Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 8
August 2023

An Efficient Scene Text Spotter and Detector Using Deepnet

Dhirendra Kumar Yadav[1], Sonalal Yadav[2], Pintu Chauhan[3]
[1]Parul University
[1,3]Parul University, Vadodara, India.
[2]Mid-West Uni versity, Surkhet, Nepal.

Abstract
Scene text detection is a complex and challenging task due to various environmental factors, such as varying illuminations, lighting conditions, tiny and curved texts, and more. Many existing works in scene text detection focus on improving model accuracy but often overlook the need for efficiency, resulting in heavy-weight models that require significant processing resources. This paper introduces a novel efficient model to address the objectives of improving accuracy and efficiency in scene text detection.This paper proposes a new hybrid model for text detection in images that uses ResNet50 with AtrousSpatial Pyramid Pooling (ASPP) based on Efficient and Accurate Scene Text (EAST) algorithm with A technique for suppressing duplicate text detections that is more lenient than traditional non-maximum suppression.The proposed method is designed to improve the efficiency and accuracy of text detection, Experiments on the IIIT-ILST, ICDAR2015 and ICDAR2019, MSRA-TD500 dataset, show that the proposed method achieves state-of-the-art results

Index Terms: ASPP,Text detection, EAST, Soft Non-Max Suppression, Text Recognition, Text Spotting

1. INTRODUCTION

The difficult procedure of scene text identification is looking for words inside picturesque scenes. Due to several environmental restrictions, such as poor image quality, blurriness, uneven lighting, inconsistent appearance, and varied backgrounds in the images, this task is difficult. The presence of small, curved letters makes it even harder to recognize and extract meaningful text signals from the sceneries.Despite these challenges, recent advancements in deep learning and the combination of neural network[1]models have brought significant progress to the field of scene text detection. Techniques based on deep learning have proven to be highly effective in classifying text, leading to notable advancements in this area. Specifically, the focus has been on addressing classification problems related with text that is asymmetrical, including vertical or curved text, which has several real-time applications.For example, scene text detection plays a vital role in applications like automobile navigation, where street-facing cameras equipped with this technology assist in navigation tasks.

In scene text reading, the standard approach involves breaking down the process into two main section: Text detection and text recognition. These two tasks are works as separate and independent steps. Among these, text detection[2] plays a critical role as it serves as the foundation for the subsequent stages of text information extraction and understanding. Accurate text detection is essential to identify and locate text regions inside the scene images, enabling the subsequent text recognition step to process and interpret the text content correctly. By ensuring reliable text detection, the overall process of scene text reading becomes more effective and reliable, leading to successful text information extraction and understanding

Many efficient techniques have been developed as a result of the widespread usage of semantic segmentation in scene text detection. These method can be divided generally into two types. The EAST[3,4] technique illustrates the first category by directly regressing text bounding boxes from the pixel-level data and predicting a score map. The Pixel Link[5] technique is an example of the second category, which, after predicting the score map, combines pixel instances using post-processing to produce the text bounding boxes.. The encoder-decoder structure is used by both categories to extract dense features.

By repeatedly striding or pooling feature maps, the network initially reduces their spatial resolution in the encoding module.This is carried out to widen the receptive field. and enable the model to gather more contextual data. The detail of spatial resolution is then recovered via deconvolution or up-sampling to provide dense features in the decoder module.However, this method may result in the loss of vital background data, which is necessary for precise text detection. The model's capacity to correctly detect and comprehend text instances in complicated scene may be hampered by the loss of context.

To address this issue, further research and advancements are being made to enhance the encoder-decoder structure and preserve important context information during the feature extraction process. The goal is to develop more robust and accurate scene text detection methods that can effectively handle challenging environments with diverse text orientations, sizes, and backgrounds.We addressed the difficulties in scene text detection in this study. by making enhancements to the existing EAST module. We provide an entirely original multi-oriented scene text detector that makes use of arous convolution.[6], a method that broadens the model's receptive area while retaining more context input.

