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

Revue d'Intelligence Artificielle
Vol. 36, No. 3, June, 2022, pp. 459-466
Journal homepage: http://iieta.org/journals/ria

DEFUSE: Deep Fused End-to-End Video Text Detection and Recognition

Chaitra Yuvaraj Lokkondra[1,2*], Dinesh Ramegowda[3], Gopalakrishna Madigondanahalli Thimmaiah[2], Ajay Prakash Bassappa Vijaya[4]

[1] Department of CSE, Jain University, Bengaluru 560069, Karnataka, India
[2] Department of CSE, SJB Institute of Technology, Affiliated to VTU, Bengaluru 560060, Karnataka, India
[3] Department of ISE, Jain University, Bengaluru 560069, Karnataka, India
[4] Department of CSE, Visvesvaraya Technological University, Regional Centre, Belagavi 590018, Karnataka, India

Corresponding Author Email: ylchaitra@gmail.com

https://doi.org/10.18280/ria.360314

Received: 6 March 2022
Accepted: 8 June 2022

Keywords:
text detection, text recognition, DeepEAST, Keras OCR, Boltzmann energy distribution technique, Benchmark datasets

ABSTRACT

Detecting and recognizing text in natural scene videos and images has brought more attention to computer vision researchers due to applications like robotic navigation and traffic sign detection. In addition, Optical Character Recognition (OCR) technology is applied to detect and recognize text on the license plate. It will be used in various commercial applications such as finding stolen cars, calculating parking fees, invoicing tolls, or controlling access to safety zones and aids in detecting fraud and secure data transactions in the banking industry. Much effort is required when scene text videos are in low contrast and motion blur with arbitrary orientations. Presently, text detection and recognition approaches are limited to static images like horizontal or approximately horizontal text. Detecting and recognizing text in videos with data dynamicity is more challenging because of the presence of multiple blurs caused by defocusing, motion, illumination changes, arbitrarily shaped, and occlusion. Thus, we proposed a combined DeepEAST (Deep Efficient and Accurate Scene Text Detector) and Keras OCR model to overcome these challenges in the proffered DEFUSE (Deep Fused) work. This two-combined technique detects the text regions and then deciphers the result into a machine-readable format. The proposed method has experimented with three different video datasets such as ICDAR 2015, Road Text 1K, and own video Datasets. Our results proved to be more effective with precision, recall, and F1-Score.

1. INTRODUCTION

In the development of the digital era, video-text detection and recognition have risen to prominence for real-time applications to assist blind people while traveling on roads, monitor vehicle license plates, and for surveillance which often needs text-recognition accuracies greater than 90% [1-3]. A robot navigation system is another fascinating text detection and recognition [4]. There are two different kinds of text in the videos: graphic and scene types [5]. Scene text in videos suffers from poor resolution, background complexity, color flashing, contrast variances, text size variations, different orientations, and motion blur compared to graphical text. As a result, effective performance for scene text in videos and images is difficult to achieve [6-10].

Visual aspects can be used to analyze text contained in images and videos. Visual aspects such as shape, texture, and color can be determined using static-based and motion-based methods [11]. Static-based methods learn the semantics of a scene by extracting frames from a video or local feature in frames/images. In the motion-based method, the main focus is the video text's motion (movement) aspects. The limitations of static-based methods are primarily concerned with detecting text in static images or frames. Motion blur, out-of-focus, and poor-quality images problems are not addressed in static images or frames. Methodologies developed for static-based methods do not always perform well in the video domain, and they do not take advantage of the additional details present in the video. With the rapid development of deep neural networks, Image/video scene text detection has advanced significantly in recent years [12-18]. Despite many methods in deep learning, text detection and recognition in the video remain a critical problem due to variations in fonts and motion blur. Hence, there is a need for an efficient deep learning model to detect and recognize text in videos. The video frames challenges of text in various orientations are depicted in Figure 1.

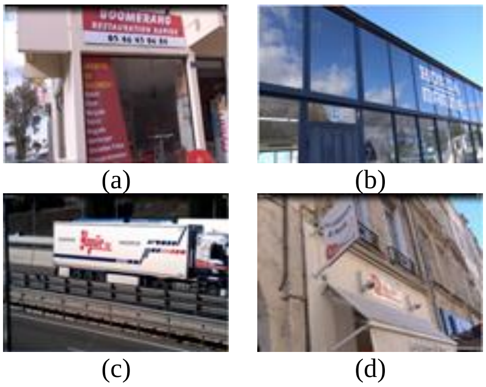

(a) (b)
(c) (d)

Figure 1. Video frames of text in various orientations. Video frame with (a) Vertical and Slant text (b) Non-horizontal text (c) Motion blur text (d) Multi-oriented/arbitrarily oriented text

The main contributions of this work are summarized herewith:

- We investigated state-of-the-art deep learning-based object-text detectors and restructured them to detect and recognize textual video content.
- The proposed framework gives good results for ICDAR 2015 videos, Road Text 1K text videos, and Own video dataset by running several iterations with various epochs and other hyper-parameters choices.
- The proposed DEFUSE model achieves excellent results when text with data dynamicity, blurry, defocusing, motion, illumination changes, arbitrarily shaped, and occlusion text with less training time.

