

A deep learning framework for robust text extraction from heterogeneous image sources

Dayananda Kodala Jayaram^{1,2}, Puttegowda Devegowda¹

¹Department of Computer Science and Engineering, ATME College of Engineering affiliated to Visvesvaraya Technological University, Belagavi, India

²Department of Computer Science and Engineering, Vidyavardhaka College of Engineering affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Jul 10, 2025

Revised Jul 30, 2026

Accepted Aug 12, 2026

Keywords:

Artificial intelligence

Convolutional neural network

Deep learning

Optical character recognition

Scene text recognition

Transformer

ABSTRACT

Text extraction from an image is characterized by various challenges that result in various evolving solutions at current times. Existing solutions using artificial intelligence (AI) do have some promising outcomes, but also have their own shortcomings associated with accuracy and efficiency. The manuscript introduces a simplified and yet innovative AI framework towards extracting text by unifying character region-aware text detection. The model uses a convolutional neural network (CNN) along with transformer-centric optical character recognition (OCR) towards better performance on scene text recognition. Analysis is carried out on the standard and benchmarked International Conference on Document Analysis and Recognition 2015 (ICDAR 2015) dataset to find that the proposed model accomplished 94.7% accuracy and 0.82 seconds per image of response time, which are much significant improvements in performance, while observed in contrast to existing AI models, proving the model to deliver high-performance and cost-effective outcomes.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Dayananda Kodala Jayaram

Department of Computer Science and Engineering

ATME College of Engineering affiliated to Visvesvaraya Technological University

Bannur Road, Mysuru-570028, Karnataka, India

Email: dayananda.kem@gmail.com

1. INTRODUCTION

The digital process of turning different document types into machine-readable text is known as text extraction [1]. Text extraction is crucial for data management, search, and storage, particularly for large organizations that hold thousands of important documents like invoices and contracts. To separate any background writing, the image may also benefit from being binarized, or converted to a black and white format. Depending on the complexity of the data, text can be detected and extracted from the photos using optical character recognition (OCR) or machine learning algorithms [2].

Natural language processing (NLP) techniques may be employed in the final stage to fix any formatting problems that may have arisen throughout the extraction process. For instance, characters may have been encoded incorrectly, or specific phrases or letters may have been overlooked during the detection and extraction procedure [3]. The precision and dependability of the results may be affected by different challenges related to text extraction from images. Poor image quality can obfuscate text and make it challenging for OCR systems to accurately recognize characters [4].

Recognition is made more difficult by variations in font style, size, and orientation, including handwritten or stylized writing. Accurate extraction may also be hampered by background clutter and poor

text-to-background contrast. In addition, complex processing techniques are needed to effectively comprehend multilingual text (MLT) and complicated layouts, as those found in magazines or forms with numerous columns and tables. In real-world situations, robust and generalized text extraction is a difficult operation due to an array of lighting situations, perspective distortion, and occlusions.

Artificial intelligence (AI) is essential for increasing text extraction from photos since it increases efficiency, accuracy, and adaptability. AI-powered systems, in contrast to conventional OCR techniques, employ deep learning models to better understand handwritten or deformed text, manage a variety of fonts, and identify intricate patterns [5]. These models are more resilient in practical applications since they can also adjust to different image circumstances like noise, blur, or low contrast. AI also makes context-aware text recognition possible, which helps systems comprehend the structure and meaning of content that has been extracted. This is particularly helpful for documents with intricate structures [6].

The related work reviewed in this process is as follows. From the perspective of text extraction from images, various studies have been performed. Malashin *et al.* [7] have extracted text from images while using NLP for information processing, along with the adoption of a genetic algorithm and OCR. The study towards leveraging OCR is noted in the work of Patil *et al.* [8], where semantic segmentation has been used. Ueda *et al.* [9] have presented a captioning method based on text that uses OCR, while Bisen *et al.* [10] have extracted text from a document image using JPEG2000 compression, and maximally stable external regions have been used for layout extraction.

Kim and Kim [11] have introduced a mechanism for encoding the distance information captured from the scene text associated with text boundaries. Existing studies towards scene text recognition using convolutional neural networks (CNN) were emerging. Huang and Feng [12] have proposed a self-attention method for capturing scene text, where a CNN is used for extracting local features. Tang *et al.* [13] have developed a computational model that can perform scene text recognition, while the adoption of a deep convolutional recurrent neural network has been seen in the work of Chandio *et al.* [14] with the idea to maximize the text recognition accuracy. In text-based image editing/synthesis-based methods, Park *et al.* [15] have used a U-Net structure as well as a pre-trained model for performing text-based editing associated with the input image.