Our proposed approach offers several key contributions:

- we integrated new atrous convolution to the backbone network, the model was able to extract additional context from the input image. By strategically adjusting the dilation rates, the model effectively handles various text orientations and backgrounds.
- On top of feature maps, we introduced the Atrous Spatial Pyramid Pooling (ASPP)[7] model to further increase text detection accuracy. The network's ability to recognize texts of various sizes and orientations is improved by ASPP, which enables the network to detect text information at various scales.
- We thoroughly assessed our proposed approach on benchmark datasets commonly used in scene text detection research. The results demonstrated highly competitive performance, on the MSRA-TD500 benchmark dataset, significantly outperforming the prior state-of-the-art technique.

Our study modifies the EAST module and introduces a completely novel atrous convolution network based multi-oriented text detector uses handle the difficulties of scene text detection. The integration of atrous convolution and ASPP enriches the model's context-awareness, leading to improved text detection accuracy across various challenging scenarios. The promising benchmark results highlight the value and potential impact of our proposed method in advancing the field of scene text detection.

2. LITERATURE SURVEY

Over the past few decades, researchers have explored various text detection methods, including deep learning techniques. In this section, several papers relevant to this field are discussed.[8]Using a single neural network to recognize arbitrary forms and quadrilateral shapes in an image,We proposed the EAST pipeline, which successfully handles challenging scene text scenarios. On the dataset ICDAR 2015, the COCO-Text, and the MSRA-TD500, their technique received remarkable F1 scores of 0.7820[9]. presented aFast Oriented Text Spotting (FOTS) network is a trainable complete technique for text detection and recognition. On the ICDAR 2013 dataset, ICDAR 2015 dataset, and ICDAR 2017 MLT dataset, their solution beat stateofthe-art approaches, with more than a 5% enhancement in text detection outcomes on ICDAR 2015.[10] introduced the method known as TextSnake, which was created to address the practical issues such that text freeform curved presented to earlier techniques.

TextSnake produced comparable results on datasets like Total-Text, ICDAR2015,SCUT-CTW1500, and MSRATD500 while successfully representing texts in curved, vertical and horizontal formats. On the Total-Text dataset, it performed noticeably better than the baseline by over 40% in terms of F-measure. presented TextBoxes[11], a trainable, end-to-end scene text detector renowned for its dependability in recognizing scene text of any orientation with a

single network in forward pass. Their approach achieved impressive F1 scores of 0.817 on the dataset ICDAR 2015 and 0.5591 on the COCOText images dataset.

In recent times, methods based on deep learning have taken over as the focal point for scene text identification, attracting significant interest from researchers. These approaches harness the power of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) which are types of deep neural networks, to directly learn intricate patterns and features from input images. By training on a sizable dataset of South Indian language images, these models exhibit remarkable proficiency in detecting and localizing text regions with precision.

One particular study [12] concentrates on detecting Telugu text in scene images, specifically in horizontal orientation. To accomplish this goal, the researchers make crucial modifications to the SSD (Single Shot MultiBox Detector) approach, customizing it for horizontal text detection. Additionally, they incorporate elements from the prediction procedure employed in Textboxes to handle multiple vertical offsets. In lieu of using the conventional VGG-16 network for feature extraction, they opt for the more advanced Densenet, which significantly enhances the model's ability to capture salient features.

To address the challenge of prioritizing foreground text over easily recognizable backgrounds, the researchers introduce SSD Focal Loss, a variant of the standard loss function. To enhance the model's dense predictions, they also make use of forecasts at various vertical offsets and better resolution inputs.Their method's use of K-means clustering to establish the default bounding box aspect ratios is distinctive and differentiating it from traditional methods like Textboxes.

Our proposed method's encoder-decoder module is comparable to those found in EAST, Pixel-Link, and PSENet[13], which employ Feature Pyramid Networks (FPN)[14].efficient feature extraction and representation of the input images.To address this limitation, both PixelLink and PSENet tackle the text instance detection differently. They identify text instances by connecting adjacent text pixels, which allows them to handle longer text sequences more effectively. This approach enhances the model's performance in detecting extended and connected text regions.

