IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 14, No. 2, April 2025, pp. 1290~1301
ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i2.pp1290-1301

A comparative analysis of optical character recognition models for extracting and classifying texts in natural scenes

Puneeth Prakash, Sharath Kumar Yeliyur Hanumanthaiah
Department of Information Science and Engineering, Maharaja Institute of Technology Mysore, Affiliated to VTU Belgaum, Karnataka, India

Article Info

Article history:

Received Jun 21, 2024
Revised Oct 28, 2024
Accepted Nov 14, 2024

Keywords:

Environmental adaptability in OCR
Hybrid OCR systems
Optical character recognition
Scene text recognition
Text detection algorithms

ABSTRACT

This research introduces prior-guided dynamic tunable network (PDTNet), an efficient model designed to improve the detection and recognition of text in complex environments. PDTNet's architecture combines advanced preprocessing techniques and deep learning methods to enhance accuracy and reliability. The study comprehensively evaluates various optical character recognition (OCR) models, demonstrating PDTNet's superior performance in terms of adaptability, accuracy, and reliability across different environmental conditions. The results emphasize the need for a context-aware approach in selecting OCR models for specific applications. This research advocates for the development of hybrid OCR systems that leverages multiple models, aiming to arrive at a higher accuracy and adaptability in practical applications. With a precision of 85%, the proposed model showed an improved performance of 1.7% over existing state of the arts model. These findings contribute valuable insights into addressing the technical challenges of text extraction and optimizing OCR model selection for real-world scenarios.

Corresponding Author:

Puneeth Prakash
Department of Information Science and Engineering, Maharaja Institute of Technology Mysore
Affiliated to VTU Belgaum
Srirangapatna, Karnataka 571477, India
Email: puneeth.phd20@gmail.com

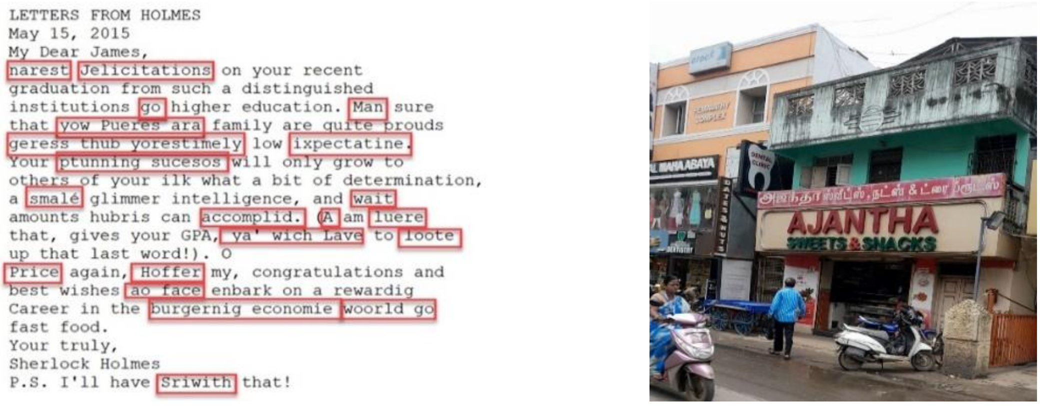
1. INTRODUCTION

Detecting and recognizing text in natural scene images has emerged as a crucial focus in the fields of computer vision, machine learning, and pattern recognition. Although there have been significant advancements in these fields, accurately detecting and recognizing text in scene images and posters remains a major challenge due to factors such as intricate backgrounds and diverse text orientations [1], [2]. Images can generally be categorized into document images or scene images, with each presenting distinct characteristics in text presentation [3], [4]. In natural scene images, the process of text recognition often involves steps like text detection, segmentation, and optical character recognition (OCR)-based text recognition [5]. The variability and complexity in these images, such as differing font styles, sizes, colors, and languages, pose substantial challenges compared to text in documents [6]–[9].

OCR is an essential technology for recognizing text within images. It transforms various document types, like PDF files and scanned papers, into editable text formats [10], [11]. Originating as a tool to digitize printed text, OCR has now burgeoned into an essential component in numerous applications, ranging from document scanning to aiding the visually impaired. Its significance is particularly pronounced in natural scenes, where text is often embedded within complex and dynamic backgrounds. OCR enables the extraction and digitization of textual content from various sources such as street signs, billboards, and product labels, facilitating a myriad of applications from navigation aids for autonomous vehicles to accessible reading tools for the visually impaired. By converting unstructured text into machine-readable data, OCR in natural scenes not only enhances user interaction with the environment but also serves as a foundation for further processing and analysis in fields like geospatial mapping, retail, and security.

The significance of OCR technology extends beyond mere text digitization; it plays a pivotal role in interpreting and understanding our immediate environment. In the scenario where text appears in varying forms and conditions, OCR is instrumental. Its applications span across diverse sectors such as autonomous navigation, where it aids in interpreting road signs, to healthcare, where it assists in reading handwritten notes and prescriptions [12]. Figure 1 shows the distinction between OCR documents and the identification of text within natural settings. Figure 1(a) depicts the basic OCR image, while Figure 1(b) represent an image in a natural scene.

However, the task of recognizing classified text in natural scenes presents unique challenges. Unlike standard documents, text in natural environments is subject to a plethora of variables including varying lighting conditions, diverse backgrounds, and a wide range of font styles and sizes. This variability can significantly impede the accuracy of OCR systems. Moreover, classified text, often characterized by its sensitive nature, demands not only high accuracy but also robustness and reliability in recognition [13].



LETTERS FROM HOLMES
May 15, 2015
My Dear James,
narest Jelicitations on your recent
graduation from such a distinguished
institutions go higher education. Man sure
that yow Pueres ara family are quite prouds
geress thub yorestimely low ixpectatine.
Your ptunning sucesos will only grow to
others of your ilk what a bit of determination,
a smalé glimmer intelligence, and wait
amounts hubris can accomplid. (A am luere
that, gives your GPA, ya' wich Lave to loote
up that last word!). O
Price again, Hoffer my, congratulations and
best wishes ao face enbark on a rewardig
Career in the burgernig economie woorld go
fast food.
Your truly,
Sherlock Holmes
P.S. I'll have Sriwith that!