Wen *et al.* [16] have used a generative adversarial network and developed a synthesis network for improving text-to-image synthesis. Zhao *et al.* [17] have used a masked transformer to carry out editing of textual content from the image, which targets text-to-image generation applications. In models towards text classification and NLP-centric tasks, He and Abisado [18] have used a CNN integrated with bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT) towards performing sentiment analysis of text. A hybrid version of CNN and long short-term memory (LSTM) has been reportedly presented in the work of Hasib *et al.* [19] for text classification using news data. Yang and Streib [20] have performed text classification of multi-labels using CNN along with thresholding for classifying medical data. There are also studies towards a transformer-based approach. Such an approach has been adopted in the work of Geng [21] for both detection and recognition of text. Adoption of a transformer is also noted in the work of Liu *et al.* [22], where text summarization is carried out. Takahashi and Kuriyama [23] have used a transformer towards the generation of layouts of text, along with an inclusion of a parallel decoding process. Finally, masked region-based convolutional neural network (R-CNN) for text localization has also witnessed various studies. Kiatphaisansophon *et al.* [24] have used R-CNN for the identification of the bounding box of text. Similar adoption of masked R-CNN is also noted in the work of Droby *et al.* [25], extracting textual lines from historical documents.

Irrespective of various advancements towards techniques of scene text extraction, there are various research problems as follows: i) the prime limitation is associated with varying lighting conditions and complex backgrounds which act as a hurdle for robust text extraction, ii) there is an increasing limitation towards contextual interpretation and semantic understanding, especially in noisy document or scene text image, iii) real-time processing with higher accuracy is actually challenging, especially for CNN and transformer models, which are actually potential out of all the AI-based models, and iv) the computational demands associated with CNN and transformer are quite high. Hence, the proposed system addresses this limitation by evolving a new design that uses both a convolution layer and a transformer in a new way and is a better substitution for existing methods. The proposed model aims to design a robust AI model that can perform detection and recognition of text from varied forms of image types, including text in them. The contributions of the proposed model are as follows: i) the proposed study combines multiple AI-based models for constructing an integrated system to be used for both detection and recognition, ii) the proposed model is trained using diverse image types to emerge as a comprehensive, adaptable, and generalizable text-extraction model, and iii) the assessment has been carried out on a benchmarked dataset and compared with frequently adopted AI models to prove its effectiveness. The organization of the manuscript is as follows: section 2 presents a discussion of the methods, section 3 presents a discussion of the results and interpretation, and section 4 presents the study conclusion.

2. METHOD

The prime motive of the proposed study is to develop a novel, simplified, cost-effective, and integrated modelling for text detection from a video dataset. The study uses an AI-driven approach in the actual methodology implementation to ensure that it can perform the detection of text from any form of complicated images. This task is performed using the inclusion of various essential modules, e.g., preprocessing, detection of text, recognition of text, and a deep learning model. The architecture is as shown in Figure 1.

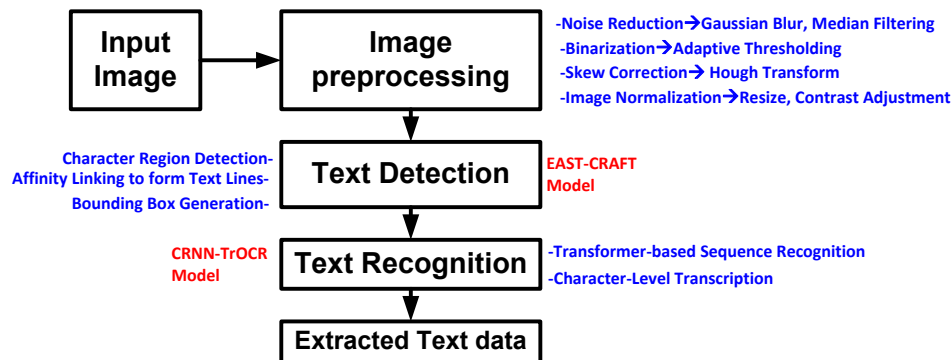


Figure 1. Architecture of proposed model

2.1. Image dataset collection and types

The prime aim of this module is to collect and make use of a diverse set of images consisting of text for depicting the practical world events that demand text extraction. For this purpose, the study initially performs an assessment of various publicly available datasets, which include artificially degraded images, handwritten notes, complex natural scenes, and clean scanned documents. According to initial observations, it is noted that the International Conference on Document Analysis and Recognition (ICDAR) 2013 consists of images with printed documents, while ICDAR 2015 and the common objects in context (COCO)-text dataset consist of natural scenes. Finally, the ICDAR 2015 dataset has been chosen, which is subjected to artificial noise and blurring to simulate real-world conditions.

