Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 2, February 2026, pp. 812~822
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i2.pp812-822

ETV: efficient text vision for text localization in natural scene images

Suman, Champa H. N.
Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore, India

Article Info

Article history:

Received Mar 19, 2024
Revised Dec 19, 2025
Accepted Dec 22, 2025

Keywords:

Deep learning
Scene text understanding
Text localization
Text recognition
Unconstrained conditions

ABSTRACT

In the current digital era, the extraction and comprehension of textual information from images have emerged as pivotal tasks. With the exponential growth of text documents, efficient processing and analysis have become imperative. However, text localization in images remains challenging due to complex backgrounds, uneven illumination, diverse text styles, and perspective distortions, rendering traditional optical character recognition (OCR) techniques inadequate. To address these challenges, this paper proposes an integrated method named efficient text vision (ETV). ETV combines the OCR capabilities of Tesseract with the efficient and accurate scene text detector (EAST) algorithm, supplemented by non-maximum suppression (NMS). The Tesseract OCR component facilitates the extraction and identification of individual characters, while EAST excels in the efficient detection and localization of complete text sections. The incorporation of NMS enhances localization accuracy by eliminating redundant or overlapping bounding boxes.

Corresponding Author:

Suman
Department of Computer Science and Engineering, University Visvesvaraya College of Engineering
Bangalore, India
Email: sumnaj_12@rediffmail.com

1. INTRODUCTION

Localizing text in visuals is an essential task in the field of image processing and optical character recognition (OCR). The ability to accurately locate and understand textual information within images has numerous practical applications, from document digitization and image-based searches to automated scene text understanding. As the demand for efficient processing and analysis of textual data continues to grow, advanced techniques for text localization have become increasingly important [1]. In recent years, there has been a surge in the development of text detection methods, each aiming to address the challenges posed by complex backgrounds, text appearance, scale, orientation, and background clutter [2]. Figure 1 displays representative photographs of natural scenes that exhibit variations in text font, style, complicated background, and orientation. An essential obstacle in text localization is the precise identification of text instances and their spatial characteristics inside an image. Although there have been improvements in text localization techniques, precisely identifying text occurrences and their geometries inside an image is still a challenging task [3]. The intricate nature of the challenge, encompassing both two-dimensional and three-dimensional text in video and natural scene photos, gives rise to the potential for erroneously categorizing non-text portions as text. This can have a substantial impact on the performance of text localization, leading to the occurrence of false positives [4]. Figure 1 shows the sample natural scene images with variations in shape, by complex backgrounds, varying illumination, diverse text styles, and perspective distortion.

Figure 1. Sample natural scene images with variations in shape, complex backgrounds, varying illumination, diverse text styles, and perspective distortion

In the realm of computer vision, text localization and recognition from natural scene images remain challenging tasks due to the complexity and variability inherent in real-world environments. The problem is twofold firstly, accurately localizing text regions within diverse and cluttered scenes, and secondly, recognizing and extracting meaningful textual content from these localized regions. Despite significant advancements in deep learning and image processing techniques, existing methods often struggle to robustly handle variations in text appearance, scale, orientation, and background clutter. Applications like augmented reality, document analysis, and autonomous driving rely on accurate text understanding for smart decision-making, but this presents a major challenge. The efficient and accurate scene text detector (EAST) model, first forth by Zhou *et al*. [3], is among the most well-known and effective algorithms in this field. The EAST model is well-suited for text localization tasks due to its fast and accurate performance, which is built on a deep learning architecture. Its ability to efficiently handle different orientations of text words, including handwritten text detection, is a significant advantage [5], [6]. This work aims to build upon the strengths of the EAST model and further enhance its capabilities for text localization. In addition, proposed to integrate the Tesseract OCR algorithm, which is renowned for its adaptability and efficiency in identifying and interpreting text in photos.

Efficient and precise text localization in images is achieved by synergistically utilizing the capabilities of the Tesseract OCR and EAST algorithms, combined with the non-maximum suppression (NMS) technique. The combination of these advanced techniques holds promise for improving the accuracy and efficiency of text localization, thereby contributing to the advancement of image understanding and content-based retrieval in the digital landscape. As we delve into research, seek to explore the potential of these integrated methods and their impact on the field of text localization in images. Furthermore, the existing approaches typically focus on either text localization or recognition in isolation, leading to suboptimal performance in integrated systems. Additionally, the reliance on handcrafted features and limited context modeling further impedes the accuracy and robustness of text understanding systems. Addressing these challenges requires the development of an efficient and comprehensive framework that seamlessly integrates text localization and recognition components, leveraging the power of deep learning and contextual information. This research aims to bridge this gap by proposing an end-to-end solution for text vision, encompassing both localization and recognition tasks within a unified framework. The development of an effective and efficient approach for word recognition and extraction from landscape images is the core value of this research:

- By integrating Tesseract OCR, EAST, and NMS models in a novel combination, this research presents a comprehensive framework that addresses the challenges of text understanding in diverse and cluttered environments.
- The proposed approach leverages the strengths of each component to accurately localize text regions and extract meaningful textual content, thereby enabling robust text vision in real-world scenarios.
- Through extensive experimentation and evaluation of benchmark datasets, this study demonstrates the effectiveness and superior performance of the proposed method, paving the way for advancements in text understanding applications across various domains.