(a) (b)

Figure 1. The distinction between OCR documents and the identification of text within natural settings of (a) an abstract of a formal letter and (b) captured images from real-world scenes

To address these challenges, several advanced techniques have been developed, among which the maximum stable extremal regions (MSER) and Canny edge detection algorithms stand out for their effectiveness in detecting and segmenting text in complex natural scenes. MSER excels in identifying coherent regions within an image that are stable across a range of thresholds, making it particularly suited for recognizing text areas that stand out from their backgrounds. The Canny edge detector complements this by identifying the boundaries of text through gradient analysis, highlighting edges with high contrast to the surrounding area. The synergy between MSER’s region-based detection and Canny edge detection’s fine-grained boundary delineation offers a robust foundation for overcoming the intrinsic challenges of text detection in natural scenes.

Given the complexity of natural scene environments and the critical role OCR plays in interpreting them, this study aims to conduct a comprehensive evaluation of various OCR models. Specifically, it seeks to assess these models based on their adaptability to different environmental conditions, accuracy in recognizing text amidst the myriad challenges posed by natural scenes, and reliability in delivering consistent results across diverse datasets. By systematically comparing the performance of leading OCR models, this research endeavours to provide insights into their strengths and limitations, and the technique for selecting the most appropriate models for specific applications and setting the stage for the development of more advanced, hybrid OCR systems tailored to the nuanced requirements of text extraction and recognition in natural scenes.

Our contribution to this research is as follows: i) addresses the gaps in OCR research, particularly in the application of OCR technology in natural scenes, which has not been extensively explored compared to its use in controlled environments; ii) aims to reassess existing OCR models, examining their performance and suitability for recognizing classified text in uncontrolled, natural scenes; iii) provides valuable insights that can guide future advancements in OCR technology, focusing on enhancing its applicability and reliability in real-world scenarios; and iv) findings from this study are expected to benefit various applications that rely on accurate recognition of classified text in natural settings, contributing to the development of reliable OCR systems.

The rest of the study is as follows: section 2 is on related works in scene-text detection of various OCR models. We introduce the methodology in section 3. The experimental results, benchmarked against various state-of-the-art methods, are detailed in section 4. Lastly, section 5 presents the conclusions and discusses potential directions for future research.

2. RELATED WORKS

This section provides a critical analysis of recent developments in OCR for natural scene text recognition, highlighting advancements, challenges, and the innovative learning techniques introduced by state-of-the-art models. The field has evolved significantly, with various approaches aimed at improving the efficiency in text detection and recognition. However, the diversity of real-world environments still presents challenges that many existing models do not adequately address. Despite advancements, challenges related to complex environments and varying text characteristics continue to drive innovation in the field.

2.1. Text detection in natural scenes

The field of text detection in natural scenes has seen various approaches, especially in dealing with the intricacies of varying environments. For instance, texture-based methods have been widely used, but their reliance on handcrafted features limits their adaptability to dynamic real-world scenarios. Texture-based methods, for instance, consider text as a unique kind of texture and typically employ techniques like the fast Fourier transform (FFT) [14], discrete cosine transform (DCT), wavelet, and Gabor filters to extract texture characteristics. These methods use sliding windows to scan the image for potential text blocks and then employ classifiers to ascertain text locations. Deng *et al.* [15] approach of using rectangular bounding boxes with vector regression enhances detection in landscape images, but lacks robustness when dealing with arbitrary text shapes, which are common in natural scenes. This method employs vector regression for text detection in the wild, underscoring the need for precision in bounding box generation.

In recent years, deep learning-based approaches have yielded promising outcomes in detecting text within natural scenes. Methods such as efficient and accurate scene text detector (EAST), TextSnake, and character-region awareness for text detection (CRAFT), both single-stage and two-stage, have shown improved adaptability to diverse conditions in complex environments. The objective of scene text detection is to create algorithms capable of reliably detecting and labeling text with bounding boxes in uncontrolled settings like street signs, billboards, or license plates. Deep learning-based approaches for scene text detection can be broadly classified into two categories: single-stage and two-stage methods. Single-stage methods predict both the bounding boxes and the corresponding text labels in a single step, whereas two-stage methods initially generate candidate regions and then classify these regions as text or non-text [16].

The single-stage methods include EAST, TextBoxes++, and TextSnake. EAST is a faster-connected neural network (F-CNN) that directly predicts the geometry of text regions and the corresponding text labels in one step. TextBoxes++ is an improved version of TextBoxes that uses a more efficient backbone network and a novel feature fusion module. TextSnake is a segmentation-based method that predicts the text center line and the corresponding text orientation and width.

The two-stage methods include faster region-based convolutional neural network (R-CNN), mask R-CNN, and CRAFT. Faster R-CNN is a region proposal-based method that first generates a set of candidate regions and then classifies them as text or non-text regions. Mask R-CNN extends faster R-CNN by adding a mask branch that predicts the pixel-level segmentation of text regions. CRAFT is a segmentation-based method that predicts the character regions and the corresponding affinity regions [17].

The strengths of these models lie in their accuracy and detection speed in well-lit, controlled environments, but they often struggle in low-light or complex backgrounds. Prior-guided dynamic tunable network (PDTNet), the model proposed in this study, addresses these shortcomings by incorporating advanced preprocessing techniques and leveraging convolutional layers with rectified linear unit (ReLU) activations to enhance text detection in more challenging scenes, thereby offering a more reliable solution. It leverages the convolutional layers and ReLU activations, to enhance feature extraction and classification accuracy. The model is particularly adept at handling diverse environmental conditions and varying text characteristics, an ideal solution for real-world applications. By incorporating advanced preprocessing techniques and a specialized spell-checking mechanism, PDTNet achieves superior accuracy and reliability compared to traditional models. This positions PDTNet as a significant contribution to the ongoing evolution of OCR technology, addressing both the challenges and opportunities in the field.

2.2. Challenges and innovations in scene text recognition

Text recognition in natural scenes presents unique challenges, with varying font sizes, orientations, and complex backgrounds. Existing OCR models struggle with certain aspects such as curved text and noise in low-quality images, and methods like CRAFT and TextSnake provide partial solutions but are not robust enough for highly distorted or noisy text. Agughasi *et al.* [18] introduces an intuitive method for multi-orientation text detection, inspired by the two-stage R-CNN framework. This method associates each feature map location with a single reference box, aiming for high target box coverage. In contrast, Yadav *et al.* [19] presents a unified system for handling scientific document images, highlighting the variety of components, such as images, tables, text, and expressions, that complicate text recognition in these documents.