2.2. Preprocessing operations

This is the second module of the proposed system that targets to perform a preprocessing operation with an agenda to improve the clarity and quality of an input image for accomplishing better reliability and accuracy while performing detection and recognition of text within the image document. Various internal operations are performed in this step, which include the adoption of median filtering and Gaussian blur for the reduction of noise, adaptive thresholding for binarization, and Hough transform for correcting skew so that text is properly aligned horizontally. Such operations of preprocessing facilitate clarity in realizing text against its background, which is much necessary for incorporating robustness. Different from conventional approaches that are found to depend mainly on manual and static preprocessing steps, this module combines both automated and adaptive preprocessing operations customized to diverse types of images. These steps of implementation make it more suitable for different levels of noise, distortions, and conditions of illumination. The novelty of this approach resides in its integrated adoption of different multiprocessing methods optimized for diverse image types that further enable the AI model to acquire text more precisely compared to conventional preprocessing operations.

2.3. Text detection

This is the third module of implementation that is meant for precise identification as well as localization of the essential region that consists of textual contents from the set of complex and diverse images. The proposed study uses efficient and accurate scene text detector (EAST) [26] and character region awareness for text detection (CRAFT) [27], which are state-of-the-art deep learning models for carrying out this text detection task. The prediction of word or bounding boxes on a text line is carried out by EAST using a fully convolutional network, while the detection of standalone regions of characters, along with linking them, is carried out by CRAFT. Hence, it assists in the precise localization of any form of text within an image. This module is then subjected to training, followed by performing a data augmentation method (color jitter, rotation, and arbitrary cropping) for enhancing robustness. The spatial relationships and learned

features are leveraged by this module, different from conventional heuristic methods, enabling it to efficiently manage complex backgrounds, fluctuating font styles, and distortions. The novelty of this model resides in the joint usage of complementation detection methods to retain an equilibrium between accuracy and efficiency while performing the text localization operation.

2.4. Text recognition

This module is responsible for precisely transcribing the textual content residing within the identified location obtained from the previous module, with an agenda to facilitate logical interpretation. The study uses convolutional recurrent neural network (CRNN) [28] and transformer-based optical character recognition (TrOCR) as deep learning architectures. The task of CRNN is to aggregate the convolution layers for carrying out feature extraction with recurrent layers for performing sequential modelling. This makes the model specifically suitable and capable for the sequences of characters present within an image with varying layouts and fonts. The proposed system also uses a transformer-based encoder-decoder model known as TrOCR for facilitating a framework towards end-to-end recognition that is capable of handling both scene text and handwritten text, which are usually of irregular and complex types, respectively. A large synthetic dataset is used for pretraining the model, while it is fine-tuned on domain-specific samples for improving its accuracy and adaptability. Different from a conventional OCR system that depends upon shallow learning methods of handcrafted features, this module supports and encourages a deep neural network for learning enriched representation from data, resulting in potentially enhanced accuracy of recognition. The novelty of this approach resides in its combination of transformer-based architecture with a complementary recognition module to offer precise contextual understanding, offering highly accurate transcription in difficult scenarios.

2.5. Artificial intelligence models used

The proposed system uses two different types of AI models specifically for text detection and text recognition, thereby leveraging the architectures of deep learning models. The following is the briefing:

- i) For text detection: the proposed system uses CRAFT, which offers superior identification of individual character-level regions, followed by linking them to generate complete text lines or words. Different from conventional bounding box techniques, CRAFT offers favorable detection of text with arbitrary shapes and curved text, too, which makes it quite effective in the presence of scene text and complex layouts. The prime potential of this AI module is towards increased precision level in the localization of text, even in the presence of a noisy background.
- ii) For text recognition: the proposed system, TrOCR [29], is a potential transformer-based OCR technique introduced by Microsoft. The method uses a transformer architecture with encoding and decoding, which is nearly equivalent to techniques adopted in the current form of large language modelling, e.g., BERT and generative pre-trained transformer (GPT). This module performs translation of image features into character sequences, thereby facilitating text-recognition end-to-end. Apart from this, the model of TrOCR offers better recognition performance for handling text with handwritten inputs, variable text length, and irregular fonts, which is mainly due to its potential contextual understanding and modelling capabilities.