The rest of this paper is laid out as follows. The second section gives an overview of related work. Section three describes the proposed method in detail; followed by Section four describes datasets and video findings. Section five summarizes the findings and gives out a plan for further work.

2. RELATED WORK

Scene text detection and recognition in videos/images is an important computer vision research area. The approaches for text detection and recognition in videos/images offered by researchers are presented in this section.

2.1 Text detection

Most traditional methods concentrate mainly on identifying text in each frame/image with connected components and sliding window approaches [19-23]. However, this approach has limited output due to handcrafted features' low representation. The proposed work uses connected-component-based text detection methods with exploring Maximally Stable Extremal Regions (MSER). The Stroke Width Transform (SWT) method was applied to determine minimal intra stroke variation in the connected components. Huang [24] uses the wavelet transform to get an edge map of the synthesized frame. The synthesized frame is then implemented using the multi-frame integration (MFI), and the candidate text regions are removed using the multi-frame verification (MFV). Qin et al. [25] propose an efficient scheme for categorizing distinct video text frame/image and symmetry features by identifying candidate- edge-components for each input frame using the Canny and Sobel edge method. The existing traditional methods are inefficient compared to deep neural networks, especially when handling complex scene images/videos.

Natural scene text detection mechanisms have seen a lot of developments, which has led to the exponential growth of deep learning. Many text detection systems based on convolutional neural networks (CNNs) have been proposed and succeeded. Sun et al. [26] introduce an efficient text detection approach utilizing color-enhanced contrasting extremal regions (CER). This approach requires working on vertical or curved text lines. Bai et al. [27] proposed a novel CNN variant for the problem of text/non-text image classification called Multi-scale Spatial Partition Network (MSP-Net). Wang et al. [28] recently used an optical flow-based approach to optimize text positions in corresponding frames. Xue et al. [29] investigate the degree of blur estimation dependent on neighboring pixel variance. This method widens the distance between text and non-text pixels by calculating a gradient for each pixel and combining it with a degree of blur. This method uses k-means clustering for the above features to extract the text candidates. Zhang et al. [30] propose a scheme that combines both object text-detection (used YOLOv3) and text recognition (employed with CRNN-Convolutional Recurrent Neural Network). Based on the above findings, it is clear that detecting text in videos is extremely challenging.

2.2 Text recognition

Several works are proposed to solve text recognition in video/images, but still, unstructured texts are not properly recognized. Mishra et al. [31] propose an efficient method to recognize the text in natural images, which provides a recognition rate of 73.26% for Street View Data (SVT) and 81.78% for ICDAR 2003 data. Phan et al. [32] propose an approach for multiple-frame (temporal) integration of video text by using SWT (Stroke Width Transform) and SIFT (scale-invariant feature transform) to improve the recognition rate. Rong et al. [33] suggested multi-frame scene text recognition on scene text character (STC) for letter estimations and conditional random field (CRF) paradigm for word configuration in several frames based on textual data tracking. Sudir and Ravishankar [34] developed a new strategy for Video-text binarization employing Wavelet-Edge-Map and Block-Eigen-values for both Superimposed and Multi-Oriented Scene text.

Importantly, some methods examine convolutional neural networks (CNN) to increase text recognition rates in video and natural scene images [35-38]. Shi et al. [35] proposed an image-based sequence recognition for end-to-end trainable neural networks and their usage in scene text recognition. Shi et al. [39] propose a new method for detecting and recognizing video text that uses a corner response feature map and transferred deep-CNN. The findings work well on newly created videos with various languages and fonts. Cheng et al. [40] proposed a coherent system named SVST (spotting video scene text) for effectively detecting scene text in videos. Then, to reliably localize text regions in scene images, a spatial-temporal video text detector (SVTD) was trained. Build a text stream scoring network (TSSN) instead of recognizing each text region in a text stream to evaluate the content of each text region and find the text region with the best quality score. Harizi et al. [41] proposed a modern CNN-based scene text reading system. They investigated how to merge the character recognition module and the word recognition module to perform better or get highly competitive results. Among these numerous approaches, there are still difficulties in scene text images/video with varying orientations and realistic distortions to enhance recognition efficiency further. As a result, this paper proposes the DEFUSE model for text detection and recognition in video, and the method works well irrespective of different orientations.

3. PROPOSED METHODOLOGY

The proposed methodology consists of three phases: preprocessing, text detection, and text recognition. The workflow of the proposed methodology is shown in Figure 2.