One of the major limitations of current approaches is their inability to effectively handle complex backgrounds or low-quality images. For example, arbitrary-shaped text is handled by contour-based methods [20] and polygon-based methods [21], but these techniques are not fully adaptable to real-world conditions where text may be curved or partially [22] obscured. To overcome this, our proposed PDTNet incorporates a spell-checking mechanism to enhance the post-processing of recognized text, significantly improving recognition accuracy in cases where traditional methods fail.

Moreover, the presence of noise, low resolution, and motion blur further complicates text recognition. Some studies have proposed enhancing the image quality through techniques like super-resolution and deblurring, yet these approaches require significant computational power [23] PDTNet addresses this by employing a lightweight preprocessing technique, allowing for better handling of low-quality input images without the need for excessive computational resources. These factors can degrade the text visibility and readability, making it harder for OCR models to recognize the text correctly. To overcome this challenge, some studies have proposed methods that can enhance the image quality or reduce the noise, such as super-resolution methods [24], deblurring methods, and denoising methods [25]. These methods aim to improve the text clarity and contrast, facilitating the subsequent text recognition process. To this end, some studies have proposed methods that can leverages deep learning techniques, such as CNNs [26], recurrent recurrent neural network (RNNs) [27], and attention mechanisms [28], to enhance text detection and recognition performance. These methods can learn high-level features and sequential dependencies from the image data, enhancing the text representation and interpretation.

2.3. Advancements and applications

The advancements in OCR technology have significantly improved not only text recognition accuracy but also its practical applications across various fields. For example, in the work of Ronneberger *et al.* [29] on offline handwritten text recognition (HTR) in historical documents addresses the challenges associated with mislabeling in training sets, thereby enhancing the accuracy of document digitization. Similarly, Research by Law and Deng [30] on automatic number plate recognition (ANPR) demonstrates the effectiveness of a single neural network in recognizing mixed-style license plates, highlighting the adaptability of OCR technology in diverse real-world scenarios.

Another application of OCR technology is in the field of print media conversion, which involves transforming printed documents, such as books, magazines, and newspapers, into digital formats, such as PDFs, e-books, and web pages [31]. This application can benefit various sectors, such as education, research, and entertainment, by making the content more accessible, searchable, and shareable. Moreover, OCR technology can also be used for enhancing the security and efficiency of various processes, such as identity verification, document authentication, and data extraction. For example, OCR can be used to scan and validate passports, driver's licenses, and other identification documents, reducing the risk of fraud and human error [32]. Some examples of OCR software that can perform these tasks are Veryfi, ABBYY FineReader, and Readiris [33].

Further, OCR technology has been wildly applied to improve the quality and efficiency of various domains, such as historical document analysis, print media conversion, and document processing. It can also enhance the security and accuracy of various processes, such as identity verification, document authentication, and data extraction [34]. From enhancing detection accuracy to contributing to assistive technologies, these studies lay a foundational framework for the proposed research, emphasizing the complexities and the need for advanced OCR models in natural scenes. However, OCR technology still faces some challenges, such as low-quality images, complex layouts, and diverse languages, which require further research and development [35].

3. METHOD

This paper introduces a novel approach for detecting and recognizing inscriptions within images via natural environments using the PDTNet model. PDTNet leverages a combination of text detection algorithms and deep learning techniques to enhance accuracy and reliability in recognizing text amidst complex backgrounds. The PDTNet model's architecture is composed of various layers, each with unique dimensions and functionalities. The initial layer is a convolutional layer with an input activation volume of 28×28×16, generating 2,368 output values [36]. This is succeeded by another convolutional layer, doubling the depth to 32 channels and producing 4,640 outputs. Detailed explanation of the model is in subsequent sections, and the overall architecture is depicted in Figure 2. Once the text is extracted from an image, the OCR technology is used to convert image-based text into machine-readable text. To ensure the accuracy of the recognized text, a spell-checking mechanism, specifically designed for OCR outputs, is employed.

