DEEP LEARNING MODEL MORAN ARCHITECTURE FOR TEXT RECOGNITION IN COMPLEX IMAGES

By

THURAKA GNANA PRAKASH * SUMALATHA L. ** SUJATHA B. ***

*-** Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Kakinada, East Godavari, Andhra Pradesh, India.
*** Department of Computer Science and Engineering, Godavari Institute of Engineering and Technology, Rajahmundry, East Godavari, Andhra Pradesh, India.

https://doi.org/10.26634/jit.13.3.21202

Date Received: 26/09/2024 Date Revised: 30/09/2024 Date Accepted: 04/10/2024

ABSTRACT

Recognizing text in images poses significant challenges, particularly in the presence of complex backgrounds. This technology plays a crucial role in assisting visually impaired individuals and interpreting semantic content. This survey explores various techniques developed over the past decade to address text recognition in complex images. It provides an overview and analysis of accumulated works and evaluates the performance of these recognition methods. While image complexity is difficult to quantify, it can be described using parameters such as background details, noise levels, lighting conditions, textures, and fonts. Furthermore, the survey highlights several benchmark datasets employed in the reviewed studies. By examining these works, challenges in the field are identified and compared.

Keywords: Complex Images, Image Processing, Text Recognition, CNN, MORAN Architecture, Preprocessing.

INTRODUCTION

Multimedia plays a crucial role in daily life, with a significant amount of information conveyed through images, videos, and other media. Text recognition becomes relatively straightforward when the text color contrasts sharply with the background. Images with plain backgrounds are generally classified as simple. However, image complexity arises from various factors, including the text, background details, lighting conditions, reflections, haze, natural surroundings, and other disturbances. Image complexity can be broadly classified based on these parameters into two types, scene complexity and text complexity. Text recognition in complex images is a critical challenge in the area of computer vision and pattern recognition (Cho et al., 2021; Zhao et al., 2020). Complex images frequently contain various obstacles such as varying fonts, sizes, orientations, backgrounds, and noise, making text extraction a nontrivial task. In the development of deep learning, there has been a prominent increase in the use of powerful tools designed to enhance the accuracy and efficiency of text recognition. Text recognition in complex images using deep learning models is a rapidly evolving field. While significant progress has been made, ongoing research continues to address the inherent challenges in achieving robust and accurate text recognition across diverse and complex visual environments (Xu et al., 2021).

i-manager Publications
Member of:
This paper has objectives related to SDGs
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

1. Literaure Review

Scene complexity refers to the various factors within an image that make text recognition challenging. These factors can include cluttered backgrounds, varying lighting conditions, multiple overlapping objects, text distortion, and diverse font styles. In complex scenes, the interplay of these elements can significantly affect the accuracy of text recognition models. Understanding and addressing scene complexity is crucial for developing robust text recognition systems.

1.1 Comparative Study

MORN training is sanctioned by the ASRN and is exclusively reliant on captions. It employs a weak detection method without the use of spectral or pixel-level visualization. To streamline the training process of this network, an initial coordinate network was established, with each pixel in the image assigned specific positional coordinates. MORN utilizes these coordinates to learn and create an offset grid, which samples pixel values to modify the image. The resulting edited image is then provided to ASRN. According to ASRN, the parsing method that emphasizes focus predicts the correct words based on the edited images. However, Alshawi et al. (2024) have indicated that manual focus methods fall short in achieving precise sharing between workspaces and targets. Consequently, a fractional retrieval method is introduced for ASRN training. This method applies various stretch scales to different segments of the feature maps, resulting in random alterations of point features during each iteration of the training phase. Through fractional training and retrieval, ASRN demonstrates enhanced robustness to contextual diversity. Testing confirms that ASRN can accurately identify target objects. Furthermore, a curriculum learning strategy has been devised for instructing MORAN. Given the reciprocal performance benefits between MORN and ASRN, the former will be adjusted to optimize the latter. Ultimately, both MORN and ASRN will undergo global optimization to enhance overall performance (Yin et al., 2013).