3. RESULTS AND DISCUSSION

This section presents a discussion of the accomplished outcomes of the study while complying with the adopted methodology discussed in the prior section. The scripting is carried out in Python using Jupyter Notebook on a normal 64-bit Windows machine. The discussion of the dataset, accomplished outcomes, and actual interpretation is carried out in this section.

3.1. Dataset description

The assessment of the proposed study is carried out using a benchmarked ICDAR 2015 dataset [30], which is specifically suitable for real-world recognition and detection of text, as it consists of challenging images of complex form under both unconstrained and natural conditions. This dataset consists of 1,500 images that are captured from Google Glass, while the images consist of text with different fonts, sizes, and orientations that are influenced by low resolution, uneven lighting, occlusion, and motion blur. Various types of word-level bound boxes, as well as corresponding transcription, are present in the annotated dataset that facilitate both detection and recognition operations. The rationale behind choosing this dataset originates from its diverse and realistic contents that support the model, and more suitable conditions apart from those found in scanned documents.

3.2. Accomplished results

The proposed AI model for text extraction is compared with three different models, viz., i) conventional optical character recognition system (COCR), ii) CNN-based text identification and

recognition (CTIR), and iii) transformer-based text recognition system (TTRS). The adoption of these existing models is based on their frequent adoption in the literature. The evaluation is carried out using accuracy, precision, recall, and response time (seconds per image). Table 1 exhibits the study outcome benchmarked with some existing potential methods to find that the proposed model accomplishes 94.7% of the highest accuracy in contrast to COCR and minimizes response time to potentially 0.82 seconds per image. Figure 2 graphically shows the outcome accomplished by the proposed model to understand its effectiveness towards text extraction methods in contrast to other conventional schemes. Figure 2(a) shows the accuracy, Figure 2(b) presents the precision, Figure 2(c) depicts the recall, and Figure 2(d) presents the response time.

The outcome of the proposed system has also been evaluated with multiple datasets as exhibited in Table 2 to understand the generalization capability. Apart from the ICDAR 2015 dataset, the COCO-text dataset [31] consists of complex images with cluttered backgrounds and MLT. The MLT dataset [32] involves MLT with varying distortion, orientation, and environmental conditions. The numerical trend of Table 2 shows consistency in proposed performance towards detection and recognition on all datasets. A closer look into the numerical scores will also show slightly degraded performance for COCO-text and MLT datasets with respect to accuracy, yet the proposed model could maintain consistency. This is mainly due to the integration of transform-based TrOCR for recognition and CRAFT for localization in the proposed model, leading to adaptive performance towards handling the visually complex structure of text, irregular orientation, multilingual scripts, and distorted boundaries.

Table 1. Comparative performance of proposed model and existing systems

Model	Accuracy (%)	Precision (%)	Recall (%)	Response time (s)
COCR	79.3	81.0	78.5	1.75
CTIR	85.6	87.2	84.9	1.20
TTRS	90.1	91.5	89.0	1.05
Proposed model	94.7	95.8	94.0	0.82