3. METHODOLOGY

An end-to-end trainable neural network with an EAST-inspired U-shape encoder-decoder structure is provided by our proposed method. The network's output contains predictions for each pixel's likelihood of being positive as well as the position of the corresponding bounding box. To maintain simplicity and efficiency, we retain the locality-aware NMS (Non-Maximum Suppression) processing for merging bounding boxes, streamlining the model while still achieving accurate results. This U-shape architecture, combined with the locality-aware NMS, enables our network to perform text detection effectively in a straightforward and trainable manner.

The proposed approach is divided into several phases, as shown in Fig. 1. The first phase involves preparing the dataset, ensuring it is well-suited for training the model in second phase an appropriate backbone for training the model, as the backbone plays a crucial role in feature extraction. Moving on to the third phase, the text detection process takes place. Here, the trained model efficiently detects text regions in the given images, accurately localizing the text instances and last phase involves text recognition thatleveraging the proposed architecture, the model proceeds to recognize and interpret the detected text, extracting meaningful information from the images.

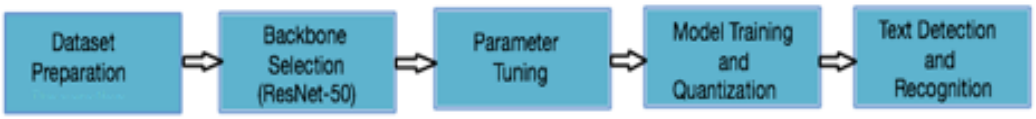

Figure 1: Scene text detection Phases

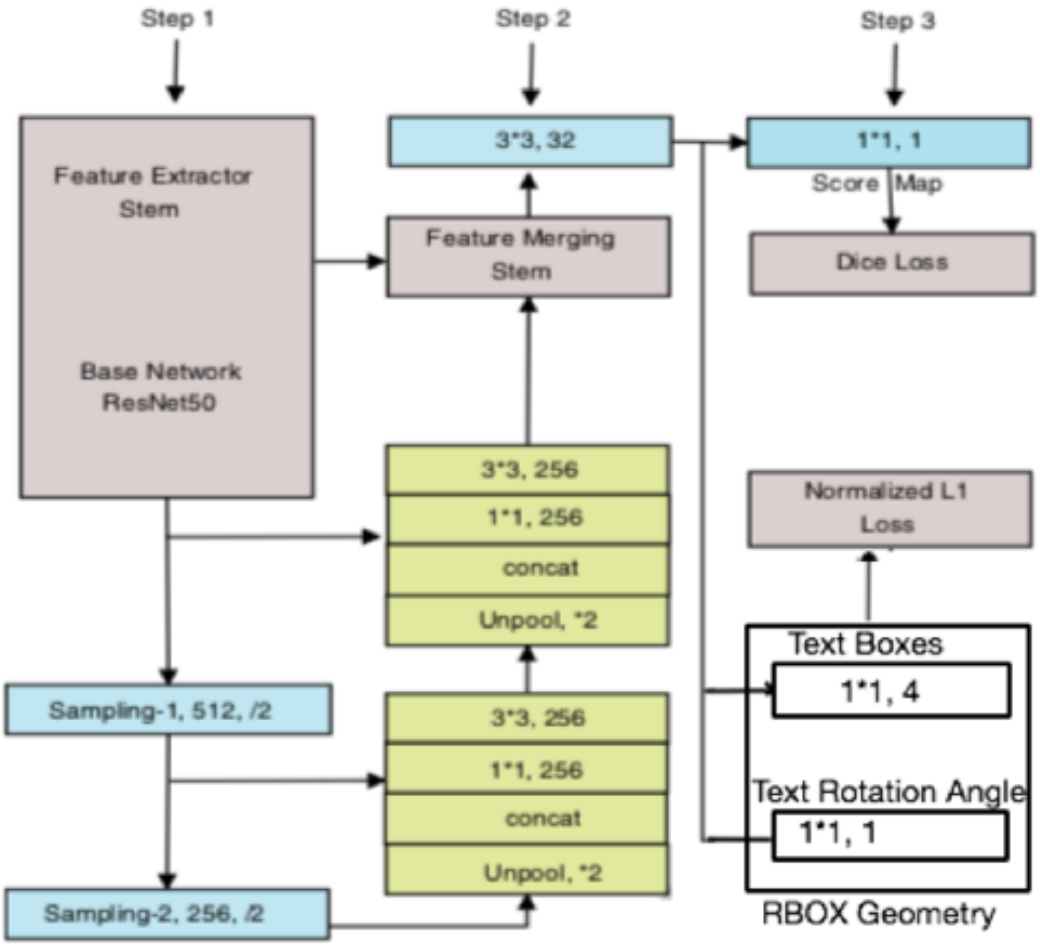

Figure 2:Architecture of EAST text detection model

- **Soft non-maximum suppression (NMS):**

Soft NMS provides a more nuanced and flexible approach to handling overlapping boxes.In soft NMS, the weight function assigns lower weights to overlapping boxes based on their proximity and degree of overlap soft NMS introduces a weight function that considers the influence of overlapping boxes on each other.

Let's assume we have two bounding boxes Box A and Box B, with scores Sa and Sb respectively and intersection-over-union (IoU) value, IoUab.

The weight function w(IoU)

w(IoU) = exp(-(IoU^2) / sigma)

The score of Box A and Box B using soft NMS

Sa_new = Sa * w(IoUab)

Sb_new = Sb * w(IoUab)

1.1 Network Design

Three main parts of the proposed network structure are a feature extractor, a component fusion module and an output layer.The component feature extractor is essential for extracting useful data from the input scene.To achieve this, we use a convolutional neural network that was