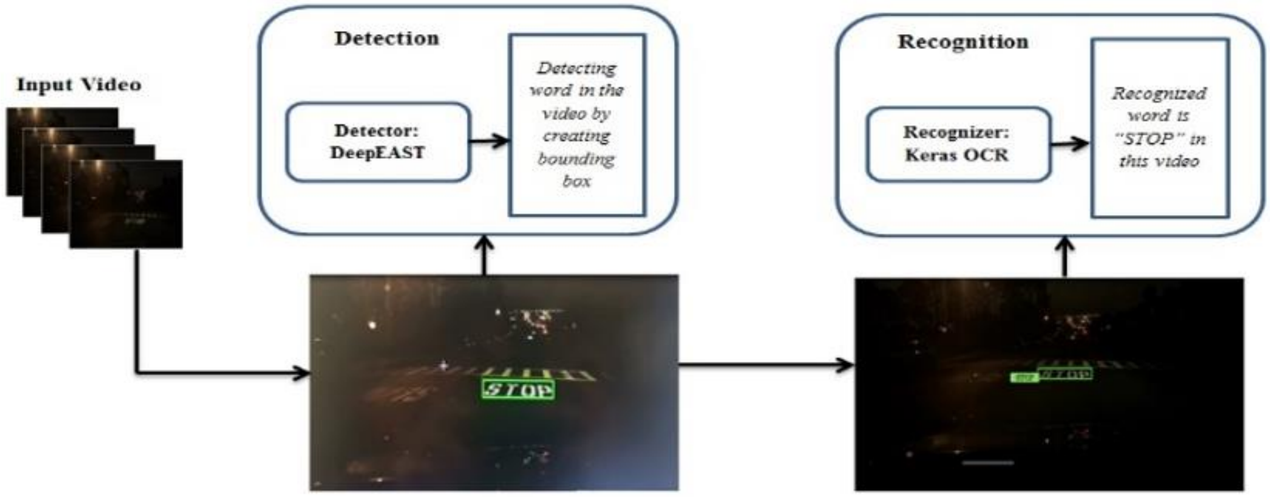

Input Video
Detection
Detector: DeepEAST
Detecting word in the video by creating bounding box
Recognition
Recognizer: Keras OCR
Recognized word is "STOP" in this video
STOP

Figure 2. Block diagram of the proposed methodology

In the first phase, input video gets converted into frames then the frames are passed to de-convolution networks to perform data preprocessing [42]. The data preprocessing is carried out to remove blurriness, noise, and haziness from the video frames. De-convolutional networks look for features that have been lost and are used to access the information in frame images. After de-convolution, it is passed to the RGB contrast map using the K-Means distribution technique, providing a cluster of high and low contrast images [43]. Clustering is carried out using features with similar regions. In the second phase, preprocessed data frames are passed to the text detection technique to localize text with a bounding box. The proposed Deep Efficient and Accurate Scene Text Detector [16] (DeepEAST) model detects the text from video frames. After text detection, the coordinates are extracted through the Boltzmann Energy distribution technique [44]. Finally, the cropped text detection part is passed to the Keras-OCR (Optical Character Recognition) model to recognize text in videos. Keras-OCR will return a (word, box) tuple when passing an image through it, with the box containing the coordinates of the word with four box corners. Here, we took three (ResNet 50 [45], Mobile Net SSD [46, 47], and YOLOv5 [48]) models for cross-analysis, in which Keras OCR [49] turned out to be the best SOTA accuracy and finally, the output generated. A DEFUSE algorithm is proposed to localize and recognize the text in the video. The detailed steps of the proposed DEFUSE algorithm are explained below:

DEFUSE Algorithm: Text Detection and Recognition in Video

While (frames! =0)

{START:

Step 1: Input video is converted into the frames. Let the video be denoted as 'V,' and individual frames be f_n [where 'n' denotes a frame number, n = (1, 2, 3………n)].

$$V = \sum_{i=1}^{n} f_i$$

Step 2: Now, the individual frame contains two regions. Let the region with the text be 'T,' and the non-text be 'N.'

$$\therefore f = (N + T)$$

So, first, text region localization is performed by using the segmentation technique. The text segments are identified using the Boltzmann Distribution Technique and can find out the maximum energy density of pixels. Let in frame f1; we have pixel map P. Now, since a frame has two regions (N, T), pixels will be different; dense (P') and non-dense pixels (P"). Here, a dense pixel gives more detailed information about the text in the image, and non-dense pixels contain less information about the image.

$$\therefore P = (P' + P'')$$

We are now applying the Boltzmann formula to extract dense pixel zone to have the text zone idea.

$$\int_{1}^{n} Pdf = \int (P')df + \int (P'')df$$

Here, it is returning the areas of dense pixel zone.