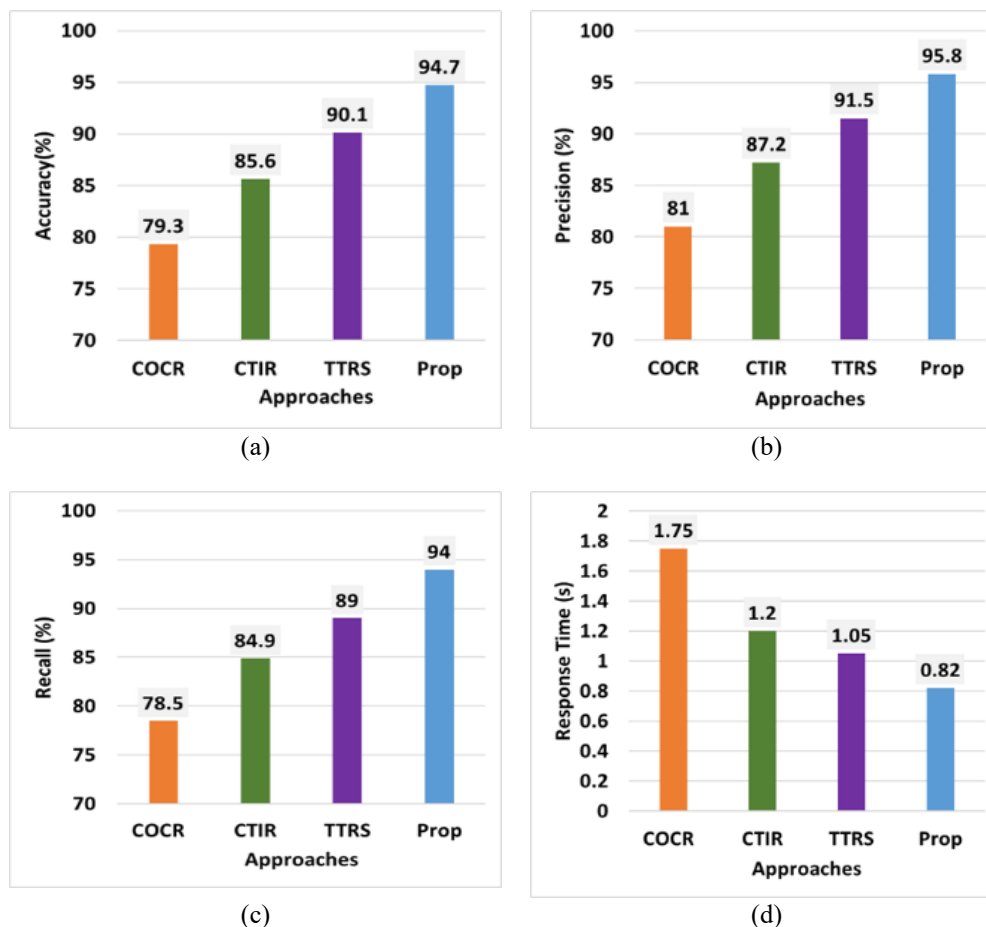


Figure 2. Accomplished study outcomes: (a) accuracy, (b) precision, (c) recall, and (d) response time

Table 2. Cross-dataset validation of proposed framework

Dataset	Challenging condition	Accuracy (%)	Precision (%)	Recall (%)	Response time (s)
ICDAR2015	Standard scene text	94.7	95.8	94.0	0.82
ICDAR2015	Motion blur	91.9	92.8	91.1	0.86
ICDAR2015	Partial occlusion	90.8	91.6	89.9	0.88
ICDAR2015	Stylized fonts	89.7	90.5	88.9	0.90
COCO-Text	Complex background	92.8	93.6	92.1	0.89
COCO-Text	Motion blur	89.5	90.4	88.6	0.93
ICDAR2017 MLT	Multilingual text	91.6	92.4	90.8	0.94
ICDAR2017 MLT	Stylized and curved text	88.9	89.8	88.0	0.97

3.3. Discussion of results

Figure 2 represents the graphical outcome of the study with respect to ICDAR 2015 dataset, as there is no much significant difference in performance on other datasets. From the perspective of accuracy outcomes, it is noted that a total of 94.7% accuracy has been witnessed to be accomplished by the proposed AI model for text extraction. This score is 19.4% higher than the COCR approach (accuracy = 79.3%), 9.1% higher than the CTIR approach (accuracy = 85.6%), and 4.6% higher than the TTRS approach (accuracy = 90.1%). This outcome eventually showcases the effective performance of the proposed study model towards handling diverse image types and challenging text layouts. The COCR has exhibited poor performance because it is highly dependent on rule-based image processing and doesn't possess any adaptability towards distorted input or scene text. This issue is overcome in the CTIR approach that learns the visual features, but it still reports struggling towards curved baselines, irregular fonts, and complex text structures. Further, the TTRS model offers stronger recognition performance but is restricted by sub-optimal text detection components. The novelty of the outcome of the proposed system resides in its integration of complementary AI models to carry out detection (EAST and CRAFT) and recognition (CRNN and TrOCR), along with its usage of domain-adaptive and diverse training data. Hence, a robust generalization over image types has been noted for different types of images. In terms of response time outcomes, the proposed scheme accomplishes a response time of 0.82 seconds per image, which is significantly less than the COCR model by 53% faster (response time = 1.75 s), less than the CTIR model by 31.7% (response time = 1.20 s), and also less than the TTRS model by 21.9% (response time = 1.05 s). The accomplishment of these enhancements shows that the proposed model maintains equal emphasis for both accuracy and computational efficiency. The conventional CTIR models using CNN are characterized by increased latency, which is mainly due to their lack of model-level optimization, while TTRS models using transformers suffer from overloaded computation while performing the recognition task, mainly due to their sophisticated self-attention methods. The novelty of the proposed model is: i) an optimized method towards distinguishing detection and recognition tasks, and ii) a lightweight combination of AI models with optimal performance, which maintains necessary computational speed without sacrificing the quality of output.