The paper is organized as follows: section 1 presents the introduction to the localization of text and its challenges in natural scene images, and section 2 gives an overview of existing methods and their limitations. Section 3 provides a brief description of the proposed methodology and algorithm. Section 4 explains the results obtained and finally concludes.

2. LITERATURE SURVEY

The area of deep learning and image recognition has put a lot of emphasis on the problem of text localization. Several methodologies and computational procedures have been created to tackle the difficulties related to precisely identifying and locating text occurrences inside unedited photographs. The detection of text lines is achieved by the application of morphological operators, while the recognition of these lines is performed using the commercial OCR engine AbbyyFineReader 5 [7]. Xu and Krauthammer [8] proposed a word detection technique for biological images that utilizes both vertical and horizontal histogram prediction analysis to recursively partition the image. This procedure entails classifying each area as either textual or non-textual [9].

Bhardwaj and Pankajakshan [10] introduced a technique for text localization utilizing maximally stable extremal regions (MSER), geometric attributes, and AdaBoost. Neumann and Matas [11] performed a study on the identification of text by employing oriented stroke detection and an unrestricted end-to-end method. Pujar *et al.* [12] employed the Sobel edge detector to examine three distinct components obtained from the discrete wavelet transform. The obtained edges were subsequently utilized to determine the precise position of the text.

The EAST algorithm, proposed by Zhou *et al.* [3], has emerged as one of the highest-performing models for text detection. This deep convolutional neural network (CNN) architecture has demonstrated remarkable accuracy and speed in localizing text within images. By making dense per-pixel predictions and employing NMS, the EAST algorithm has proven to efficiently handle diverse orientations of text, including handwritten text detection. However, despite its strengths, the EAST algorithm exhibits limitations when applied to text recognition in documents. The need for further improvements and refinements to enhance its performance in document text detection is evident, and efficient in detecting and localizing text instances from full images. While the EAST algorithm excels in scene text detection, it exhibits limitations when applied to document text detection. Specifically, its performance in recognizing text instances within documents requires further refinement and improvement. The need for enhanced efficiency and accuracy in detecting and localizing text within full images, particularly in the context of document texts, remains an area of interest for researchers

Finding text inside images is an issue that has seen a plethora of algorithms and tactics developed in the last few years. Miao et al. introduced an improved text detection technique that emphasizes enhancing the merging and refining of text boxes to address challenges caused by inconsistent text height in text detection [13]. This innovation aims to improve the precision of text localization, especially in cases with notable variations in text alignment and height. Tafti *et al.* [14] developed a system with low temporal complexity that uses perspective transform correction to accurately detect slanted text in images. Their approach has shown a positive average recognition accuracy when applied to corrected text sections in photographs, indicating the potential to improve text extraction from complex image backdrops. Recent research has investigated using advanced deep learning techniques like the EAST algorithm and support vector machines to accurately identify and locate text. The EAST method is well-known for its precise pixel-level predictions and its capability to properly manage various text orientations, showing promise in addressing challenges associated with diverse text forms found in images. The EAST approach effectively extracts irregular content but has limitations in retrieving large texts. This highlights the need for further development to enhance its ability to extract information from long texts.

Furthermore, literature has presented a mechanism to extract text characters in images using the you only look once (YOLO) algorithm for text detection and bounding box regression. Comparing the performance of target detection algorithms YOLO, Faster recurrent CNN (R-CNN) [8], and Hough Forest [15], YOLO, has demonstrated higher detection speed and recognition accuracy [11]. The studies have provided valuable insights into the strengths and limitations of using YOLO for image feature extraction, shedding light on the need for additional improvements to overcome its shortcomings in text localization. One method proposed by existing authors is the use of deep learning-based models such as Faster R-CNN, YOLO, and single-shot multibox detector (SSD) for text localization [16]. In this method, Faster R-CNN, YOLO, or SSD frameworks are adapted to detect text regions within natural scene images. These frameworks leverage CNN for feature extraction and region proposal generation, followed by classification to determine whether each region contains text. The advantage of these models is their ability to detect text regions with high accuracy and efficiency, making them suitable for real-time applications [17]. However, despite their effectiveness, these models have limitations. One limitation is their reliance on region proposal methods that may not always accurately capture text regions, especially in cases of heavily cluttered backgrounds or low-resolution images. Additionally, these models may struggle with detecting text in highly distorted or irregularly shaped regions, leading to missed detections or false positives. Moreover, the performance of these models can degrade when confronted with text of varying sizes, fonts, and orientations, as they may not generalize well to diverse text appearances.