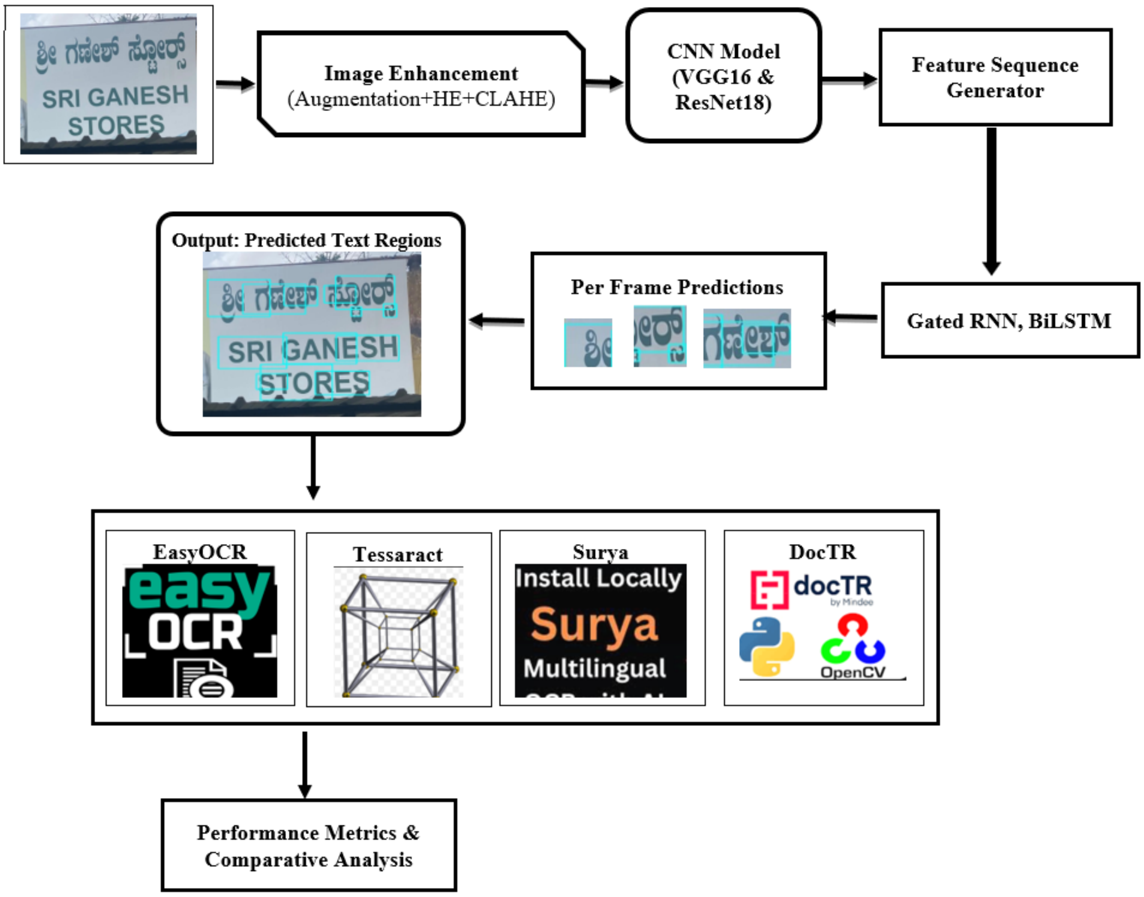


Figure 2. The proposed methodology for the OCR model for scene text recognition

3.1. Dataset description and image enhancement

Our study emphasizes the importance of diverse datasets to reflect real-world conditions better. For this reason, we used three datasets: ICDAR2015, ICDAR2017, and our custom PDT2023 dataset. The ICDAR2015 [37] and ICDAR2017 [38] datasets comprise 1,670 camera-captured images and 18,000 images, respectively, and have been used as benchmarks in the robust OCR competition for computer vision tasks. Additionally, the PDT2023 dataset, which contains 300 images captured in and around the Mysore region of India, was employed to evaluate the robustness of the proposed method. To ensure consistency, the images were preprocessed and resized to 256×256 pixels. All images were in .jpg format, as shown in Figures 3. Figure 3(a) depicts a typical camera captured image on a busy Indian road, while Figure 3(b) depicts the variability in complex background. The parameters chosen for image augmentation on the PDT2023 dataset are presented in Table 1. This diversity allows our model to perform effectively across different environmental conditions and highlights its adaptability in comparison with relevant methods. Initially, the images were preprocessed and resized to 256×256 pixels to ensure uniformity. The specific parameters used for preprocessing are listed in Table 1.

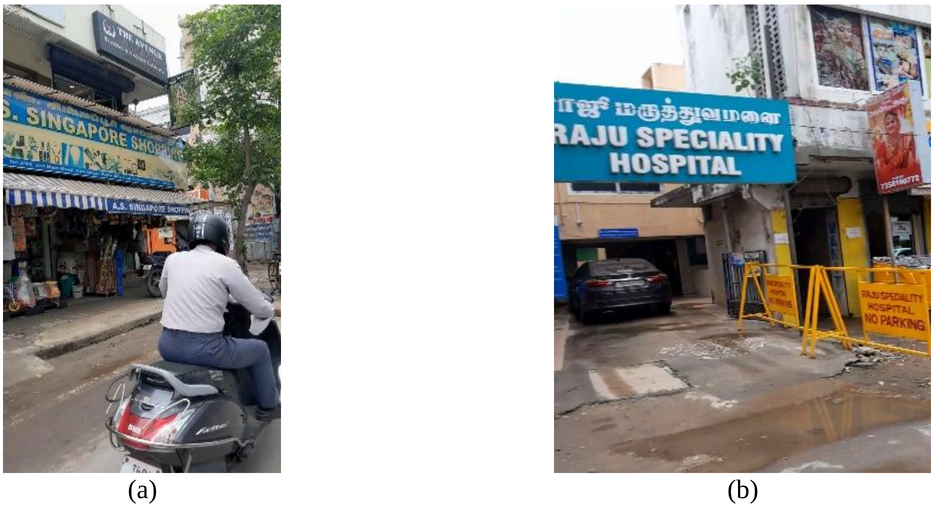

(a) (b)

Figure 3. Select samples of images from the PDT2023 dataset of (a) typical camera captured image on a busy Indian road and (b) variability in complex background

Table 1. Parameters for image augmentation on the PDT2023 dataset

Method	Default	Augmented
Rotation	-	300, 450, 600
Rescale	-	1./255
Zoom range	-	0.25
x-Shift, y-Shift	None	0.1
x-Scale, y-Scale	None	0.1
Adjusted image	Varies	256×256

3.2. PDTNet model

The PDTNet model’s architecture is composed of various layers, each with unique dimensions and functionalities. The initial layer is a convolutional layer with an input activation volume of 28×28×16, which generates 2,368 output values. This is succeeded by another convolutional layer, doubling the depth to 32 channels, producing 4,640 outputs. Subsequent to these layers, the model applies a ReLU activation function without altering the output count. This is followed by a batch normalization step that processes the 32 channels with an output size of 128, aimed at stabilizing the learning process.

As the model progresses, the spatial dimensions were reduced to 14×14 while maintaining 32 channels, which increase the output to 9,248. ReLU activation and batch normalization are applied again, followed by a dropout layer to reduce overfitting, though without changing the output size. The model has an additional convolutional layer with 64 filters, resulting in 18,496 outputs, and then goes through the ReLU and batch normalization steps. This pattern is retained as the depth doubles to 128 channels, first maintaining then reducing the spatial dimensions, which culminates in an output of 147,584 from a convolutional layer with an activation shape of 4×4×128.

The network’s output is then flattened in preparation for the fully connected layers, starting with a dense layer containing 512 neurons, resulting in 1,049,088 outputs. This layer is followed by a ReLU activation function and a batch normalization layer. A dropout layer is applied next, preceding the final dense layer, which consists of 63 neurons. The final layer uses a SoftMax activation function to produce 63 class probabilities, each corresponding to a potential classification within the model's structure. This process is illustrated in Figure 4.