Step 3: Now, we got the area where the text is present. In order to create a boundary box, let us consider area coordinators of P$'$ be A_x and A_y. So, to plot the bounding box, we need to plot the vector. Here, the vector determines the positions of points. Let the angle made by vectors be θ.

$$\therefore \theta = \cos^{-1}\left(\frac{A_x.A_y}{|A_x||A_y|}\right)$$

$$\frac{d\theta}{dP'} = \frac{d}{dP'}\cos^{-1}\left(\frac{A_x.A_y}{|A_x||A_y|}\right)$$
$$= Total\ area\ of\ bounding\ box.$$

Step 4: After the localization, the localized text is de-blurred using the de-convolution technique. Since the dense pixel is P' and all pixels are a collection of sub-pixel. Let sub-pixel of P$'$ be P_i where (i =1, 2,……n).

$$\therefore P' = \prod_{i=1}^{i=n} P_i$$

Now each P_i is de-convoluted by differentiations.

Step 5: Finally, the Keras OCR technique recognizes text, which runs on the frame-wise localized text. Then understand the pixel distribution, cross tallies the training database, and make an 'n' prediction sample to select the best pair by weight-based sort.

: STOP}

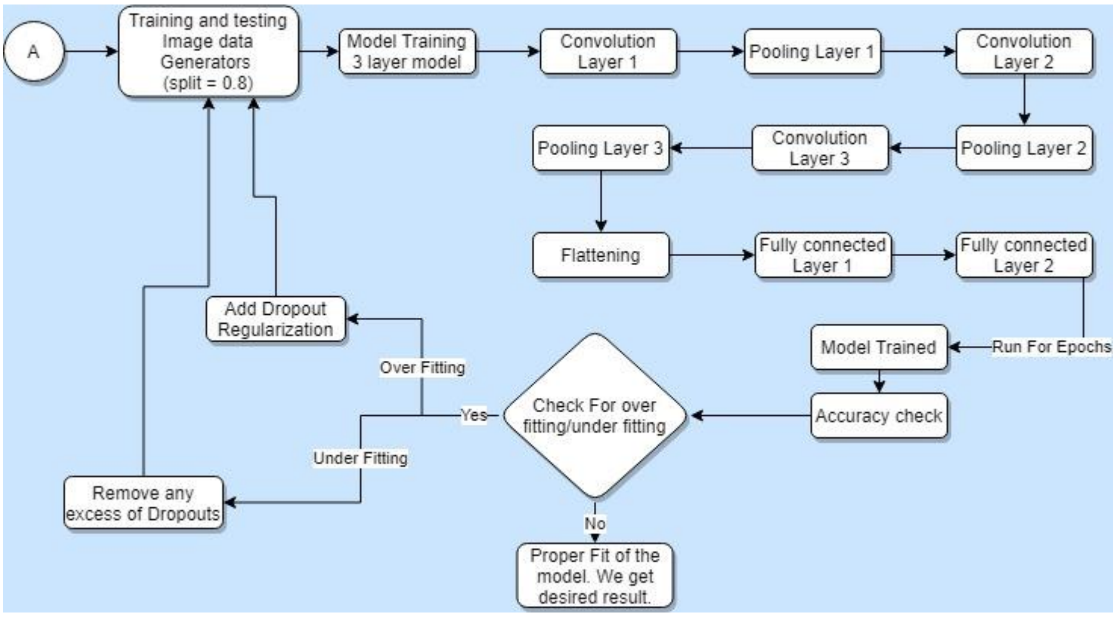


Figure 3. Deep fused model: DeepEast + Keras OCR

3.1 Develop DeepEAST and Keras OCR model

The DeepEAST (Deep Efficient and Accurate Scene Text Detector) and Keras OCR (Keras Optical Character Recognition) model are fused to localize and recognize the text in video frames and is influenced by a neural network. Common layers in neural networks are convolutional layers, pooling layers, and dropout. The seven-layer convolutional is used to extract the features, and Pooling layers help classify the valuable and improper features. The features which are not required it is immediately passed to dropout. The dropout normalizes the rest of the feature map, and improper features will be removed. The kernel size of each layer is three, the stride used is two, and as the activation function, ReLU is used. The DeFuse DeepEAST and Keras OCR model architecture is shown in Figure 3.

'A' is the video input that will be converted into frames. Then three-layer model training is carried out by simply extracting the feature map (individual coordinate points of text), passing the convolutional layer, and next passing into pooling. Totally ten steps training done, initially three and next seven-layer training steps for DeepEAST. After the seventh layer of training, we have a refined feature map and pass to flattening, where individual feature analysis is carried out. Now individual attention will be done, and feature enhancement will occur; the DeepEAST model fully learned a good feature and then passed to a fully connected layer to get a complete refining process. Finally, the model gets trained. Thus, the text detection and recognition process were completed using DeepEAST and Keras OCR.

4. IMPLEMENTATION AND RESULTS

Initially, the implementation is carried out with Python 3.6 in a Google Colab and then moved to Google Colab Pro. The standard video datasets used for experimenting are ICDAR-2015 [50], RoadText-1K [51], and own video datasets. Table 1 shows the video dataset summary. The models are trained at different levels of epochs with Adam optimizers [52]; the loss function used cross-entropy to minimize loss/error and the other parameters- with batch_size of 32 and the learning rate set to 0.1. We employ NMS [16] to remove multiple bounding boxes, and Precision, Recall, and F-Score are used to evaluate performance.