previously trained with the popular ImageNet dataset.[15] We choose the ResNet-50 network [16] specifically as our backbone. The deep network structure of the ResNet-50 makes it possible for it to extract in-depth semantic data from the input scene image.Moreover, its differential architecture effectively addresses issues like vanishing gradient or gradient explosion during training, enhancing the stability and convergence of our model. The fusion module serves as an interface between the feature extractor and the output layer. It consolidates the extracted features from the backbone network, enabling the model to capture and combine important information from various network layers. The output layer is finally responsible for producing the final predictions. It analyzes the fused features to forecast the likelihood that each pixel will be a positive and creates the location of the bounding box for each text occurrence.

By substituting all of the 33 convolutions in the final block with atrous convolutions, our method deviates from the traditional ResNet-50 network. We acknowledge the importance of contextual information in text detection, therefore we add atrous convolutions to the feature extractor to account for this. This will allow us to enhance the feedback field while maintaining spatial resolution of feature maps under our control.

The final block uses atrous convolution to lower the feature maps' spatial resolution from the initial 1/32 to 1/16 of the input image. With this change, the performance is improved overall by enhancing the detection of small texts.On utilizing the same dilation rate, the "gridding effect" is created, which results in some pixels being removed from calculations., we set different dilation rates (2, 5, 7) for the three 3×3 convolutions in the last block. Additionally, we incorporate the ASPP structure from DeepLabv3 on top of feature maps. The ASPP module, consisting of atrous convolutions with different dilation rates (3, 9, 12, 18), effectively detects texts of various scales.

To enhance computational efficiency, we utilize depth separable convolution for the convolutions within the ASPP structure.In order to include image-level data on the feature maps, we further integrate Global Average Pooling (GAP).We use the FPN structure in the feature fusion module. After being sampled by a factor of 2, equivalent low level features on the network backbone with similar spatial resolution have been combined to high level features. To reduce the number of channels in order to minimise computational costs, an 11 convolution layer has been implemented on low level features.

Through 1:1 convolution, In the output layer, a map prediction based on 1 channel's score and an 8 channel text localization forecast is automatically generated. The localization map's channels hold the four vertices that surrounds text bounding box as the coordinate offsets for each pixels, whereas the score map indicates the likelihood of pixel that contains a text instance.By employing these modifications and incorporating atrous convolutions, ASPP, and FPN structures, our approach aims to effectively capture context and features for precise and effective detection of scene text and localization.