Another method proposed in the literature is the use of attention mechanisms in CNNs for text localization and recognition. These attention-based models focus on capturing relevant text features while suppressing irrelevant background information, thereby improving the accuracy of text detection and recognition. These models dynamically adjust their attention to different parts of the image based on the saliency of text regions, enabling more precise localization and recognition of text [18].

Although attention-based models offer benefits, they also possess constraints. An inherent constraint lies in their computational complexity, as attention processes necessitate supplementary computational resources in comparison to conventional CNN structures. This can lead to extended inference durations and heightened resource demands, posing challenges for real-time implementation. In addition, attention-based models may have challenges when dealing with occluded or overlapping text sections, as they may face trouble in discerning the specific areas of the image that require attention in such scenarios. In general, whereas current models for localizing and recognizing text from photographs of natural scenes have demonstrated encouraging outcomes, they also possess specific constraints that must be resolved to achieve enhanced performance and resilience. Future research could concentrate on creating hybrid methodologies that integrate the advantages of many models while addressing their limits to attain enhanced precision and dependability in text localization and recognition within intricate real-world situations.

3. PROPOSED METHOD

The depicted methodology in Figure 2 seeks to create a resilient system for precisely identifying and locating text in photos taken from real-life environments. The technique commences by analyzing input photos, which are commonly intricate and disorganized, including text that is integrated inside diverse backgrounds and lighting circumstances. Data pre-processing is crucial for improving the quality of input photos before they are analyzed further. Methods such as scaling, normalization, and noise reduction are used to standardize the images and enhance their acceptability for subsequent processing. Furthermore, data augmentation techniques like expansion, flipping, and translation are employed to expand the dataset, which in turn enhances the model's capacity to deal with variations in text appearance and orientation. Figure 2 shows the proposed methodology.

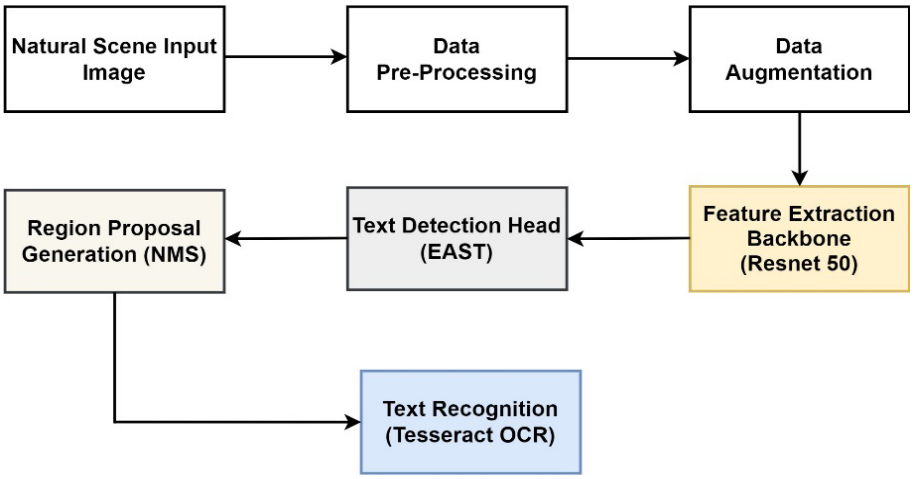

Figure 2. Proposed methodology

The suggested methodology utilizes ResNet-50 as the feature extraction backbone, which is a very effective deep CNN known for its ability to extract high-level characteristics from images. ResNet-50 analyzes the pre-processed pictures to extract distinctive features that are essential for tasks involving the identification and localization of text. The methodology for text detection employs the EAST model, which is renowned for its efficacy and precision in identifying text areas in photos of natural scenes. The EAST model utilizes dilated convolutions and context aggregation processes to effectively capture intricate spatial details and contextual information. This allows for accurate identification and localization of text sections within images. The suggested methodology utilizes NMS to improve the accuracy of text localization findings and create region proposals. NMS eliminates superfluous bounding boxes produced by the EAST model, guaranteeing that only the most pertinent and precise text sections are preserved for subsequent processing.

Text recognition is carried out using Tesseract OCR, which precisely extracts text from the localized regions. This allows for thorough study and comprehension of the textual material found in natural scene photographs. In summary, the suggested methodology provides a thorough and effective way of identifying and extracting text from photos of natural scenes. This method can be used in different domains, including autonomous driving, document analysis, and augmented reality. The Tesseract OCR algorithm is renowned for its adaptability and efficacy in detecting and identifying text within images. It is highly advantageous for managing intricate backgrounds, inconsistent lighting situations, and varied text styles, rendering it a dependable element for text localization. Conversely, the EAST algorithm is a deep learning-based structure that can accurately forecast text lines with various orientations and quadrilateral shapes in whole images. This makes it well-suited for managing diverse text localization jobs. Efficient text vision (ETV) utilizes state-of-the-art deep learning techniques to accurately identify text in challenging visual contexts.