Table 1. Comparison of different text video datasets

Datasets	ICDAR-2015	RoadText-1K	Own Dataset
Source	Egocentric	Car-mounted	College-Campus view
Length (Seconds)	Varying	Varying	Varying
Videos	49	1000	12
Resolution	720 * 480	1280 * 720	varying
Annotated Frames	27,824	300,000	70
Total Text Instances	143,588	1,280,613	≈100
Text type	Scene Text	Scene Text	Scene Text
Unique Words	3,563	8,263	≈30-40

ICDAR 2015 Video Dataset contains 49 videos ranging in length from ten seconds to one minute; the shot was taken outdoors and indoors. The dataset consists of 25 videos used for training, 24 for testing, and sizes ranging from 720 x 480 to 1280 x 960. Challenges facing in ICDAR 2015 video dataset are motion blur and out-of-focus issues. The dataset consists of different languages (Spanish, French, English), but the proposed method is focused on English video text.

RoadText-1k Dataset consists of 1000 videos in outdoor scenes, which is 20 times the size of the existing video text. The image size is 1280 x 720 pixels, including 700 videos for training and 300 videos for testing. Challenges faced in the RoadText-1k dataset are low contrast, motion blur, out-of-focus, and distortions. Text instances are three categories:

illegible text, English, and Non-English text.

Own Video Dataset consists of 12 videos in outdoor scenes captured using iPhone XR. The frame size varies from 1280 * 720 to 1920 * 1080, consisting of the English language with different font sizes. The duration of the video is 25 to 30 seconds. The challenges faced in the own video dataset are motion blur, fast text movement, and out-of-focus issues.

4.1 Result analysis

A comparative study is performed by comparing existing methods such as FREE [53] and YORO [54]. Precision, Recall, and the F1-score are used to assess the proposed model's accuracy. Precision is a text detected correctly (Tr.p) divided by the text detected correctly (Tr.p), and non-text detected as text (Fp). A recall is text detected correctly (Tr.p) divided by the text detected correctly (Tr.p) and not having text but detected with text (Fn). F1-score is used to measure test accuracy. The Precision, Recall, and F1-score can be computed using the following equations:

$$\text{Precision} = \frac{Tr.p}{Tr.p + Fp} \tag{1}$$

$$\text{Recall} = \frac{Tr.p}{Tr.p + Fn} \tag{2}$$

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

where, Tr.p: True-Positive, Fp: False-Positive, Fn: False-Negative.

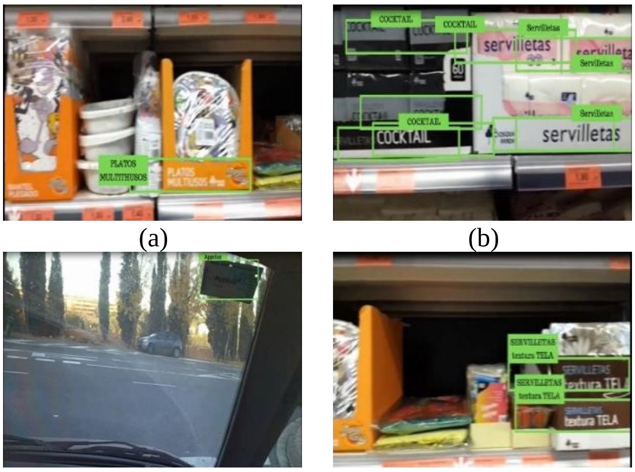
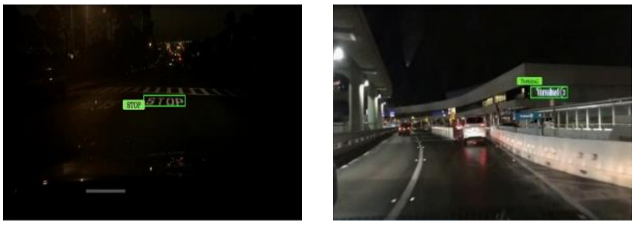

(a) (b)
(c) (d)

Figure 4. Output text was detected and recognized from videos on ICDAR 2015 dataset-(a), (b), (c) and (d). The green region is a true positive result, and the proposed method performs admirably for dense, horizontal, moving, and slanted text


(a) (b)

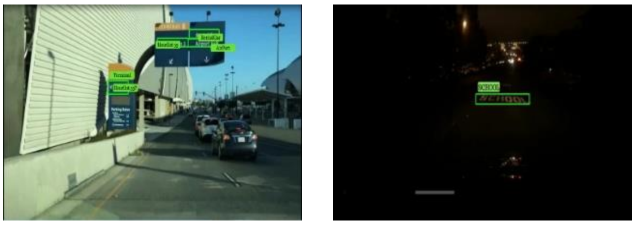

(c) (d)