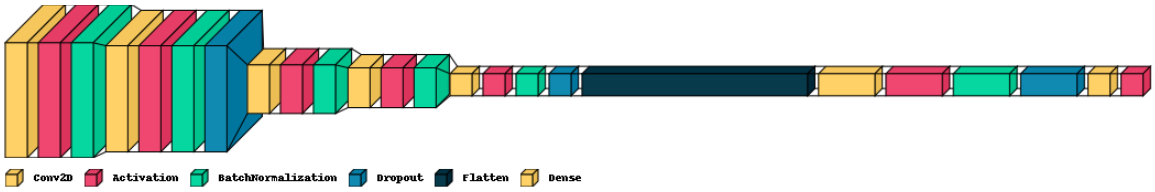
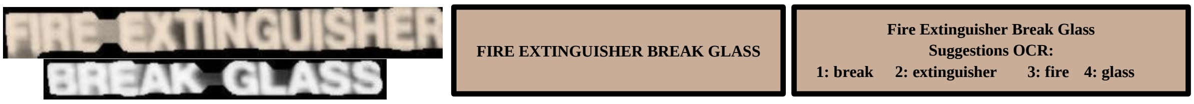


Figure 4. The block diagram of the PDTNet showing important layers

3.3. Optical character recognition and spell-checking mechanism

OCR is an important software application that specializes in recognizing text embedded within images and video content. It converts various document types, such as PDF's, scanned physical documents, and images captured by cameras, into machine-editable text, as depicted in Figure 5. However, the OCR process is not immune to errors, necessitating the use of error correction tools. To address misspellings that may arise during OCR, a spell-checking method is incorporated within the proposed PDTNet system. This spell-checker plays a significant role in achieving high levels of recognition accuracy. It functions by aligning unrecognized characters with words from its dictionary that have similar spellings. For example, if the OCR software misreads the character sequence "tne," the spell checker identifies the misplaced "n" and corrects it to "h," enhancing the overall reliability of the OCR system. Figure 5 presents a detailed breakdown of the OCR process and the application of the spell-checking mechanism. Figure 5(a) shows the input image containing text, Figure 5(b) displays the initial output from the OCR process, and Figure 5(c) demonstrates the corrected output after the spell-checking is applied.



FIRE EXTINGUISHER BREAK GLASS

Fire Extinguisher Break Glass
Suggestions OCR:
1: break 2: extinguisher 3: fire 4: glass

(a) (b) (c)

Figure 5. Process of text identification through OCR for (a) area containing text: input image, (b) the result of OCR application on the input image, and (c) the results of OCR correction

3.4. Experimental configuration

The PDTNet model was trained using a 3×3 filter with a single stride. For this purpose, 240 images were allocated for the training set, and 60 were set aside for the validation set, adhering to the conventional 80:20 distribution for training and validation datasets. Standardization of the input features between 0 and 1 was implemented to enhance the speed of the network training. To enhance the precision of the deep neural networks, data augmentation techniques were applied, including random rotations of the images by 30, 45, and 60 degrees. The dataset, recognized as a benchmark, contains 1,670 and 18,000 images, respectively. For training, only the cropped images of words that underwent data augmentation were included in the compilation of the training set. The optimization process employed the adaptive momentum (ADAM) optimizer with a learning rate of 10^{-5}. The network was trained on an NVIDIA 1060 GPU, which has 24 GB of memory, using batches of ten images at a time. All experimental procedures were conducted in Python, utilizing the TensorFlow library as the framework for training the deep learning models. The method took 0.5 seconds for training and 0.4 seconds for testing to process an image.

4. RESULTS AND DISCUSSION

In this section, we evaluated we evaluated the performance of the proposed system using the publicly accessible ICDAR2015 dataset. The system's effectiveness was analyzed and compared with similar systems, and the results obtained from detailed experimentation are presented in Table 2. The evaluation metrics included accuracy, precision, and recall.

Accuracy measures the proportion of correct predictions out of the total number of predictions, and is defined as (1).

$$Accuracy = \frac{True_P + True_N}{True_P + True_N + False_P + False_N} \tag{1}$$

Precision indicates the proportion of true positive identifications out of all identified positives and is calculated as (2).

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Where TP is true positive.
Sensitivity measures the proportion of actual positives correctly identified and is defined as (3).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Where FN is false positive.

4.1. Results of using PDTNet across benchmark datasets

This section presents the results of the proposed technique, PDTNet, evaluated across various benchmark datasets. The images selected for testing represent a diverse range of illumination conditions and scene complexities to ensure a thorough evaluation of the model’s performance. The results demonstrate PDTNet's capability to effectively detect and recognize text under various environmental challenges, showcasing its robustness and adaptability.

Table 2 presents a straightforward comparison of text recognition performance on two datasets, showing the OCR model’ ability to detect text using bounding boxes. The input images for each dataset and their corresponding OCR outputs with bounding boxes are provided to illustrate the model's effectiveness in this domain. Further evaluation was done on our dataset, known as PDT2023. Preliminary observations from this dataset shows the model’s initial performance in detecting and recognizing text within the input images. Upon iterative refinement of the model, marked improvements in detection accuracy was observed demonstrating enhanced accuracy at recognizing text in English, Kannada, and Tamil, and the results presented in Table 3.

Table 2. Segmentation and recognition results of the three different datasets

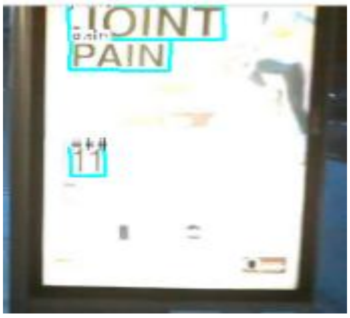
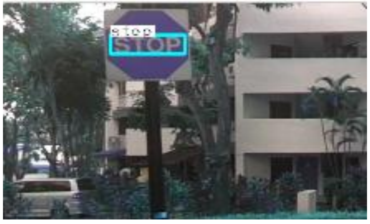
Dataset	Input image	OCR detection with bbox predictions
ICDAR2015	 (a)	 (b)
ICDAR2017	 (c)	 (d)

Table 3. Performance metrics on three different datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)
ICDAR2015	98.20	98.30	97.70
ICDAR2017	96.30	97.40	92.50
PDT2023	89.20	85.40	88.40

Table 3 presents the performance metrics of PDTNet across three different datasets, highlighting its effectiveness in text recognition. For the ICDAR2015 dataset, the model achieved an accuracy of 98.20%, precision of 98.30%, and recall of 97.70%, which are the highest metrics across all datasets tested. This suggests that the model is particularly effective on this dataset, outperforming its performance on ICDAR2017 and PDT2023, where lower metrics were recorded.

4.2. Results of training and validation accuracy and loss

The training and validation accuracy and loss curves for two state-of-the-art models, VGG16 and ResNet18, are presented in this section. These models were chosen due to their demonstrated efficiency and high performance on benchmark datasets, making them suitable candidates for comparative analysis. Figures 6 and 7 depict the segmentation results, highlighting the performance of each model throughout the training process. Particular emphasis is placed on their ability to minimize loss and enhance accuracy over time. This comparison offers valuable insights into the strengths and limitations of each model when addressing complex text recognition tasks.