Apart from analysis of responsiveness, the different attributes viz. scalability, memory consumption, and graphics processing unit (GPU) utilization have also been investigated on varying ranges of images. With complete deployment being carried out on an NVIDIA RTX system with a GPU where batch processing is optimized, the observation found around 69% of GPU utilization, while a maximum of 4.2 GB is noted as memory footprint while performing inference operations when subjected to images. The model is noted with scalable performance on varying and diverse forms of input images and video frames. However, there is a slight increase in computational effort when image resolution is increased, and yet stable performance is noted throughout the experiment.

The study has also investigated the scalability of the model under the input condition of continuous video frames as well as large-scale processing of documents. A stable inference performance has been noted when exposed to increasing streams of image frames and large documents. Computational demands slightly increased with high-resolution images and yet the proposed model is noted with highly controlled utilization of memory, while optimized processing has been noted by the combined effort of CNN and transformer. The outcome infers the practical suitability of the proposed study model in a practical deployment scenario.

3.4. Novelty

An integrated and robust deep learning model has been presented by the proposed study towards detection and recognition right from different image-centric embedded text sources. It does so by integrating localization methods using CNN and contextual recognition methods using transformers. Unlike classical methods that usually relies on standalone recognition methods and preprocessing, the proposed system jointly harnesses the potential of EAST and CRAFT towards accurate localization of characters. It also performs recognition of irregular and distorted structures of text using CRNN and TrOCR, respectively. The prime motive of this framework is towards facilitating higher degree of balance among adaptability, computational

efficiency, and recognition accuracy when subjected to a complex form of an image. Therefore, the proposed study model contributes to bridge the gap existing between real-world deployment demands with research area of document analysis using AI towards robust extraction of text.

3.5. Ablation study

Table 3 showcases the ablation study being carried out with 4 configurations for the detection and recognition module. Lower performance is noted for the baseline configuration via EAST and CRNN which is mainly due to issues in handling contextual recognition and irregular boundaries of text. The system witnessed good improvement in recognition accuracy and contextual understanding by adoption of TrOCR, which is a transformer-based model. Finally, the complete combination (4th row in Table 3) showcases the rationale behind the effective performance of the proposed system.

Table 3. Outcome of ablation study

Configuration	Detection module	Recognition module	Accuracy (%)	Precision (%)	Recall (%)	Response time (s)
Config-1	EAST	CRNN	87.5	86.3	86.7	0.75
Config-2	EAST + CRAFT	CRNN	89.5	90.2	87.6	0.80
Config-3	CRAFT	TrOCR	92.1	93.0	91.4	0.85
Config-4(Prop)	EAST + CRAFT	CRNN + TrOCR	94.7	95.8	94.0	0.82

4. CONCLUSION

The research presented here built a deep learning-based text extraction system that outperformed existing and traditional OCR models, handling a variety of image formats. Higher accuracy and shorter processing time were attained by the suggested system by combining the advantages of CRAFT and TrOCR with focused preprocessing methods and a varied training dataset. The outcomes validate the system's capacity to generalize across a variety of text formats and intricate visual settings. Further research in multilingual OCR, low-resource environments, and real-time deployment on mobile or embedded platforms is made possible by this study. The contributions of the proposed study are as follows: a novel AI framework is presented for text extraction that uses detection (using EAST and CRAFT) and recognition (using CRNN and TrOCR), while it accomplishes 94.7% accuracy, which is the highest performance on a real-world dataset (ICDAR 2015) with faster response time exceedingly better than conventional CNN or transformer-centric models. The model has also witnessed a cost-effective and generalized framework that can manage with diverse and large number of image types. The possible limitation of this model could be that it would be a computationally complex task for processing video sequences with higher definition that could demand more optimization. Another possible limitation associated with the model is its challenges towards recognizing much stylish text, occluded text, or extremely curved text. All these limitations can be addressed as part of future work. Future work will also be carried out to explore Grad-CAM and attention visualization as an explainable model to enhance interpretability in the detection and recognition of text in an image.

FUNDING INFORMATION

The authors attest that no outside funding or financial assistance was used in the conduct of this study.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dayananda Kodala	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Jayaram														
Puttegowda Devegowda		✓				✓		✓	✓	✓	✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. There are no financial or personal relationships that could inappropriately influence or bias the content of this work.

DATA AVAILABILITY

The corresponding author, [DKJ], can provide the datasets created and/or examined for this study upon reasonable request.