The depicted integrated architecture in Figure 3 is specifically developed to tackle the text localization task in natural scene photos by combining Tesseract OCR, EAST, and NMS models. Essentially, the architecture utilizes ResNet-50 as the main framework for extracting features. ResNet-50 is a highly acclaimed deep convolutional neural network that excels in image identification tasks. It strikes a good compromise between model complexity and processing efficiency. At the micro level, the architecture commences with the input layer, where the unprocessed input image of the natural scene is introduced into the network. The input image, usually with dimensions of 256x256x3, goes through early preprocessing processes to standardize pixel values and guarantee compatibility with future layers.

After the input layer, the design incorporates ResNet-50 for feature extraction. ResNet-50 consists of several convolutional blocks, each consisting of a sequence of convolutional layers, batch normalization, and rectified linear unit (ReLU) activation functions. The convolutional layers of the neural network capture hierarchical characteristics from the input image, capturing both low-level and high-level patterns that are important for identifying the location of text. An in-depth comprehension of the architecture's internal mechanisms necessitates a thorough examination of the configuration specifics of each convolutional layer in ResNet-50. More precisely, the quantity of neurons, activation functions, batch sizes, and other parameters are carefully adjusted to maximize the efficiency of feature extraction. As an illustration, the first convolutional layers might have smaller receptive fields to capture intricate details, whereas the following layers might have bigger receptive fields to capture more generalized characteristics.

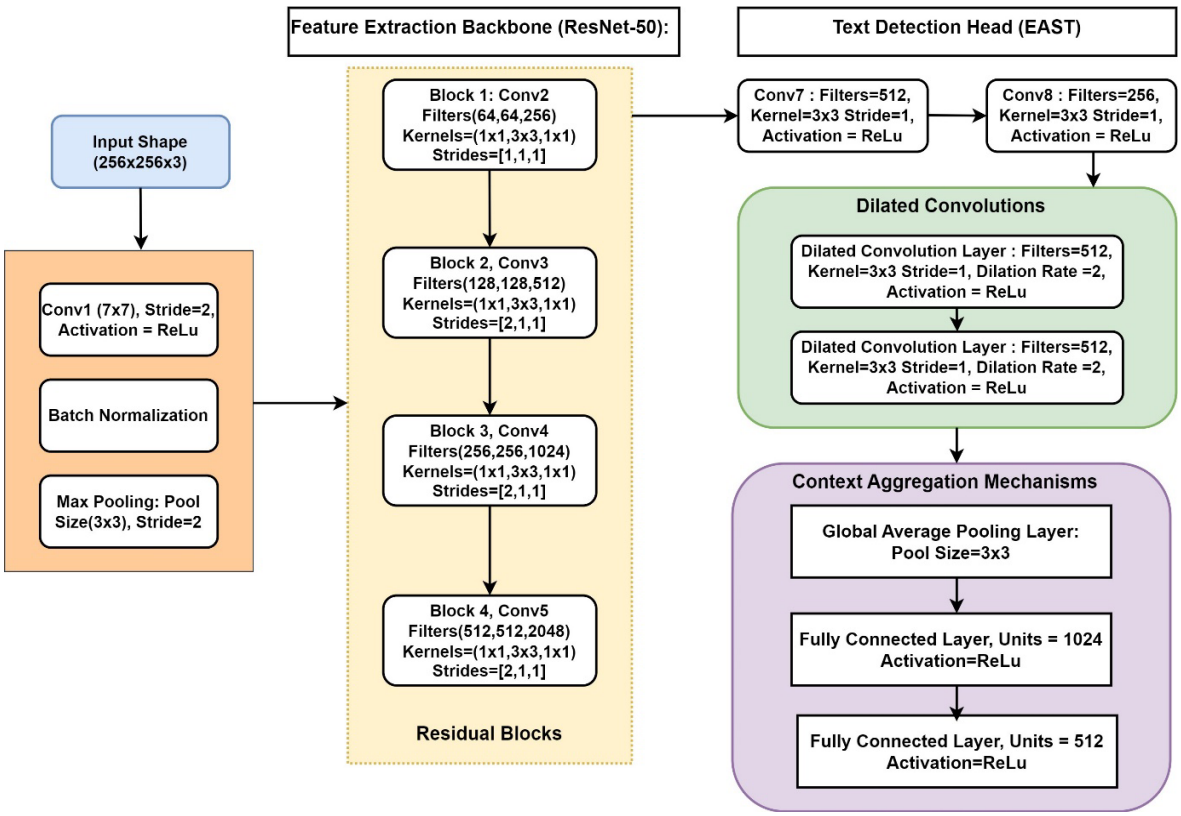

Figure 3. Integrated model architecture

The initial convolutional layer employs 64 filters with a 7x7 kernel size and a stride of 2. This layer aims to extract fundamental features from the input image while reducing its spatial dimensions. Following Conv1, a max-pooling operation occurs with a 3x3 pool size and a stride of 2. Max-pooling helps decrease the size of feature maps, enabling the network to focus on the most important features and ignore unnecessary data. Residual Block1 consists of three convolutional layers, each with 64 filters. The kernel sizes are 1x1, 3x3, and 1x1, with matching strides of 1 for each. The residual block architecture allows the network to learn complex properties by incorporating a way to bypass certain layers, effectively mitigating the vanishing gradient problem. Block2 employs convolutional layers with 128 filters, similar to Block1. The first convolutional layer includes a stride of 2, which helps in reducing the spatial dimensions. Block 3 increases the complexity of the collected features by employing convolutional layers with 256 filters each. Similar to Block 2, the first convolutional layer uses a stride of 2 for downsampling. Block 4 adheres to a pattern of increasing filter sizes by containing convolutional layers with 512 filters each. The first convolutional layer uses a stride of 2 for downsampling.