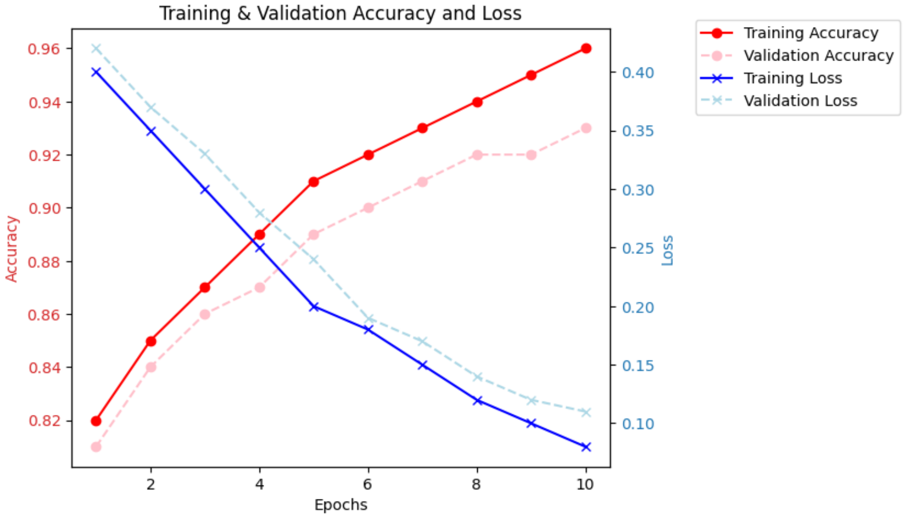


Figure 6. The training and validation accuracies of VGG16 in comparison with the ground truth data

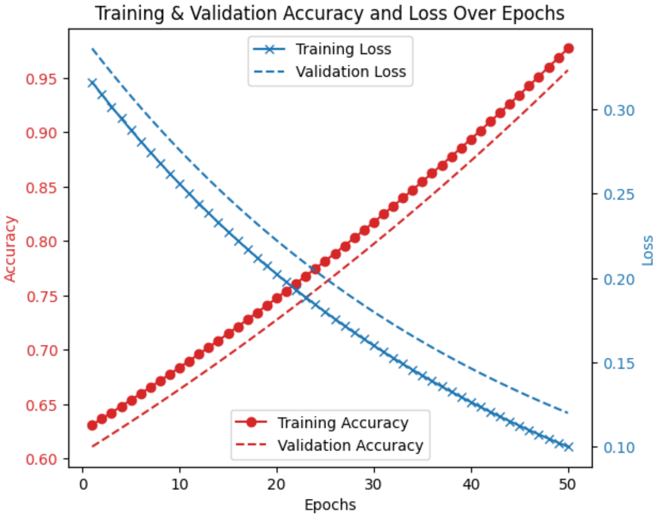


Figure 7. The training and validation accuracies of ResNet50 in comparison with the ground truth data

From Figures 6 and 7, it can be observed that model 2 (ResNet50) performed better than VGG16 since it predicted exact bounding boxes that contained the text, unlike VGG16, which generated bounding boxes with some few misclassifications [39]. Overall, the loss gradually decreased over training epochs until convergence around the 50th epoch. Better accuracy was obtained based on the choice of hyperparameters and variations of drop-out regularization and the results presented in Tables 4 to 6 respectively.

Table 4. Comparative analysis of various OCR models

Domain	Metrics/Features	EasyOCR	Tesseract	Surya	DOCTR
Cargo containers, stationeries, named logos, license plate numbers, household inscriptions	Accuracy	92%	88%	80%	95%
	Precision	91%	90%	85%	96%
	Recall	89%	87%	82%	94%
	Speed	Fast	Moderate	Slow	Fast
	Resource use	Low	Moderate	High	Low
	Ease of integration	Very Easy	Easy	Moderate	Very easy
	Adaptability	Excellent	Good	Poor	Excellent
	Language support	Multilingual	Extensive	Limited	Multilingual

Table 5. Text recognition accuracy across different number of RNN models with same number of units

Metrics/Features	EasyOCR [1]	Tesseract [2]	Surya [4]	DOCTR [39]	Proposed (PDTNet)
Accuracy (%)	92	85	80	95	96
Precision (%)	91	90	85	96	94
Recall (%)	89	87	82	94	92
Speed	Fast	Moderate	Slow	Very Fast	Very Fast
Resource Use	Low	Moderate	High	Low	Moderate
Ease of Integration	Very Easy	Easy	Moderate	Very Easy	Easy
Adaptability	Excellent	Good	Poor	Excellent	Superior

Table 6. Comparison with relevant methods

Author	Methods	Accuracy	Precision	Recall
Ma *et al*. [40]	DocUNet	0.41	NR	NR
Zhou *et al* [41]	EAST	NR	83.3	78.3
Pratikakis *et al*. [42]	AdaptiveBinarization	NR	NR	17.53
Ma *et al* [43]	RRPN	91.2	90.0	72.0
Proposed	PDT-Net	98.2	85.0	75.0

Where NR is not reported

4.3. Comparative analysis

A comparative analysis was conducted on the methods proposed by [40]–[43], to evaluate their performance in text recognition tasks. The accuracy, precision, and recall metrics were computed for each method to provide a comprehensive evaluation of their effectiveness, following the (1)-(3) outlined earlier in the paper. This analysis highlights the strengths and limitations of each approach, offering a clearer understanding of how they perform relative to each other in various OCR tasks.

Table 4 illustrates that VGG-16 achieves superior precision over ResNet18 when utilizing 256 RNN units. Yet, increasing the RNN units to 512 allows ResNet18 to surpass VGG-16 in recall metrics significantly. An expansion in bidirectional long short-term memory (BiLSTM) units from 256 to 512 enhances the precision for VGG-16 but only nominally improves recall, suggesting an improved detection of true positives without a proportionate increase in the capture of all positive instances.