1.2 Ground Truth Generation

a) For generating the ground truth score map:

We use a strategy similar to EAST, whichincludes assigning ground truth to text regions using a shrunk polygon technique [3].Specifically, we shrink the original text region to create a smaller polygon, and only the pixels within this shrunk area are labeled as text-positive. These pixels are considered as belonging to the text region. However, as it can include confusing or incomplete text information, the space between the reduced polygon and the real bounding box is discarded.

b) For generation for localization map :

This approach helps in discriminating text objects that are closely located, as it focuses on the precise text-positive area while excluding any potential overlaps or ambiguities. By generating the ground truth score map in this manner, our model gains a better understanding of the actual text regions in the input images, leading to more accurate and effective text detection results.

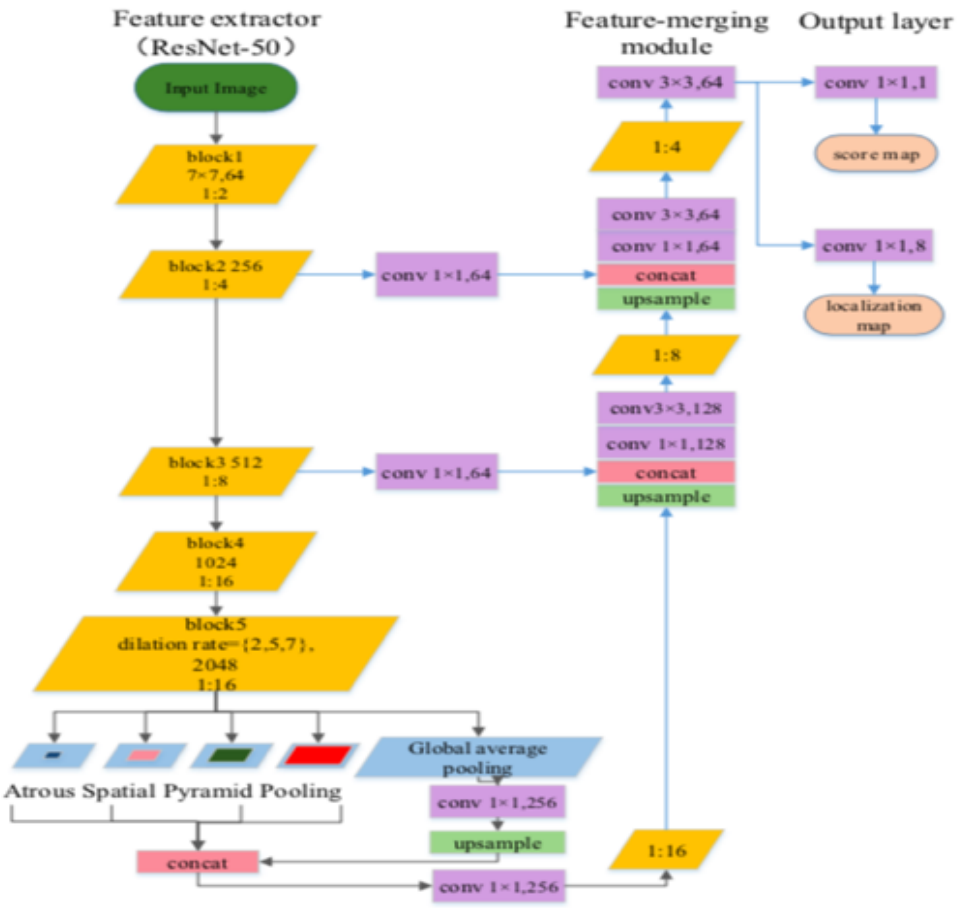

Figure 3:Network Architecture.

c) Loss Function :

The loss function can be expressed as the weighted sum of the localization and score map losses.