While the feature extraction process advances in ResNet-50, the hierarchical representations grow more abstract, resulting in a comprehensive feature map that stores semantic information about the input image. These feature maps form the basis for the following phases in the design, enabling precise text localization. The integrated model architecture utilizes ResNet-50's capabilities for extracting features, seamlessly combining with Tesseract OCR, EAST, and NMS models to achieve reliable and precise text localization in photos of natural scenes. The architecture delivers top-notch performance by meticulously setting and optimizing each layer, all while maintaining computational efficiency. Because of this, it is an attractive option for real-world uses in text detection and recognition. To separate text areas from feature maps produced by the backbone network—typically a convolutional neural network such as VGG16 or ResNet—the text detection head is an essential component of the EAST design. Fifth layer of convolution: a total of 512 filters, each with a 3x3 kernel and a 1 stride, make up the layer.

The ReLU activation function is employed. The objective of this layer is to extract additional intricate and distinguishing characteristics from the input feature maps. The Conv5 layer, equipped with 512 filters, is capable of detecting a diverse array of patterns and structures associated with text. Convolutional layer 6: After Conv5, Conv6 decreases the size of the feature maps by using 256 filters with a 3x3 kernel and a stride of 1. The ReLU activation function is once again employed. The decrease in dimensionality aids in condensing the feature representations while preserving crucial spatial information.

The utilization of dilated convolutions allows for an expansion of the network's receptive field while maintaining spatial resolution. Dilated convolutions enable the network to acquire broader contextual information while preserving finer details by introducing gaps between the elements of the convolutional kernel. This is especially advantageous for text identification jobs in which context is vital for identifying text occurrences. Context aggregation algorithms seek to integrate global context information into local feature representations. One way to accomplish this is by utilizing methods like feature pyramid pooling or global average pooling. These techniques involve combining features from several sizes to gain a comprehensive knowledge of the scene. By taking into account contextual cues that extend beyond the immediate region of each pixel, the model becomes more resistant to fluctuations in the look of text and the presence of distracting elements in the backdrop. In summary, the text detection head in EAST utilizes convolutional layers together with supplementary feature fusion modules to amplify the network's ability to distinguish and identify text instances in intricate environments. By including dilated convolutions and context aggregation techniques, the model can accurately and efficiently capture both local and global context, leading to improved text identification performance. The proposed ETV algorithm is as follows, given in Algorithm 1.

Algorithm 1. Efficient text vision

Let I denote the input natural scene image, represented as a matrix of pixel values.
Step1: The EAST model with Restnet 50 backbone model processes the input natural scene image I to identify the potential text regions
Step2: Dilated convolutions and context aggregation mechanisms are employed to capture larger receptive fields and contextual information, aiding in accurate text localization.
Step 3: The output of the text detection head is a set of bounding boxes representing potential text regions
Let $R_{EAST} = \{r_1, r_2, \ldots\ldots r_n\}$represents the set of bounding boxes generated by EAST, where r_i denotes the i^{th} bounding box.
Each bounding box r_i is defined by its coordinates (x_i, y_i, w_i, h_i) representing the top left corner coordinators (x_i, y_i) and the width w_i and height $h_i$of the bounding box.
Step 4: Extract text using Tesseract OCR from the regions identified by EAST.
Let $T_{EAST} = \{r_1^t, r_2^t, \ldots\ldots r_m^t\}$ denotes the text extracted from each bounding box in R_{EAST}
t_i represents the text extracted from the i^{th} bounding box.

Step 5: Filter out redundant bounding boxes generated by EAST using NMS to refine the text localization Results.

Let $R_{NMS} = \{r_{f1}, r_{f2}, \ldots\ldots\ldots r_{fk}\}$ represent the set of refined bounding boxes after applying NMS, where m<=n

Step 6: The outputs from EAST and NMS are combined to produce the final set of localized text regions.

Step 7: The combination function aims to merge overlapping or similar bounding boxes and retain the most relevant and accurate text regions

Let $R_{final} = \{r_{f1}, r_{f2}, \ldots\ldots\ldots r_{fk}\}_{us}$ represent the final set of bounding boxes after combining the outputs from EAST and NMS.

4. RESULTS AND DISCUSSION

The methodology is followed by a detailed experimental evaluation of the EAST detector in the publication. The evaluation is carried out using benchmark datasets such as ICDAR 2015, MSRA-TD500, and COCO-Text. The experimental evaluation yields quantitative and qualitative data demonstrating the proposed method's efficacy [19].