Figure 5. Output text was detected and recognized from videos on RoadText-1k dataset-(a), (b), (c) and (d). The green region is a true positive result, and the proposed method performs admirably for horizontal, dense, moving text and dark time

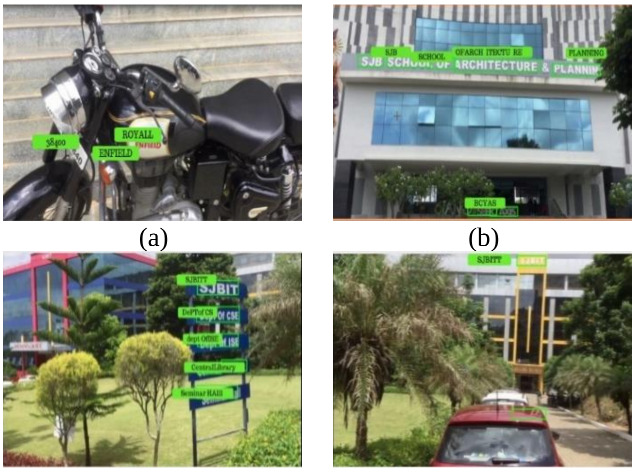

(a) (b)
(c) (d)

Figure 6. Output text was detected and recognized from videos on Own dataset-(a), (b), (c) and (d). The green region is a true positive result, and the proposed method performs admirably for horizontal, dense, moving, and far away text. We noted some bounding box, which does not have text

Table 2. Comparison of the proposed approach with the existing methods for different videos dataset

Dataset	Methods	Precision	Recall	F1-score
ICDAR-2015	YORO [54]	68.28	67.21	67.74
	FREE [53]	71.52	65.48	68.36
	Proposed Method	76	67	71.21
RoadText-1K	FREE [53]	34.82	20.74	26
	Proposed Method	41.22	32.84	36.55
Own Dataset	Proposed DEFUSE Method	37.46	24.11	29.34

The proposed method reaches state-of-the-art performance for the video datasets, as shown in Table 2. We can see the results by comparing them with different methods, and the results are minor because of the complexity of challenging video frames. The processing time consumption of the proposed model is 2.30 FPS-Frames Per Second with reasonable accuracy. Figures 4 and 5 show the video output frames for the ICDAR 2015 and RoadText-1k video datasets by the proposed DEFUSE method. Some of the text in the RoadText-1k is illegible/ occluded and contains the non-English text; hence results are low compared to the ICDAR-2015 video dataset. Figure 6 shows the video text frames for the own video datasets. Here, the green regions represent the localized text's bounding box, and the recognized text is annotated on top of the bounding box.

5. CONCLUSIONS

Text Detection and recognition from the video is a tough challenge due to the high complexity in the appearance of the text and data dynamicity problem, in which the screen moves from one place to another in videos. No prior information will be available about the text's location, direction, and different text angles in videos, which is the most challenging task. In this concern, a combination model is proposed named DeepEAST and Keras OCR, which came out with extensive results for ICDAR 2015 video dataset. The suggested model outperforms the existing techniques in terms of accuracy. RoadText-1K dataset needs to be improved in further extension of the work. As a result, text detection and video recognition are still difficult, and additional research is needed. The model will be extended in the future to integrate with sophisticated pretrained-models to detect and recognize text in various conditions like complex background, curved like text, and blurry text, and experiments will be carried out on various video datasets.

ACKNOWLEDGMENTS

We thank the research community and our well-wishers.

REFERENCES

[1] Zhang, J., Kasturi, R. (2014). A novel text detection system based on character and link energies. IEEE Transactions on Image Processing, 23(9): 4187-4198. https://doi.org/10.1109/TIP.2014.2341935

[2] Sain, A., Bhunia, A.K., Roy, P.P., Pal, U. (2018). Multi-oriented text detection and verification in video frames and scene images. Neurocomputing, 275: 1531-1549. https://doi.org/10.1016/j.neucom.2017.09.089

[3] Ye, Q., Doermann, D. (2014). Text detection and recognition in imagery: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(7): 1480-1500. https://doi.org/10.1109/TPAMI.2014.2366765

[4] Zarechensky, M. (2014). Text detection in natural scenes with multilingual text. Proceedings of the Tenth Spring Researcher's Colloquium on Database and Information Systems, Veliky Novgorod, Russia.

