Patil, P., & Hiremath, P. S. (2022). Text Localization and Script Identification in Natural Scene Images and Videos. *Proceedings of the 2022 International Conference on Connected Systems and Intelligence, CSI 2022*. DOI: https://doi.org/10.1109/CSI54720.2022.9924044

Vaidya, G., Vaidya, K., & Bhosale, K. (2020). Text recognition system for visually impaired using portable camera. *2020 International Conference on Convergence to Digital World - Quo Vadis, ICCDW 2020*. DOI: https://doi.org/10.1109/ICCDW45521.2020.9318706

Wei, Y., Shen, W., Zeng, D., Ye, L., & Zhang, Z. (2018). Multi-oriented text detection from natural scene images based on a CNN and pruning non-adjacent graph edges. *Signal Processing: Image Communication*, *64*, 89–98. DOI: https://doi.org/10.1016/j.image.2018.02.016

Zhao, R., Zheng, X., Ying, Z., & Fan, L. (2022). Localization of Pointed-At Word in Printed Documents via a Single Neural Network. *IEICE Transactions on Information and Systems*. DOI: https://doi.org/10.1587/transinf.2021EDP7143

Vijay Prasad
Assam Don Bosco University,
Guwahati, Assam, India
vpd.vijay82@gmail.com
ORCID 0000-0003-1996-4910

Pranab Das
Assam Don Bosco University,
Guwahati, Assam, India
pranab17@gmail.com
ORCID 0000-0003-0359-6615

Y. Jayanta Singh
NIELIT Guwahati, Assam,
India
yjayanta@gmail.com
ORCID 0000-0002-2886-1625