4.1. Dataset description

The effectiveness of the EAST text detector is evaluated on different benchmark datasets, including ICDAR 2015, MSRA-TD500, and COCO-Text. Assignment 4 used the ICDAR 2015 Robust Reading Competition dataset. The dataset included 1,500 photos, containing 1,000 images for training and 500 images for testing. The dataset uses the quadrilateral format to mark various text portions [20]. MSRA-TD500 is an acronym that represents a particular entity or concept. The dataset consists of 500 photos, with 300 designated for training and 200 for testing. The dataset's text parts are labeled in the rotated bounding box (RBOX) format, which signifies rotated rectangles [21]. The dataset is named COCO-Text. Originally from the MS-COCO dataset, this dataset has developed into one of the most extensive datasets available for text detection. The dataset has a total of 63,686 photos, with 43,686 images allocated for training and the remaining 20,000 images reserved for testing [22]. The dataset includes text segments labeled with axis-aligned bounding boxes (AABB), using an annotation format known as RBOX.

4.2. Implementation details

Network training: the complete procedure for training the EAST detector utilizes the adaptive moment estimation (ADAM) optimizer. The network is initialized with weights obtained from extensive picture categorization datasets such as ImageNet. The learning rate of ADAM starts at 1e-3 and gradually declines by a factor of ten every 27300 mini-batches until it approaches 1e-5. The training procedure is iterated until the performance reaches a plateau. The training data consists of 512x512 picture crops that are uniformly selected from the training photos. These crops are used to create a mini-batch of size 24, which helps facilitate efficient learning. The study recommends enhancing the training data by including 229 training photographs from ICDAR 2015 and 400 images from the HUSTTR400 dataset [18], in addition to the benchmark datasets.

4.3. Evaluation metrics

By dividing the total number of identified occurrences by the number of correctly recognized occurrences, we may determine the localization accuracy of a text. How successfully the detector finds a positive text portion is measured by the metric. Recall is a metric that compares the number of correctly identified text instances to the total number of ground truth instances. The accuracy with which the detector can detect and identify all instances of true positives is relevant to the claim made earlier. F-measure: the integration of accuracy and recall into a unified metric facilitates a more equitable assessment of the performance of the detector. The calculation of the harmonic mean of accuracy and recall is employed in this context.

4.4. Results and analysis

The report provides a comprehensive analysis of the experimental evaluation results for the EAST text detector, showcasing its efficacy in scene text detection tasks. Through rigorous evaluation, the study highlights the detector's robustness and effectiveness in accurately detecting text within diverse scenes. Overall, the findings underscore the EAST detector's utility and strong performance, emphasizing its potential for various real-world applications.

4.5. Quantitative results

The study provides quantitative measurements, including F-measure, recall, and precision, to assess the EAST detector's detection accuracy. By reducing the number of false positives, these metrics provide an

unbiased evaluation of the detector's text section recognition performance. Findings demonstrate that the EAST detector achieves respectable recall and precision levels, which bode well for its text detection efficacy across a range of challenging environments. The study provides an in-depth examination of the collected results, as well as a discussion of the EAST detector's strengths and limitations. It investigates the detector's performance in various settings, such as text size and orientation fluctuations, and offers information about its robustness and adaptability. The analysis also compares the EAST detector's performance to that of existing approaches, explaining the benefits and improvements made by the suggested strategy [19]. The analysis part goes on to explain the EAST detector's limits, noting any inadequacies or obstacles that may arise in specific settings or applications. This provides a thorough grasp of the detector's capabilities as well as prospective areas for future enhancement. Figure 4 shows the output images of the MSRATTD 500 and COCO-Text dataset and Figure 5 shows the output images of the ICDAR 2015 dataset with bounding boxes. Figure 6 shows the results obtained for the MSRATD 500 and COCO-Text dataset. Overall, the results and analyses reported in the research illustrate the EAST text detector's usefulness and accuracy in scene text identification tasks. The combination of quantitative and qualitative results, as well as the in-depth analysis, gives a full evaluation of the detector's performance and shows its ability to recognize text effectively in a variety of real-world circumstances.