1.2 Factors Contributing to Scene Complexity

- *Background Clutter:* Complex scenes frequently contain numerous objects, patterns, and textures that can obscure or blend with the text. For instance, recognizing text on a signboard in a busy street scene may be difficult due to overlapping elements like trees, people, or vehicles. Background clutter can create noise that confuses the text recognition model, leading to errors in detection and classification.
- *Lighting Variations:* Lighting conditions in an image can vary widely, from bright sunlight causing reflections and glare to dim or uneven lighting resulting in shadows and low contrast. These variations can distort the appearance of text, making it harder for models to differentiate text from the background or recognize characters accurately. Adaptive techniques that account for different lighting scenarios are essential for handling such complexities.
- *Text Distortion:* In real-world images, text is frequently distorted due to perspective, curvature, or motion. For example, text on a curved surface, such as a cylinder or sphere, can appear stretched or compressed. Similarly, text captured at an angle may suffer from perspective distortion. These distortions complicate the recognition process, requiring advanced models capable of normalizing or compensating for these effects.
- *Font and Style Variability:* Text in complex scenes can appear in a wide range of fonts, sizes, and styles, including decorative or handwritten fonts. This variability can challenge models trained primarily on standard fonts, as they must generalize across different appearances. Recognizing stylized or artistic fonts frequently requires models to learn more abstract features rather than relying solely on pixel patterns.
- *Text Orientation and Layout:* In contrast to the usual horizontal alignment of standard printed text, text in complex scenes can appear in diverse orientations, including vertical, diagonal, or circular. Additionally, text may be laid out in non-linear arrangements, such as around the edge of an object or in a curved path. These irregular orientations and layouts require models to be flexible and capable of handling non-traditional text arrangements.
- *Occlusion and Overlapping Objects:* Text in complex scenes is frequently partially obscured by other objects, such as foliage, people, or additional signage. This occlusion can hide parts of the text, making it difficult for models to recognize the entire

word or phrase. In some cases, the overlapping objects may be semi-transparent, further complicating recognition.

- *Low-Resolution and Noisy Images:* Complex scenes captured in low resolution or with noise (e.g., motion blur, sensor noise) can significantly degrade text quality. Recognizing text in such conditions is challenging because fine details may be lost, and characters may merge together, leading to misinterpretation by the model.

1.3 Addressing Scene Complexity

- *Multi-Scale Feature Extraction: To* address the varying sizes and resolutions of text in complex scenes, deep learning models commonly utilize multi-scale feature extraction. Techniques such as multi-scale convolutional neural networks (CNNs) or pyramid networks enable the model to detect and recognize text across various scales, enhancing performance in scenes where text appears in different sizes.
- *Attention Mechanisms:* These mechanisms allow models to concentrate on key areas of an image, reducing the influence of background noise and unrelated elements. By directing the model's attention to areas most likely to contain text, these mechanisms can improve accuracy in recognizing text within complex scenes. This approach is particularly useful when dealing with cluttered backgrounds or when the text is surrounded by distracting elements.
- *Data Augmentation:* Training deep learning models with augmented data can help them generalize better to complex scenes. Techniques like adding noise, applying perspective distortions, or simulating different lighting conditions can prepare the model to handle a wide range of real-world scenarios. Data augmentation increases the model's robustness to scene complexity by exposing it to diverse training examples.
- *Scene Text Detection Models:* Specialized models like EAST (Efficient and Accurate Scene Text) and CTPN (Connectionist Text Proposal Network) are designed to detect text in complex scenes by generating text proposals from different parts of the image. These models focus on localizing text regions even in challenging conditions, providing accurate bounding boxes for subsequent recognition stages.
- *Transfer Learning and Pre training:* Transfer learning allows models to leverage knowledge from related tasks or domains, improving their ability to recognize text in complex scenes. Pretraining on large datasets, such as ImageNet, can help models learn robust features that are useful for recognizing text in diverse and challenging environments.
- *Use of Synthetic Datasets:* Synthetic datasets, like SynthText, simulate complex scenes with varying text properties, providing a rich source of training data for deep learning models. These datasets can include a wide range of scene complexities, such as different backgrounds, lighting, and text orientations, helping models generalize better to real-world scenarios.
- *Street Sign Recognition:* Recognizing text on street signs in urban environments involves dealing with complex scenes characterized by background clutter, varying lighting conditions, and occlusion by objects like trees or vehicles. Models used in this application must be robust to these challenges to provide accurate recognition for applications like autonomous driving or navigation systems.
- *Document Digitization in Real-World Conditions:* Document digitization typically involves scanning or photographing documents under less-than-ideal circumstances, such as poor lighting, skewed angles, or low resolution. Text recognition in such scenarios demands models capable of handling distortions and noise while accurately extracting textual information (Zhao et al., 2022b).
- *Augmented Reality (AR) Applications:* AR applications that overlay translated text or annotations onto real-world scenes must deal with complex environments where text may appear in