[5] Lokkondra, C.Y., Ramegowda, D., Gopalkrishna, M.T., Vijaya, A.P.B., Shivananjappa, M.H. (2021). ETDR: An exploratory view of text detection and recognition in images and videos. Rev. d'Intelligence Artif., 35(5): 383-393. https://doi.org/10.18280/ria.350504

[6] Sharma, N., Pal, U., Blumenstein, M. (2012). Recent advances in video based document processing: A review. In 2012 10th IAPR International Workshop on Document Analysis Systems, pp. 63-68. https://doi.org/10.1109/DAS.2012.72

[7] Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X. (2018). Arbitrary-oriented scene text detection via rotation proposals. IEEE Transactions on Multimedia, 20(11): 3111-3122. https://doi.org/10.1109/TMM.2018.2818020

[8] Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., Jia, J. (2019). Learning shape-aware embedding for scene text detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 4234-4243. https://doi.org/10.1109/CVPR.2019.00436

[9] Jiang, Y., Zhu, X., Wang, X., Yang, S., Li, W., Wang, H., Luo, Z. (2017). R2CNN: Rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv:1706.09579.

[10] Deng, D., Liu, H., Li, X., Cai, D. (2018). Pixellink: Detecting scene text via instance segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12269

[11] Pour, A.K., Seng, W.C., Palaiahnakote, S., Tahaei, H., Anuar, N.B. (2021). A survey on video content rating: taxonomy, challenges and open issues. Multimedia Tools and Applications, 80(16): 24121-24145. https://doi.org/10.1007/s11042-021-10838-8

[12] Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 20-36. https://doi.org/10.1007/978-3-030-01216-8_2

[13] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X. (2016). Multi-oriented text detection with fully convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4159-4167. https://doi.org/10.1109/CVPR.2016.451

[14] Wu, W., Xing, J., Yang, C., Wang, Y., Zhou, H. (2020). A scene text detector for text with arbitrary shapes. Mathematical Problems in Engineering, 2020: 8916028. https://doi.org/10.1155/2020/8916028

[15] Zhang, C., Liang, B., Huang, Z., En, M., Han, J., Ding, E., Ding, X. (2019). Look more than once: An accurate detector for text of arbitrary shapes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10552-10561.

[16] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J. (2017). East: An efficient and accurate scene text detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5551-5560. https://doi.org/10.1109/CVPR.2017.283

[17] Sharma, N., Shivakumara, P., Pal, U., Blumenstein, M., Tan, C.L. (2012). A new method for word segmentation from arbitrarily-oriented video text lines. In 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 1-8. https://doi.org/10.1109/DICTA.2012.6411703

[18] Chaitra, Y.L., Dinesh, R., Gopalakrishna, M.T., Prakash, B.V. (2021). Deep-CNNTL: Text localization from natural scene images using deep convolution neural network with transfer learning. Arabian Journal for Science and Engineering, pp. 1-12. https://doi.org/10.1007/s13369-021-06309-9

[19] Yin, X.C., Yin, X., Huang, K., Hao, H.W. (2013). Robust text detection in natural scene images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(5):

970-983. https://doi.org/10.1109/TPAMI.2013.182

[20] Kim, K.I., Jung, K., Kim, J.H. (2003). Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12): 1631-1639. https://doi.org/10.1109/TPAMI.2003.1251157

[21] Epshtein, B., Ofek, E., Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2963-2970. https://doi.org/10.1109/CVPR.2010.5540041

[22] Neumann, L., Matas, J. (2013). Scene text localization and recognition with oriented stroke detection. In Proceedings of the IEEE International Conference on Computer Vision, pp. 97-104. https://doi.org/10.1109/ICCV.2013.19

[23] Pan, Y.F., Hou, X., Liu, C.L. (2010). A hybrid approach to detect and localize texts in natural scene images. IEEE Transactions on Image Processing, 20(3): 800-813. https://doi.org/10.1109/TIP.2010.2070803

[24] Huang, X. (2012). Automatic video text detection and localization based on coarseness texture. In 2012 Fifth International Conference on Intelligent Computation Technology and Automation, pp. 398-401. https://doi.org/10.1109/ICICTA.2012.106

[25] Qin, L., Shivakumara, P., Lu, T., Pal, U., Tan, C.L. (2016). Video scene text frames categorization for text detection and recognition. In 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 3886-3891. https://doi.org/10.1109/icpr.2016.7900241

[26] Sun, L., Huo, Q., Jia, W., Chen, K. (2015). A robust approach for text detection from natural scene images. Pattern Recognition, 48(9): 2906-2920. https://doi.org/10.1016/j.patcog.2015.04.002

[27] Bai, X., Shi, B., Zhang, C., Cai, X., Qi, L. (2017). Text/non-text image classification in the wild with convolutional neural networks. Pattern Recognition, 66: 437-446. https://doi.org/10.1016/j.patcog.2016.12.005

[28] Wang, L., Wang, Y., Shan, S., Su, F. (2018). Scene text detection and tracking in video with background cues. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 160-168. https://doi.org/10.1145/3206025.3206051

[29] Xue, M., Shivakumara, P., Zhang, C., Lu, T., Pal, U. (2019). Curved text detection in blurred/non-blurred video/scene images. Multimedia Tools and Applications, 78(18): 25629-25653. https://doi.org/10.1007/s11042-019-7721-2