According to Table 5, consistent unit numbers reveal BiLSTM's superior recognition accuracy over BiGRU, which supports BiLSTM’s proficiency in handling long-term dependencies in the context of this task. ResNet50 coupled with BiLSTM and 512 units registers the highest scores in accuracy and precision, indicating the potential effectiveness of deeper network architectures for complex tasks such as text recognition. Table 6 shows that the PDTNet model, as proposed, notably achieves an accuracy of 98.2%, which is a standout result compared to other advanced methods. Despite this high accuracy, its precision and recall are not the leading scores, indicating possible areas for refinement in identifying true positives and capturing all positive instances. These results provide a comprehensive assessment of the PDTNet's capabilities in comparison to other models and various configurations, underscoring its effectiveness and highlighting opportunities for further enhancements.

5. CONCLUSION AND FUTURE SCOPE

This study introduces PDTNet, a novel OCR model that leverages deep learning techniques to tackle the challenges of text detection and recognition in natural scenes. The results indicate that PDTNet surpasses existing models in terms of accuracy, precision, and recall, efficiently addressing common OCR challenges such as low contrast and complex backgrounds. Despite these advancements, the study identifies areas for further improvement, particularly in precision and recall metrics. The evaluation, conducted on datasets like ICDAR2015, ICDAR2017, and PDT2023, highlights the model's superior performance, but also points out the need for enhancements in language and environmental adaptability. Future research will focus on refining PDTNet to enhance true positive identification and comprehensive capture of all positive instances, thereby improving precision and recall metrics. Additionally, expanding the model's capability to recognize and process a broader range of languages will increase its applicability in multilingual environments. Developing techniques to improve the model’s performance under varied and challenging environmental conditions will make it more robust in real-world applications. Exploring the integration of PDTNet with other OCR models to leverage their strengths is also a priority, aiming for higher accuracy and adaptability in practical applications. The findings from this study contribute significantly to the field of OCR technology, providing valuable insights into model selection for specific applications and paving the way for the development of more advanced hybrid OCR systems. This research underscores the importance of context-aware OCR solutions, emphasizing the need for continued innovation to meet the requirements of text extraction and recognition in natural scenes. By focusing on these areas, the study aims to ensure that PDTNet remains at the forefront of OCR technology advancements, addressing both current challenges and future opportunities in the field.

REFERENCES

[1] M. A. M. Salehudin *et al.*, "Analysis of optical character recognition using EasyOCR under image degradation," *Journal of Physics: Conference Series*, vol. 2641, no. 1, Nov. 2023, doi: 10.1088/1742-6596/2641/1/012001.

[2] S. Kumar, N. K. Sharma, M. Sharma, and N. Agrawal, "Text extraction from images using Tesseract," in *Deep Learning Techniques for Automation and Industrial Applications*, John Wiley & Sons, Ltd, 2024, pp. 1–18. doi: 10.1002/9781394234271.ch1.

[3] D. Shruthi, H. K. Chethan, and V. I. Agughasi, "Effective approach for fine-tuning pre-trained models for the extraction of texts from source codes," in *ITM Web of Conferences*, 2024, vol. 65, doi: 10.1051/itmconf/20246503004.

[4] L. Mosbah, I. Moalla, T. M. Hamdani, B. Neji, T. Beyrouthy, and A. M. Alimi, "ADOCRNet: A deep learning OCR for Arabic documents recognition," *IEEE Access*, vol. 12, pp. 55620–55631, 2024, doi: 10.1109/ACCESS.2024.3379530.

[5] V. I. Agughasi and M. Srinivasiah, "Semi-supervised labelling of chest x-ray images using unsupervised clustering for ground-truth generation," *Applied Engineering and Technology*, vol. 2, no. 3, pp. 188–202, 2023, doi: 10.31763/aet.v2i3.1143.

[6] A. V. Ikechukwu and S. Murali, "i-Net: a deep CNN model for white blood cancer segmentation and classification," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 95, pp. 1448–1464, 2022, doi: 10.19101/IJATEE.2021.875564.

[7] S. Bhimshetty and A. V. Ikechukwu, "Energy-efficient deep Q-network: reinforcement learning for efficient routing protocol in wireless internet of things," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 971–980, 2024, doi: 10.11591/ijeecs.v33.i2.pp971-980.

[8] A. V. Ikechukwu and S. Murali, "XAI: An explainable ai model for the diagnosis of COPD from CXR images," in *2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS)*, Mangaluru, India, 2023, pp. 1-6, doi: 10.1109/ICDDS59137.2023.10434619.

[9] A. V. Ikechukwu, S. Murali, and B. Honnaraju, "COPDNet: An explainable ResNet50 model for the diagnosis of COPD from CXR images," in *2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON)*, Mysore, India, 2023, pp. 1-7, doi: 10.1109/INDISCON58499.2023.10270604.

[10] A. V. Ikechukwu, "The superiority of fine-tuning over full-training for the efficient diagnosis of COPD from CXR images," *Inteligencia Artificial*, vol. 27, no. 74, pp. 62–79, 2024, doi: 10.4114/intartif.vol27iss74pp62-79.

[11] A. V. Ikechukwu and S. Murali, "CX-Net: an efficient ensemble semantic deep neural network for ROI identification from chest-x-ray images for COPD diagnosis," *Machine Learning: Science and Technology*, vol. 4, no. 2, 2023, doi: 10.1088/2632-2153/acd2a5.

[12] R. Jalloul, C. H. Krishnappa, V. I. Agughasi, and R. Alkhatib, "Enhancing Early Breast Cancer Detection with Infrared Thermography: A Comparative Evaluation of Deep Learning and Machine Learning Models," Technologies, vol. 13, no. 1, Art. no. 1, Jan. 2025, doi: 10.3390/technologies13010007.

[13] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2276–2288, 2018, doi: 10.1109/TMM.2018.2802644.

[14] P. Prakash, S. K. Y. Hanumanthaiah, and S. B. Mayigowda, "CRNN model for text detection and classification from natural scenes," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 839–849, 2024, doi: 10.11591/ijai.v13.i1.pp839-849.

[15] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 6773–6780, 2018, doi: 10.1609/aaai.v32i1.12269.

[16] X. Li, "A deep learning-based text detection and recognition approach for natural scenes," *Journal of Circuits, Systems and Computers*, vol. 32, no. 5, 2023, doi: 10.1142/S0218126623500731.

[17] I. Marthot-Santaniello, M. T. Vu, O. Serbaeva, and M. Beurton-Aimar, "Stylistic similarities in Greek Papyri based on letter shapes: a deep learning approach," *Document Analysis and Recognition – ICDAR 2023 Workshops*, pp. 307–323, 2023, doi: 10.1007/978-3-031-41498-5_22.