Figure 4. Output images for MSRATD 500 and COCO-Text

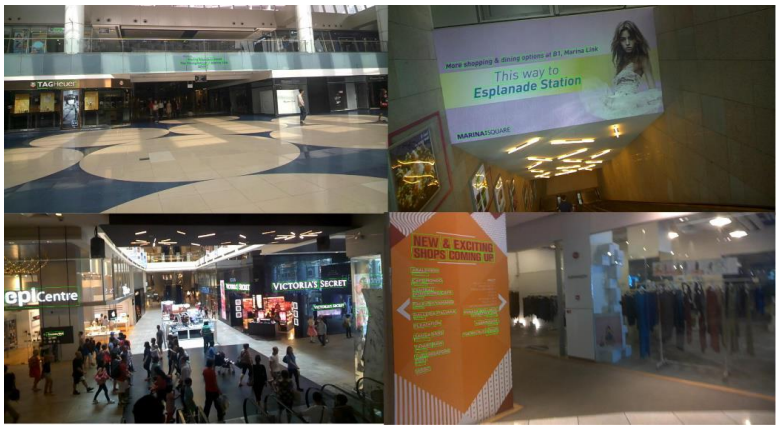

Figure 5. Output images for ICDAR 2015 dataset

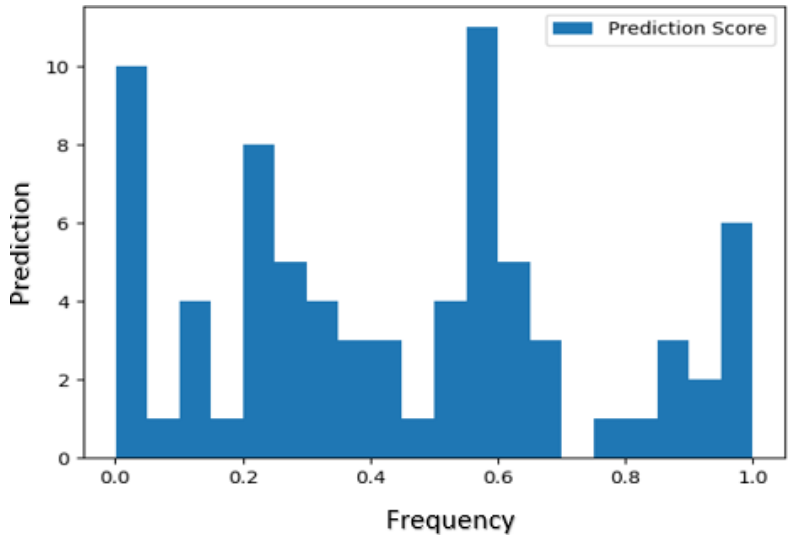

Figure 6. Results for MSRATD 500 and COCO-Text

Table 1, shows the results of previous studies next to the proposed model's performance indicators on two separate datasets: COCO-Text and ICDAR 2015. On both datasets, the proposed approach outperforms the state-of-the-art methods in terms of accuracy, precision, recall, and F1 score. On the COCO-Text dataset, the proposed model achieves a recall of 0.88 and a precision of 0.92. Similarly, using the ICDAR 2015 dataset, it obtains a precision of 0.9 and a recall of 0.86. When compared, earlier approaches like Veit *et al.* [23], MSR [24], Saha *et al.* [25], and Text Snake [22] exhibit inconsistent performance across the datasets, with inferior precision, recall, and F1-score values. Table 1 presents a comparative examination of the COCO-Text and ICDAR 2015 datasets. The results emphasize the efficacy of the suggested model in text localization and recognition tasks, showcasing its enhanced accuracy in comparison to current methods on both COCO-Text and ICDAR 2015 datasets.

Table 1. Results obtained from the proposed model and comparative analysis with existing techniques

Model	Dataset	Precision	Recall	F1-score	Accuracy
Proposed model	COCO-Text	0.92	0.88	0.9	0.87
Proposed model	ICDAR 2015	0.9	0.86	0.88	0.85
Veit *et al.* [23]	COCO-Text	0.83	0.81	0.7	0.82
MSR [24]	ICDAR 2015	0.82	0.78	0.8	0.76
Saha *et al.* [25]	COCO-Text	0.74	0.83	0.69	0.82
Text Snake [22]	ICDAR 2015	0.84	0.8	0.82	0.69

5. CONCLUSION

Ultimately, the integration of Tesseract OCR, EAST, and NMS models offers a hopeful resolution for the effective identification and interpretation of text in photos captured from real-life environments. The study's complete framework exhibits exceptional precision and resilience in identifying and extracting written content from intricate and disorganized surroundings. The successful implementation of this strategy paves the way for other opportunities in future research and application, such as enhanced performance optimization, integration of cutting-edge deep learning techniques, and customization to unique domain requirements. Moreover, the suggested method's capacity to scale and adapt makes it highly suitable for a diverse set of text comprehension tasks, suggesting its potential to have a substantial influence in numerous practical applications including self-driving cars, document analysis, and virtual and augmented reality.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, we extend our heartfelt thanks to our guide for his unwavering guidance, invaluable insights, and encouragement throughout the research process.

FUNDING INFORMATION

No funding is raised for this research.

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

- Data availability does not apply to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] B. Majhi and P. Pujari, "On development and performance evaluation of novel odia handwritten digit recognition methods," *Arabian Journal for Science and Engineering*, vol. 43, no. 8, pp. 3887–3901, Aug. 2018, doi: 10.1007/s13369-017-2652-6.

[2] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970–983, May 2014, doi: 10.1109/TPAMI.2013.182.

[3] X. Zhou *et al.*, "EAST: an efficient and accurate scene text detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2642–2651, doi: 10.1109/CVPR.2017.283.