$L = L_s + \lambda_{loc} L_{loc}$

The terms Ls and L_{loc} respectively refer for the loss functions on a scoring map and a localization map, respectively. Our experiment's λ_{loc} parameter is set at 1.

d) Map Loss score

In EAST, we utilize a class-balanced cross-entropy loss to train the model. However, to accelerate the convergence speed during training, We also use dice coefficient loss which is used widely for semantic segmentation tasks..

$$L_s = \text{dice-coefficient}(\hat{\mathbf{P}}, \mathbf{P}^*) = 1 - \frac{2|\hat{\mathbf{P}} \cap \mathbf{P}^*| + 1}{|\hat{\mathbf{P}}| + |\mathbf{P}^*| + 1}$$

where p̂is score map prediction and P* is the actual truth. |p̂ ∩ P*|denote the point whrep̂ and P* meet . where |p̂| & |P*|are predicted pixels in the score map and actual truth

e) localization-related loss:

To handle the variability in text sizes, aspect ratios, and orientations, ensuring that our model learns to detect and localize text regions more effectively .We're using the loss L1 or L2 regression function, which makes the loss biased to a larger and longer text area, and we're using the loss L1 in the East to transpose the quadrilateral bounding box Q.

$$C_Q = \{x_1, y_1, x_2, y_2, \ldots, x_4, y_4\}$$

f) Loss function on localization

$$L_{loc} = \min \sum_{\substack{c_i \in C_{\hat{Q}} \\ \tilde{c}_i \in C_{Q^*}}} \frac{\text{smoothed}_{L1}(c_i - \tilde{c}_i)}{8 \times N_{Q^*}}$$

where $N_{Q*}$is the short age of Q* and the bounding box $\hat{Q}$of the of Q* after prediction and it is the called actual truth .

2 Investigations

a) The reference Datasets

We use one dataset to train our network and four datasets for the test.

ICDAR2019: The 1000 training samples and 500 testing examples in the ICDAR2019[17] dataset are all word-by-word annotated.The usual standardized resolution for the images in this dataset is 1280 720 pixels. On the other hand, the ICDAR2019 dataset is more diverse, consisting of samples from ten different languages. 10,000 scene pictures with ground truth files with coordinates of it's bounding box for each text word in the image are in this training data set. An image scaling method based on the long-side has been used to achieve uniformity because the image resolutions in both datasets differ.To make the images compatible with the proposed model, all images in the ICDAR2019 dataset have been resized to resolution of 720 × 1280 pixels, for maintaining it's original aspect ratios.

ICDAR 2015: There are 1000 training images and 500 test images in the ICDAR 2015 [18]Incidental Text dataset. These pictures have a low resolution and were taken with Google glasses. The dataset contains text instances with various orientations, making it challenging for text detection algorithms. The dataset's images each have word-level annotations, providing bounding box information for each text instance present. The presence of multi-oriented text instances and the low resolutions of the images pose difficulties for the task of text detection and recognition on this dataset. However, the word-level annotations enable more accurate evaluation and training of models, allowing researchers to develop robust and effective scene text detection algorithms that can handle diverse text orientations and resolutions.

MSRA-TD500 :500 photos from a pocket camera, including a mixture of interior and outside locations make up the MSRA-TD500 dataset. The inside images feature caution plates, signage, and door plates in workplace and mall settings, while the outside photographs primarily feature street sceneries with guided billboards and boards in English and Chinese. These outdoor images present complex backgrounds, posing challenges for text detection and recognition. With its diverse and real-world scenarios, the MSRA-TD500 dataset serves as a valuable resource for evaluating text detection algorithms. Researchers can use this dataset to develop and assess models that effectively handle text instances in various environments, considering factors such as orientation, language, and intricate backgrounds. By exploring this dataset, advancements in scene text detection and recognition can be made, catering to the needs of practical applications in both indoor and outdoor settings.

IIIT-ILST: The IIIT-ILST[19] dataset is a comprehensive collection of nearly 1000 real images for each script, each accompanied by annotations for scene text bounding boxes and transcriptions. This dataset is particularly valuable for researchers and programmers working on text detection and recognition algorithms for Indian languages. The ILST dataset offers a diverse range of challenges commonly encountered in scene text detection, including variations in lighting conditions, different font styles, overlapping of foreground and background elements, and occlusions. As a result, it provides an invaluable resource for developing and testing robust text detection and recognition algorithms tailored for Indian languages. The dataset is structured to include both images and XML files that store the corresponding annotation information. These annotations facilitate the evaluation of text detection algorithms using important performance metrics such as precision, recall, and F-measure. Researchers can utilize this dataset to benchmark and compare their algorithms, ensuring that their solutions are effective and accurate for a wide

range of real-world scene text scenarios in Indian languages.