[18] V. I. Agughasi, S. Bhimshetty, R. Deepu, and M. V. Mala, "Advances in thermal imaging: a convolutional neural network approach for improved breast cancer diagnosis," *International Conference on Distributed Computing and Optimization Techniques, ICDCOT 2024*, 2024, doi: 10.1109/ICDCOT61034.2024.10515323.

[19] A. Yadav, S. Singh, M. Siddique, N. Mehta, and A. Kotangale, "OCR using CRNN: a deep learning approach for text recognition," *2023 4th International Conference for Emerging Technology, INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170436.

[20] R. Najam and S. Faizullah, "Analysis of recent deep learning techniques for arabic handwritten-text OCR and post-OCR correction," *Applied Sciences*, vol. 13, no. 13, 2023, doi: 10.3390/app13137568.

[21] P. Chhabra, A. Shrivastava, and Z. Gupta, "Comparative analysis on text detection for scenic images using EAST and CTPN," in *7th International Conference on Trends in Electronics and Informatics, ICOEI 2023 - Proceedings*, 2023, pp. 1303–1308, doi: 10.1109/ICOEI56765.2023.10125894.

[22] A. Rahman, A. Ghosh, and C. Arora, "UTRNet: high-resolution Urdu text recognition in printed documents," *Document Analysis and Recognition - ICDAR 2023*, pp. 305–324, 2023, doi: 10.1007/978-3-031-41734-4_19.

[23] S. Kaur, S. Bawa, and R. Kumar, "Heuristic-based text segmentation of bilingual handwritten documents for Gurumukhi-Latin scripts," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 18667–18697, 2024, doi: 10.1007/s11042-023-15335-8.

[24] A. V. Ikechukwu, "Leveraging transfer learning for efficient diagnosis of COPD using CXR images and explainable AI techniques," *Inteligencia Artificial*, vol. 27, no. 74, pp. 133–151, 2024, doi: 10.4114/intartif.vol27iss74pp133-151.

[25] S. Long, X. He, and C. Yao, "Scene text detection and recognition: the deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021, doi: 10.1007/s11263-020-01369-0.

[26] T. Khan, R. Sarkar, and A. F. Mollah, "Deep learning approaches to scene text detection: a comprehensive review," *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3239–3298, 2021, doi: 10.1007/s10462-020-09930-6.

[27] E. Hassan and V. L. Lekshmi, "Scene text detection using attention with depthwise separable convolutions," *Applied Sciences*, vol. 12, no. 13, 2022, doi: 10.3390/app12136425.

[28] X. Liu, G. Meng, and C. Pan, "Scene text detection and recognition with advances in deep learning: a survey," *International Journal on Document Analysis and Recognition*, vol. 22, no. 2, pp. 143–162, 2019, doi: 10.1007/s10032-019-00320-5.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, May 2015, doi: 10.1109/ACCESS.2021.3053408.

[30] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 642–656, 2020, doi: 10.1007/s11263-019-01204-1.

[31] N. Otsu, "Threshold selection method from gray-level histograms," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979, doi: 10.1109/tsmc.1979.4310076.

[32] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition*, vol. 39, no. 3, pp. 317–327, 2006, doi: 10.1016/j.patcog.2005.09.010.

[33] N. Phansalkar, S. More, A. Sabale, and M. Joshi, "Adaptive local thresholding for detection of nuclei in diversity stained cytology images," in *ICCSP 2011 - 2011 International Conference on Communications and Signal Processing*, 2011, pp. 218–220, doi: 10.1109/ICCSP.2011.5739305.

[34] F. Z. A. Bella, M. El Rhabi, A. Hakim, and A. Laghrib, "An innovative document image binarization approach driven by the non-local p-Laplacian," *Eurasip Journal on Advances in Signal Processing*, vol. 2022, no. 1, 2022, doi: 10.1186/s13634-022-00883-2.

[35] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," *IEEE Transactions on Image Processing*, vol. 7, no. 6, pp. 918–921, 1998, doi: 10.1109/83.679444.

[36] Y. C. Wei and C. H. Lin, "A robust video text detection approach using SVM," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10832–10840, 2012, doi: 10.1016/j.eswa.2012.03.010.

[37] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10 SPEC. ISS., pp. 761–767, 2004, doi: 10.1016/j.imavis.2004.02.006.

[38] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *Proceedings of the IEEE International Conference on Computer Vision*, ICCV 2015, pp. 4651–4659, doi: 10.1109/ICCV.2015.528.

[39] P. Batra, N. Phalnikar, D. Kurmi, J. Tembhurne, P. Sahare, and T. Diwan, "OCR-MRD: performance analysis of different optical character recognition engines for medical report digitization," Int. J. Inf. Technol., vol. 16, no. 1, pp. 447–455, Jan. 2024, doi: 10.1007/s41870-023-01610-2.

[40] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, "DocUNet: document image unwarping via a stacked U-Net," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4700–4709, doi: 10.1109/CVPR.2018.00494.

[41] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 2642–2651, 2017, doi: 10.1109/CVPR.2017.283.

[42] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICDAR2017 competition on document image binarization (DIBCO 2017)," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2017, vol. 1, pp. 1395–1403, doi: 10.1109/ICDAR.2017.228.

[43] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018, doi: 10.1109/TMM.2018.2818020.

BIOGRAPHIES OF AUTHORS

Puneeth Prakash is an Assistant Professor at Maharaja Institute of Technology Mysore. He has 8 years of teaching and research experience. He has done bachelor's degree in information science and a master in computer science from VTU Belagavi. His key area of interest is image processing, machine learning, and computer vision. He mainly works on scene text-related images and has published papers in national and international journals. He is keen on multiprogramming paradigm implementation. He can be contacted at email: puneeth.phd20@gmail.com.

Sharath Kumar Yeliyur Hanumanthaiah is a Professor and Head in Department of Information Science and Engineering, Maharaja Institute of Technology Mysore. His areas of interest are image processing, pattern recognition, and information retrieval. He has around 12 years of experience in teaching. He completed a B.E. in computer science and engineering from VTU and an M.Tech. in computer cognition technology from the University of Mysore. Further, completed Ph.D. from the University of Mysore. Published 50 research articles in reputed conferences and journals. He served as the BoE and BoS of the University of Mysore from 2016 to 2020. He can be contacted at email: sharathyhk@gmail.com.