REFERENCES

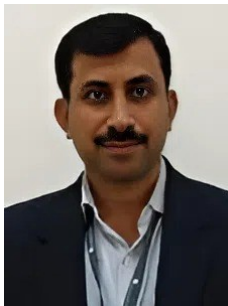
- [1] R. S. Patil, G. Hanji, and R. Huded, "Enhanced scene text recognition using deep learning based hybrid attention recognition network," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 4, Dec. 2024, doi: 10.11591/ijai.v13.i4.pp4927-4938.
- [2] A. Shinde, E. Q. Shahra, S. Basurra, F. Saeed, A. A. AlSewari, and W. A. Jabbar, "SMS scam detection application based on optical character recognition for image data using unsupervised and deep semi-supervised learning," *Sensors*, vol. 24, no. 18, Sep. 2024, doi: 10.3390/s24186084.
- [3] N. Ahmed, A. K. Saha, M. A. Al Noman, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Deep learning-based natural language processing in human-agent interaction: applications, advancements and challenges," *Natural Language Processing Journal*, vol. 9, Dec. 2024, doi: 10.1016/j.nlp.2024.100112.
- [4] K. Okarma and P. Lech, "A method supporting fault-tolerant optical text recognition from video sequences recorded with handheld cameras," *Engineering Applications of Artificial Intelligence*, vol. 123, Aug. 2023, doi: 10.1016/j.engappai.2023.106330.
- [5] F. Shahid, M.-H. Hsu, Y.-C. Chang, and W.-S. Jian, "Using generative AI to extract structured information from free text pathology reports," *Journal of Medical Systems*, vol. 49, no. 1, Mar. 2025, doi: 10.1007/s10916-025-02167-2.
- [6] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *Journal of Big Data*, vol. 11, no. 1, Jul. 2024, doi: 10.1186/s40537-024-00948-z.
- [7] I. Malashin, I. Masich, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Image text extraction and natural language processing of unstructured data from medical reports," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 1361–1377, Jun. 2024, doi: 10.3390/make6020064.
- [8] S. Patil *et al.*, "Enhancing optical character recognition on images with mixed text using semantic segmentation," *Journal of Sensor and Actuator Networks*, vol. 11, no. 4, Oct. 2022, doi: 10.3390/jsan11040063.
- [9] A. Ueda, W. Yang, and K. Sugiura, "Switching text-based image encoders for captioning images with text," *IEEE Access*, vol. 11, pp. 55706–55715, 2023, doi: 10.1109/ACCESS.2023.3282444.
- [10] T. Bisen, M. Javed, P. Nagabhushan, and O. Watanabe, "Segmentation-less extraction of text and non-text regions from JPEG 2000 compressed document images through partial and intelligent decompression," *IEEE Access*, vol. 11, pp. 20673–20687, 2023, doi: 10.1109/ACCESS.2023.3249961.
- [11] H. J. Kim and H. G. Kim, "Leveraging text signed distance function map for boundary-aware guidance in scene text segmentation," *IEEE Access*, vol. 13, pp. 78850–78863, 2025, doi: 10.1109/ACCESS.2025.3566045.
- [12] B. Huang and X. Feng, "Scene text detection based on multi-headed self-attention using shifted windows," *Applied Sciences*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/app13063928.
- [13] S. Tang, Y. Cao, S. Liang, Z. Jin, and K. Lai, "Scene text recognition that eliminates background and character noise interference," *Applied Sciences*, vol. 15, no. 7, Mar. 2025, doi: 10.3390/app15073545.
- [14] A. A. Chandio, M. Asikuzzaman, M. R. Pickering, and M. Leghari, "Cursive text recognition in natural scene images using deep convolutional recurrent neural network," *IEEE Access*, vol. 10, pp. 10062–10078, 2022, doi: 10.1109/ACCESS.2022.3144844.
- [15] J. Park, D. Jeong, H. Lee, S. Han, and J. Paik, "Prompt-based learning for image variation using single image multi-scale diffusion models," *IEEE Access*, vol. 12, pp. 158810–158823, 2024, doi: 10.1109/ACCESS.2024.3487215.
- [16] H. Wen, S. You, and Y. Fu, "LucIE: language-guided local image editing for fashion images," *Computational Visual Media*, vol. 11, no. 1, pp. 179–194, Feb. 2025, doi: 10.26599/CVM.2025.9450310.
- [17] H. Zhao, W. Li, Z. Chang, and C. Yang, "Diff-KT: text-driven image editing by knowledge enhancement and mask transformer," *IEEE Access*, vol. 12, pp. 112948–112965, 2024, doi: 10.1109/ACCESS.2024.3442296.
- [18] A. He and M. Abisado, "Text sentiment analysis of douban film short comments based on BERT-CNN-BiLSTM-Att model," *IEEE Access*, vol. 12, pp. 45229–45237, 2024, doi: 10.1109/ACCESS.2024.3381515.
- [19] K. M. Hasib *et al.*, "MCNN-LSTM: combining CNN and LSTM to classify multi-class text in imbalanced news data," *IEEE Access*, vol. 11, pp. 93048–93063, 2023, doi: 10.1109/ACCESS.2023.3309697.
- [20] Z. Yang and F. Emmert-Streib, "Threshold-learned CNN for multi-label text classification of electronic health records," *IEEE Access*, vol. 11, pp. 93402–93419, 2023, doi: 10.1109/ACCESS.2023.3309157.
- [21] T. Geng, "Transforming scene text detection and recognition: a multi-scale end-to-end approach with transformer framework," *IEEE Access*, vol. 12, pp. 40582–40596, 2024, doi: 10.1109/ACCESS.2024.3375497.
- [22] Z. Liu, H. Wang, C. Zhang, and S. Zhang, "A text summarization approach to enhance global and local information awareness of transformer," *IEEE Access*, vol. 12, pp. 118744–118755, 2024, doi: 10.1109/ACCESS.2024.3449280.
- [23] H. Takahashi and S. Kuriyama, "Text2Layout: layout generation from text representation using transformer," *IEEE Access*, vol. 12, pp. 136319–136328, 2024, doi: 10.1109/ACCESS.2024.3452957.
- [24] P. Kiatphaisansophon, D. Wanvarie, and N. Cooharajanone, "Efficient text bounding box identification using mask R-CNN: case of Thai documents," *IEEE Access*, vol. 12, pp. 49306–49328, 2024, doi: 10.1109/ACCESS.2024.3383911.
- [25] A. Droby, B. K. Barakat, R. Alaasam, B. Madi, I. Rabaev, and J. El-Sana, "Text line extraction in historical documents using mask R-CNN," *Signals*, vol. 3, no. 3, pp. 535–549, Aug. 2022, doi: 10.3390/signals3030032.
- [26] G. Yin, S. Huang, T. He, J. Xie and D. Yang, "Mirrored EAST: an efficient detector for automatic vehicle identification number detection in the wild," in *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 4594–4605, March 2024, doi: 10.1109/TII.2023.3322158.
- [27] Y. Baek, B. Lee, D. Han, S. Yun and H. Lee, "Character region awareness for text detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9357–9366, doi: 10.1109/CVPR.2019.00959.