- **Performance is Evaluation**

Text detection assessment index is based on Precision (P), Recall (R), and F-score (FM), which are each defined as:

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

$$F = 2\times\frac{P\times R}{P+R}$$

Where the true positive, false positive, and false negative values are denoted as TP, FP, and FN, respectively. In the context of text detection, a true positive is one in which the detected bounding box has an intersection over union with regard to the ground truth bounding box greater than the threshold. False positives are defined as incorrect bounding box predictions, and false negatives as missed bounding boxes. The trade-off between precision and recall can be measured using the F-measure.

b) ImplementationDetails

The proposed method's performance was assessed using OpenCV on a computer with a 1.8 GHz speed of processor and 16 GB RAM. The evaluation utilized the ILST [7] dataset to measure key performance metrics such as Precision (P), Recall (R)and F-measure (F). These metrics consider both true positive and false positive detections, providing insights into the approach's ability to accurately locate and categorize text regions within images. Stochastic Gradient Descent (SGD) was used to train the model with TensorFlow, and a momentum of 0.92 was used.

4. RESULTS AND DISCUSSION:

Results for three benchmark datasets using the suggested approach are shown in Fig. 4, where the quadrilateral areas stand in for the recognized text portions.In particular, the algorithm performs effectively in difficult conditions including harsh lighting, poor resolution, and image distortion.It is capable of accurately detecting long Hindi text, as showcased in Fig. 4,5,6 and 7.To further validate the proposed algorithm's effectiveness, Using the ICDAR2015, ICDAR2013, and MSRA-TD500 reference datasets, a comparison with various top text detection algorithms was conducted. Tables I and II provide a summary of the findings. It is important to note that the information in these tables represents the best outcomes each algorithm was able to provide.

Upon inspecting the tables, it becomes evident that the proposed algorithm outperforms the state-of-the-art approaches significantly in terms of precision, while maintaining comparable recall rates. This substantial improvement in precision demonstrates the effectiveness and superiority of the proposed algorithm over existing methods in text detection tasks.

Figure 4ICDAR2019 test result

Figure 5 : ICDAR2015 test results

Figure 6: ICDAR2015 test results

Figure 7 : Other script test

Table 1 : (a) and (b)

Approach	ICDR 2019		
	Recall	Precision	F-score
CPTN	0.516	0.742	0.609
PvaNet	0.852	0.732	0.782
MobileNet	0.833	0.746	0.786
Efficient Net	0.06	0.09	0.07
Resnet	0.846	0.772	0.807
Ours	0.812	0.760	0.789

Approach	ICDR 2015		
	Recall	Precision	F-score
CPTN	0.514	0.712	0.609
PvaNet	0.768	0.731	0.750
MobileNet	0.817	0.829	0.823
Efficient Net	0.735	0.836	0.782
Resnet	0.816	0.752	0.779
Our	0.783	0.862	0.821

Table 2: (a) and (b)

Approach	ICDR 2013		
	Recall	Precision	F-score
CPTN	0.511	0.612	0.629
PvaNet	0.830	0.930	0.877
MobileNet	0.830	0.877	0.853
Efficient Net	0.875	0.886	0.881
Resnet	0.827	0.926	0.874
Our	0.831	0.952	0.897

Approach	MSRA-TD500		
	Recall	Precision	F-score
CPTN	-	-	-
PvaNet	0.730	0.834	0.774
MobileNet	0.700	0.860	0.772
Efficient Net	0.732	0.830	0.778
Resnet	0.674	0.873	0.761
Our	0.754	0.882	0.813