various orientations, fonts, and lighting conditions. Deep learning models employed in these applications must be capable of recognizing and interpreting text in dynamic and frequently unpredictable scenes (Liu et al., 2019). Scene complexity poses significant challenges for text recognition, requiring advanced deep learning techniques to ensure accurate and reliable results. By addressing factors such as background clutter, lighting variations, and text distortion, developers can create more robust text recognition systems capable of performing well in diverse and challenging environments. The continuous evolution of deep learning models, combined with innovative approaches to handling scene complexity, will drive further improvements in the field of text recognition (Wang et al., 2022).

1.4 Text Complexity

Text complexity refers to the inherent challenges associated with the textual content itself, which can significantly affect the performance of text recognition systems. Unlike scene complexity, which involves the surrounding environment, text complexity focuses on the characteristics of the text that make it difficult to detect, segment, or recognize accurately. Factors contributing to text complexity include font style, size, orientation, language, and the presence of distorted or low-quality text. Understanding and addressing text complexity is crucial for enhancing the robustness and accuracy of deep learning models in text recognition tasks (Shi et al., 2016).

1.4.1 Factors Contributing to Text Complexity

- *Font Variability:* Text can appear in a wide range of fonts, from standard and well-structured fonts to highly stylized or decorative ones. This variability can pose a significant challenge for recognition models, especially when dealing with fonts that deviate from those seen during training. Additionally, the use of bold, italic, or underlined text can further increase complexity, as these styles may alter the shape and spacing of characters.
- *Text Size and Resolution:* Text in images can vary greatly in size, ranging from small labels on products to large billboards. Small text is particularly challenging, as it tends to appear blurry or pixelated in images, especially at lower resolutions. Recognizing small text requires models that can detect subtle features and distinguish characters that are close together or partially obscured.
- *Text Orientation and Skew:* In complex images, text may not always be aligned horizontally; it could be oriented vertically, diagonally, or even circularly. Skewed or rotated text, such as that found on packaging or in artistic layouts, adds to the complexity, as traditional models trained on horizontal text may struggle to recognize characters correctly.
- *Language and Script Diversity:* Recognizing text in multiple languages or scripts within the same image adds a layer of complexity. Different languages have unique character sets, word structures, and writing systems (e.g., Latin, Cyrillic, Arabic, Chinese). Some scripts, like cursive handwriting or complex ideograms, present additional challenges due to their intricate and connected character forms. Multilingual text recognition models must be capable of identifying and processing these diverse scripts accurately.

2. Proposed Methodology

Multimedia plays a crucial role in daily life, with a substantial amount of information shared through images, videos, and other forms of media. Although these images vary, text recognition becomes more efficient when the color of the text contrasts with the background color. Figure 1 shows the proposed methodology implemented to address this challenge. By leveraging techniques that enhance text visibility through optimal color contrast, the methodology aims to improve the accuracy and efficiency of text recognition across diverse multimedia applications. Typically, an image with a plain background is considered a simple image. However, complexity in an image arises from various

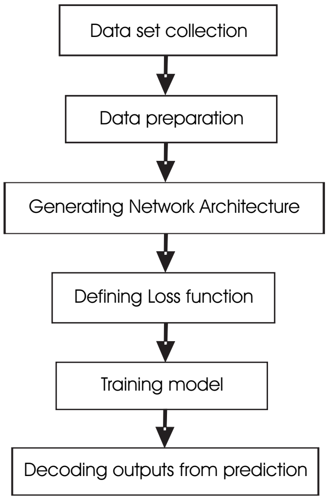

Figure 1. Proposed Methodology Implemented

factors such as the text itself, the background, lighting conditions, reflections, haze, natural surroundings, and other disturbances. Image complexity can be broadly classified based on these parameters.

Various methods are available for detecting text in images, each designed to address different types of image challenges. These techniques are applied to images with hazy backgrounds, complex environments, and other variations. Saha et al. (2020) and Zhao et al. (2022a) explored several text detection methods, including character-based, word-based, and text-line-based approaches, along with their respective applications. Their framework works in two stages: initially, the input image is split into two sets, with one being the word stroke, which predicts the text region. However, these predicted text regions may include small false positives. Despite this, the model generates word and line-level results, demonstrating enhanced performance in detecting multi-oriented and inconsistent text using standard datasets (Yin et al., 2013; Zhang et al., 2016).