- [28] A. Mali, N. Sathe, E. Bootwala, A. Khodkumbhe, H. Magar and M. Mali, "A hybrid CRNN and NLP-based PostCorrection framework for optical character recognition," *2026 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2026, pp. 1-6, doi: 10.1109/SCEECS68810.2026.11429872.
- [29] E. S.-Nielsen and I. M.-Herrera, "Transformer-based OCR system for handwritten signature verification in lawmaking petitions: a proof of concept," in *IEEE Access*, vol. 13, pp. 177573-177594, 2025, doi: 10.1109/ACCESS.2025.3620686.
- [30] M. Ibrayim, Y. Li, and A. Hamdulla, "Scene text detection based on two-branch feature extraction," *Sensors*, vol. 22, no. 16, Aug. 2022, doi: 10.3390/s22166262.
- [31] R. Gomez, B. Shi, L. Gomez, L. Numann, A. Veit, J. Matas, "ICDAR2017 robust reading challenge on COCO-text," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1435-1443, doi: 10.1109/ICDAR.2017.234.
- [32] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification – RRC MLT," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1454-1459, doi: 10.1109/ICDAR.2017.237.

BIOGRAPHIES OF AUTHORS



Dayananda Kodala Jayaram    is working as an assistant professor in the Department of Computer Science, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India. He has received his M. Tech. and B.E. from Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India. He is pursuing his Ph.D. from Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India. His teaching and research interests are in the field of data mining, machine learning, and image processing. He has a total teaching experience of 11 years. He can be contacted at email: dayananda.kem@gmail.com.



Dr. Puttegowda Devegowda    is working as a professor and head in the Department of Computer Science and Engineering, ATME College of Engineering, Mysuru, Karnataka, India. He has received his Ph.D. from Mysuru University, Mysuru, Karnataka, India. His teaching and research interests are in the field of data mining, video processing, machine learning, and image processing. He has a total teaching experience of 20 years. He can be contacted at email: puttegowda.77@gmail.com.