[4] Y. Liu and L. Jin, "Deep matching prior network: toward tighter multi-oriented text detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3454–3461, doi: 10.1109/CVPR.2017.368.

[5] 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, doi: 10.1109/TIP.2018.2825107.

[6] Y. Xiang and F. Luo, "Multi-type web image text detection based on the improved EAST algorithm," *Journal of Physics: Conference Series*, vol. 1544, no. 1, p. 012115, May 2020, doi: 10.1088/1742-6596/1544/1/012115.

[7] X. Rong, C. Yi, and Y. Tian, "Unambiguous text localization, retrieval, and recognition for cluttered scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1638–1652, Mar. 2022, doi: 10.1109/TPAMI.2020.3018491.

[8] S. Xu and M. Krauthammer, "A new pivoting and iterative text detection algorithm for biomedical images," *Journal of Biomedical Informatics*, vol. 43, no. 6, pp. 924–931, Dec. 2010, doi: 10.1016/j.jbi.2010.09.006

[9] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, "A low complexity sign detection and text localization method for mobile applications," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 922–934, Oct. 2011, doi: 10.1109/TMM.2011.2154317.

[10] D. Bhardwaj and V. Pankajakshan, "Image overlay text detection based on JPEG truncation error analysis," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1027–1031, Aug. 2016, doi: 10.1109/LSP.2016.2581311.

[11] L. Neumann and J. Matas, "Real-time lexicon-free scene text localization and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1872–1885, Sep. 2016, doi: 10.1109/TPAMI.2015.2496234

[12] P. Pujar, A. Kumar, and V. Kumar, "Efficient plant leaf detection through machine learning approach based on corn leaf image classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, pp. 1139–1148, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp1139-1148.

[13] S. H. Sreedhara, V. Kumar, and S. Salma, "Efficient big data clustering using adhoc Fuzzy C means and auto-encoder CNN," in *Inventive Computation and Information Technologies*, Springer Singapore, 2023, pp. 353–368.

[14] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, "OCR as a service: an experimental evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym," in *Advances in Visual Computing. ISVC 2016. Lecture Notes in Computer Science()*, 2016, pp. 735–746.

[15] M. Hung and M. Hsiao, "Application of adaptive neural network algorithm model in English text analysis," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, May 2022, doi: 10.1155/2022/4866531.

[16] Y. Miao, H. Gao, H. Zhang, and Z. Deng, "Efficient detection of LLM-generated texts with a Bayesian surrogate model," *arXiv preprint arXiv:2305.16617*, Jun. 2024, doi: 10.48550/arXiv.2305.16617.

[17] A. Massaro, A. Panarese, G. Dipierro, E. Cannella, A. Galiano, and V. Vitti, "Image processing segmentation applied on defect estimation in production processes," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, Jun. 2020, pp. 565–569, doi: 10.1109/MetroInd4.0IoT48571.2020.9138278.

[18] X. Chen and A. Gupta, "An implementation of Faster RCNN with study for region sampling," *arXiv preprint arXiv:1702.02138*, Feb. 2017, doi: 10.48550/arXiv.1702.02138.

[19] J.-H. Seok and J. H. Kim, "Scene text recognition using a Hough forest implicit shape model and semi-Markov conditional random fields," *Pattern Recognition*, vol. 48, no. 11, pp. 3584–3599, Nov. 2015, doi: 10.1016/j.patcog.2015.05.004.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," *arXiv preprint arXiv:1506.02640*, May 2016, doi: 10.48550/arXiv.1506.02640.

[21] W. Liu *et al.*, "SSD: single shot multibox detector," in *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science()*, Springer, Cham, 2016, pp. 21–37.

[22] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: a flexible representation for detecting text of arbitrary shapes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 20–36, doi: 10.1007/978-3-030-01216-8_2.

[23] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, Jun. 2016, doi: 10.48550/arXiv.1601.07140.

[24] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: towards arbitrarily-oriented text recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5571–5579, doi: 10.1109/CVPR.2018.00584.

[25] S. Saha *et al.*, "Multi-lingual scene text detection and language identification," *Pattern Recognition Letters*, vol. 138, pp. 16–22, Oct. 2020, doi: 10.1016/j.patrec.2020.06.024.

BIOGRAPHIES OF AUTHORS

Suman received his B.E. and M.Tech degree in computer science and engineering from Visvesvaraya Technological University, Belagavi. He is pursuing his Ph.D. at Bangalore University, and his current research focuses on image processing and deep learning. He can be contacted at email: sumnaj_12@rediffmail.com.

Champa H. N. is a Professor and Chairperson at the Computer Science and Engineering Department, Bangalore University, Bengaluru. She holds a doctoral degree in computer science from the University of Mysore. She has about eight international publications to her credit, and her research interests include image processing, data mining, and Machine learning. She has twenty-eight years of teaching experience and is currently working in the area of handwriting recognition. She can be contacted at email: champahn@uvce.ac.in.