[30] Zhang, F., Luan, J., Xu, Z., Chen, W. (2020). DetReco: Object-text detection and recognition based on deep neural network. Mathematical Problems in Engineering, 2020: 2365076. https://doi.org/10.1155/2020/2365076

[31] Mishra, A., Alahari, K., Jawahar, C.V. (2012). Top-down and bottom-up cues for scene text recognition. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2687-2694. https://doi.org/10.1109/CVPR.2012.6247990

[32] Phan, T.Q., Shivakumara, P., Lu, T., Tan, C.L. (2013). Recognition of video text through temporal integration. In 2013 12th International Conference on Document Analysis and Recognition, pp. 589-593. https://doi.org/10.1109/icdar.2013.122

[33] Rong, X., Yi, C., Yang, X., Tian, Y. (2014). Scene text recognition in multiple frames based on text tracking. In 2014 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. https://doi.org/10.1109/ICME.2014.6890248

[34] Sudir, P., Ravishankar, M. (2014). An effective approach towards video text recognition. Advances in Signal Processing and Intelligent Recognition Systems, pp. 323-333. https://doi.org/10.1007/978-3-319-04960-1_29

[35] Shi, B., Bai, X., Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371

[36] Lee, S.J., Kim, S.W. (2016). Recognition of slab identification numbers using a deep convolutional neural network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 718-721. https://doi.org/10.1109/ICMLA.2016.0128

[37] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. International Journal of Computer Vision, 116(1): 1-20. https://doi.org/10.1007/s11263-015-0823-z

[38] B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai. (2016). Robust scene text recognition with automatic rectification. In Proc. CVPR, pp. 4168-4176. https://doi.org/10.1109/CVPR.2016.452

[39] Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X. (2016). Robust scene text recognition with automatic rectification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168-4176. https://doi.org/10.1109/access.2018.2851942

[40] Cheng, Z., Lu, J., Xie, J., Niu, Y., Pu, S., Wu, F. (2019). Efficient video scene text spotting: Unifying detection, tracking, and recognition. ArXiv, abs/1903.03299

[41] Harizi, R., Walha, R., Drira, F., Zaied, M. (2022). Convolutional neural network with joint stepwise character/word modeling based system for scene text recognition. Multimedia Tools and Applications, 81(3): 3091-3106. https://doi.org/10.1007/s11042-021-10663-z

[42] Reshma Vijay, V.J., Deepa, P.L. (2016). Image deblurring using convolutional neural network. IOSR-JECE, 11(5): 7-12. https://doi.org/10.9790/2834-1105020712

[43] Dhanachandra, N., Manglem, K., Chanu, Y.J. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54: 764-771. https://doi.org/10.1016/j.procs.2015.06.090

[44] Alani, A.A. (2017). Arabic handwritten digit recognition based on restricted Boltzmann machine and convolutional neural networks. Information, 8(4): 142. https://doi.org/10.3390/info8040142

[45] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[46] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[47] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.,

Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. In European Conference on Computer Vision, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[48] Solawetz, J., Nelson, J. (2021). How to train yolov5 on a custom dataset. Roboflow Blog. https://blog.roboflow.com/how-to-train-yolov5-on-a-custom-dataset/, accessed on November 17, 2021.

[49] Troller, M. (2017). Practical OCR system based on state of art neural networks.

[50] Zhou, X., Zhou, S., Yao, C., Cao, Z., Yin, Q. (2015). ICDAR 2015 text reading in the wild competition. arXiv preprint arXiv:1506.03184.

[51] Reddy, S., Mathew, M., Gomez, L., Rusinol, M., Karatzas, D., Jawahar, C.V. (2020). Roadtext-1k: Text detection & recognition dataset for driving videos. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 11074-11080. https://doi.org/10.1109/ICRA40945.2020.9196577

[52] Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[53] Cheng, Z., Lu, J., Zou, B., Qiao, L., Xu, Y., Pu, S., Zhou, S. (2020). Free: A fast and robust end-to-end video text spotter. IEEE Transactions on Image Processing, 30: 822-837. https://doi.org/10.1109/TIP.2020.3038520

[54] Cheng, Z., Lu, J., Niu, Y., Pu, S., Wu, F., Zhou, S. (2019). You only recognize once: Towards fast video text spotting. In Proceedings of the 27th ACM International Conference on Multimedia, pp. 855-863. https://doi.org/10.1145/3343031.3351093

NOMENCLATURE

Abbreviations

EAST	Efficient and Accurate Scene Text Detector
OCR	Optical Character Recognition
ROI	Region of Interest
CNN	Convolutional Neural Network
ResNet	Residual Neural Network
SSD	Single-Shot Detector
YOLOv5	You Only Look Once version 5
SOTA	State-of-the-Art
DEFUSE	Deep Fused
RGB	Red Green Blue
ICDAR	International Conference on Document Analysis and Recognition
NMS	Non Maximum Suppression
FREE	Fast and Robust End-to-End
YORO	You Only Recognize Once