2.1 Collecting Dataset

The Visual Geometry Group will provide a large dataset of 10 GB of images; however, only 40,000 images have been used for training purposes.

2.2 Preprocessing

Preprocessing is carried out to ensure compatibility between the dataset and the implemented model. The steps to be followed are:

- The images are read and converted into grayscale format.
- Each image is resized to dimensions of (128, 32) using padding where necessary.
- The image dimensions are expanded to (128, 32, 1) to align with the input shape of the model architecture.
- Pixel values are normalized by dividing them by 255.

The output labels preprocessed using the following steps are:

- The text is extracted from the image titles, as the titles contain the embedded text.
- Each character of a word is encoded into numerical values using a mapping function.
- The maximum word length is computed, and each label is padded to match this maximum length, ensuring compatibility with the output shape of the RNN model.

In the preprocessing step, two additional lists are created: one for the name length and another for the input length to the RNN. The name length corresponds to the length of each output text label, while the input length, fixed at 31 in the architecture, ensures consistency for each input to the LSTM layer.

2.3 Creating Model Network Architecture

The MORAN framework, which stands for Multi-Object Rectified Attention Network, is composed of two main components: MORN (Multi-Object Rectification Network) and ASRN (Attention-based Sequence Recognition Network) (Luo et al., 2019; Shi et al., 2016). MORN is designed to reduce image complexity by adjusting and aligning characters uniformly, thereby simplifying the text for subsequent processing steps. Once the text is properly aligned, it is ready for recognition by the ASRN, which reads the text and predicts the output word using a CNN-LSTM network. This model has shown promising results, particularly for recognizing irregular text. The focus was primarily on text complexity rather than scene complexity.

Cho et al. (2021) proposed a recognition framework specifically for images affected by haze. Since haziness can obscure text and make it difficult to read, image processing techniques are applied to separate the text layer from the haze and convert it into a recognizable format. To remove the haziness, a bilateral filter is applied to the image. Figure 2 shows the creating model network architecture.

Once these disturbances are addressed and removed from the image, the text is localized more easily. This method effectively handles scene complexity. Figure 3 shows the rectification of curved text (Dai et al., 2019; Cong et al., 2019; Wang et al., 2022).

Figure 3. Rectification of Curved Text

2.4 CTC Loss Function

CTC Loss function, which is a translation layer used to predict output for each time step, is employed to train the RNN. This method helps address the alignment issue in handwritten text, as handwritten text varies in writing style across different writers. The ground truth text (what is written in the image) and BLSTM output are provided, and the loss is calculated by minimizing the negative maximum probability path. The CTC loss function only needs to know the content present in the image, ignoring both the position and width of the characters. No further processing of the recognized text is required.

2.5 Decoding Outputs From Prediction

A region proposal network was employed to detect text within the image, with the identified regions forwarded to the extraction process using a recurrent network. The approach is carried out in two stages: text and non-text classification, followed by region extraction. The focus is on handling scene complexity, text orientation, and implementing the region proposal network through which text regions are detected and subsequently processed by the recurrent network (Liu et al., 2023). This helps in extracting the text from the detected region and achieves good performance for benchmark datasets. Wang et al. (2022) divided the process into text detection, localization, and recognition. In text detection, different methods based on texture, connected components (Shi et al., 2016), and deep learning models like CNN (Convolutional Neural Network), VGG, LSTM (Long Short-Term Memory), etc., are used. Recognition methods are based on words and characters. An end-to-end recognition system is a fusion method combining text detection and recognition (Li et al., 2024; Zhao et al., 2022b). After the rectangles are formed on the image, text is extracted one by one by analyzing its neighboring areas. This process is based on texture, features, and edge information. Using these parameters, the text is recognized.

3. Experimental Results

This method is used to process an input image, applying the models mentioned above to obtain the final output.