5. CONCLUSION

This research introduces a novel multi-oriented scene text detector designed to directly locate text within images. The method's performance is enhanced by incorporating an ASPP network to improve multi-scale detection. Through extensive benchmark testing on standard datasets, the proposed approach has proven to surpass many state-of-the-art methods.The primary focus of this paper is on addressing two significant challenges in deep learning and computer vision scene text detection and quantization. To achieve this, a lightweight deep learning model based on the ResNet50 backbone has been developed for scene text detection. Compared to existing methods, the proposed model showcases a remarkable advantage by being 30 to 100 times smaller in size while maintaining a well-balanced blend of accuracy and efficiency. The model's efficacy has been validated using the ICDAR2019 dataset, which encompasses samples from 10 different languages. The potential for further expansion and improvement is evident, especially concerning mobile device applications, where the proposed model can offer significant advantages due to its compact size. Future research can explore

hybridizing two or more high-performing models to enhance overall performance. Additionally, the model's applicability can be extended to support various other languages, making it accessible to a broader range of users.

REFERENCES

[1] Divyansh Agrawal, Sachin Minocha, Suyel Namasudra, and Sathish Kumar. 2021. Ensemble Algorithm using Transfer Learning for Sheep Breed Classiication. In 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE, 199ś204.

[2] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A Single-Shot Oriented Scene Text Detector," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3676-3690, Aug. 2018.

[3] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, "EAST: an efficient and accurate scene text detector," Proc. IEEE conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp. 5551-5560.

[4] Nandam, S.R., Negi, A., Koteswara Rao, D. (2021). Telugu Scene Text Detection Using Dense Textbox. In: Sharma, H., Saraswat, M., Yadav, A., Kim, J.H., Bansal, J.C. (eds) Congress on Intelligent Systems. CIS 2020. Advances in Intelligent Systems and Computing,

[5] D Deng, H. Liu, X. Li, "Pixellink: Detecting scene text via instance segmentation," Proc. Thirty-second AAAI conference on artificial intelligence(AAAI), 2018.

[6] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, 1 April 2018, pp. 834-848.

[7] L. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprintarXiv:1706.05587, 2017.

[8] Jianxin Zhang and Yunhai Feng. 2020. Advanced Chinese Character Detection for Natural Scene Based on EAST. In Journal ofPhysics: Conference Series, Vol. 1550. IOP Publishing, 032050.

[9] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. 2018. Fots: Fast oriented text spotting with a uniiednetwork. In Proceedings of the IEEE conference on computer vision and pattern recognition. 5676ś5685.

[10] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. Textsnake: A lexible representationfor detecting text of arbitrary shapes. In Proceedings of the European conference on computer vision (ECCV). Springer, 20ś36.

[11] Minghui Liao, Baoguang Shi, and Xiang Bai. 2018. Textboxes++: A single-shot oriented scene text detector. IEEE transactionson image processing 27, 8 (2018), 3676ś3690.

[12] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape Robust Text Detection With Progressive Scale ExpansionNetwork," Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9328-9337.

[13] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. IEEEconference on computer vision and pattern recognition(CVPR), 2017, pp. 2117-2125.

[14] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi- oriented Text Detection with Fully Convolutional Networks," Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, pp. 4159-4167.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In Advances inneural information processing systems, pp. 1097–1105, 2012.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE conference on computer vision and pattern recognition(CVPR), 2016, pp. 770-778.

[17] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. 2019. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognitionÐRRC-MLT-2019. In 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 1582ś1587.

[18] A. Gupta and S. Namasudra. 2022. A novel technique for accelerating live migration in cloud computing. AutomatedSoftwareEngineering (2022).

[19] Atmakuri, V. and Dhanalakshmi, M., Advancements in Telugu Text Detection: Leveraging EAST with Soft Non-Max Suppression.

[20] Bin Li and Dimas Lima. 2021. Facial expression recognition via ResNet-50. Elsevier International Journal of CognitiveComputing in Engineering 2 (2021), 57ś64.

[21] Li-Hong Juang, Ming-Ni Wu, and Cian-Huei Lin. 2020. Afective computing study of attention recognition for the 3D guide system. CAAI Transactions on Intelligence Technology 5, 4 (2020), 260ś267.