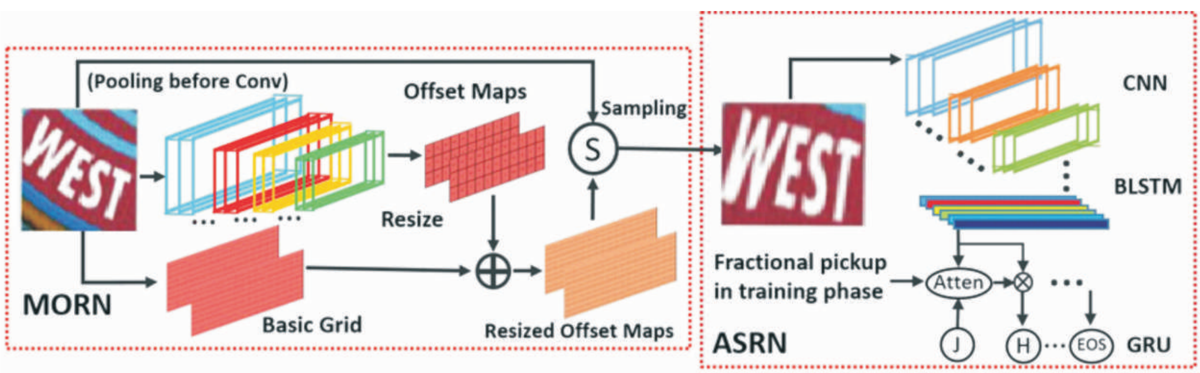

Figure 2. Creating Model Network Architecture

The discussion includes handling text, scene complexity, and techniques for removing disturbances, such as binarization of images and background removal preprocessing actions. The performance of the models is compared using benchmark datasets. The model has successfully processed blurred images and text with different orientations. Connected components were implemented to filter text and non-text regions, as text typically remains aligned and consistent within the same line, making it easier to filter. Text is localized and extracted from these regions. Image processing techniques are used for background removal, followed by connected component analysis for text extraction (Selvam et al., 2022). Figure 4 shows the application of MORAN to rectify irregular text with small curve angles.

Figure 4. MORAN to Rectify Irregular Text with Small Curve Angles

Conclusion

Text recognition systems under complex conditions and disturbances focus on extracting text from intricate backgrounds and isolating the foreground from the background using image preprocessing techniques. For images with complex backgrounds, the primary emphasis is on background and noise removal techniques. Deep learning models are commonly employed as a subsequent step in many methods. Different models have been utilized for text recognition, with their performance on benchmark datasets presented in detail. The study provides a concise analysis of the challenges associated with different datasets. Despite substantial advancements in the field, no model has yet achieved complete success for practical use, highlighting the need for further research and development in this domain.

References

[1]. Alshawi, A. A. A., Tanha, J., & Balafar, M. A. (2024). An Attention-Based Convolutional Recurrent Neural Networks for Scene Text Recognition. *IEEE Access.*

[2]. Cho, S. J., Ji, S. W., Hong, J. P., Jung, S. W., & Ko, S. J. (2021). Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4641-4650).

[3]. Cong, F., Hu, W., Huo, Q., & Guo, L. (2019, September). A comparative study of attention-based encoder-decoder approaches to natural scene text recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 916-921). IEEE.

https://doi.org/10.1109/ICDAR.2019.00151

[4]. Dai, P., Zhang, H., & Cao, X. (2019). Deep multi-scale context aware feature aggregation for curved scene text detection. *IEEE Transactions on Multimedia,* 22(8), 1969-1984.

https://doi.org/10.1109/TMM.2019.2952978

[5]. Li, N., Wang, Z., Huang, Y., Tian, J., Li, X., & Xiao, Z. (2024). A multi-scale natural scene text detection method based on attention feature extraction and cascade feature fusion. *Sensors,* 24(12), 3758.

https://doi.org/10.3390/s24123758

[6]. Liu, Y., Jin, L., Zhang, S., Luo, C., & Zhang, S. (2019). Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition,* 90, 337-345.

https://doi.org/10.1016/j.patcog.2019.02.002

[7]. Liu, Y., Wang, Y., & Shi, H. (2023). A convolutional recurrent neural-network-based machine learning for scene text recognition application. *Symmetry,* 15(4), 849.

https://doi.org/10.3390/sym15040849

[8]. Luo, C., Jin, L., & Sun, Z. (2019). Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition,* 90, 109-118.

https://doi.org/10.1016/j.patcog.2019.01.020

[9]. Saha, S., Chakraborty, N., Kundu, S., Paul, S., Mollah, A. F., Basu, S., & Sarkar, R. (2020). Multi-lingual scene text detection and language identification. *Pattern Recognition Letters,* 138, 16-22.

https://doi.org/10.1016/j.patrec.2020.06.024

[10]. Selvam, P., Koilraj, J. A. S., Romero, C. A. T., Alharbi, M., Mehbodniya, A., Webber, J. L., & Sengan, S. (2022). A transformer-based framework for scene text recognition. *IEEE Access,* 10, 100895-100910.

https://doi.org/10.1109/ACCESS.2022.3207469

[11]. Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 39(11), 2298-2304.

https://doi.org/10.1109/TPAMI.2016.2646371

[12]. Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 39(11), 2298-2304.

https://doi.org/10.1109/TPAMI.2016.2646371

[13]. Wang, T., Wu, X., & Shi, B. (2022). Revisiting scene text recognition: A data perspective. *arXiv preprint arXiv:2307.08723*.

https://doi.org/10.48550/arXiv.2307.08723

[14]. Xu, S., Chen, C., Peng, S., & Hu, X. (2021, October). CCANet: Exploiting pixel-wise semantics for irregular scene text spotting. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (pp. 1-7). IEEE.

https://doi.org/10.1109/CISP-BMEI53629.2021.9624403

[15]. Yin, X. C., Yin, X., Huang, K., & Hao, H. W. (2013). Robust text detection in natural scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 36(5), 970-983.

https://doi.org/10.1109/TPAMI.2013.182

[16]. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., & Bai, X. (2016). Multi-oriented text detection with fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4159-4167).

[17]. Zhao, F., Yu, J., Xing, E., Song, W., & Xu, X. (2022a). Real-time Scene Text Detection Based on Global Level and Word Level Features. *arXiv preprint arXiv:2203.05251*.

https://doi.org/10.48550/arXiv.2203.05251

[18]. Zhao, M., Feng, W., Yin, F., Zhang, X. Y., & Liu, C. L. (2022b). Mixed-supervised scene text detection with expectation-maximization algorithm. *IEEE Transactions on Image Processing,* 31, 5513-5528.

https://doi.org/10.1109/TIP.2022.3197987

[19]. Zhao, S., Sun, L., Li, G., Liu, Y., & Liu, B. (2020). A CCD based machine vision system for real-time text detection. *Frontiers of Optoelectronics,* 13(4), 418-424.

https://doi.org/10.1007/s12200-019-0854-0

ABOUT THE AUTHORS

T. Gnana Prakash is currently pursuing his Ph.D. from Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India. He completed his B.Tech. and M.Tech. in Computer Science and Engineering in 2006 and 2010, respectively, from Jawaharlal Nehru Technological University, Hyderabad, Telangana, India. He is an accomplished educator with 17 years of teaching experience and 7 years of research in the field of Computer Science. His research focuses on Computer Vision, Deep Learning, and Machine Learning, aiming to develop innovative solutions and applications.

Dr. Sumalatha Lingamgunta is currently a Professor in the Department of Computer Science at Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India. She has over 22 years of teaching experience at the University College of Engineering, Kakinada, Andhra Pradesh. She has served as the Head of the CSE Department twice and has held various administrative posts, including Director of Evaluation, Director of IIIPT, and Chief Coordinator of the Skill Development Center. During her tenure as Chief Coordinator for the Skill Development Center, she played a key role in establishing the APSSDC SIEMENS Centre of Excellence at UCEK. She is a faculty champion for the University Innovation Fellows Program at Stanford University, CA, USA, and the Research Lead of the AI Platoon research group, in collaboration with leadingindia.ai, an AI & Deep Learning research initiative. She is a member of several professional bodies. Her research interests include Network Security, Image Processing, Machine Learning, and Natural Language Processing.

Dr. B. Sujatha currently serves as an Associate Professor at Godavari Global University, Rajamahendravaram, Andhra Pradesh, India. She earned her Ph.D. in Computer Science from the University of Mysore, Mysuru, Karnataka, India. She completed her M.Tech. in Computer Science and Engineering from Andhra University, Visakhapatnam, Andhra Pradesh, India, in 2002, and her B.Tech. degree from Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India, in 1997. With over a decade of teaching experience, Dr. Sujatha has made significant contributions to her field, with 63 research publications in international journals. She is an active member of the SRRF-GIET, Rajahmundry, Andhra Pradesh. Her research focuses on Image Processing and Pattern Recognition, and she is a Life Member of the Indian Science Congress Association (ISCA).