Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250403012
Volume 4, Issue3 (September-December 2025), PP: 62-66.
www.indjcst.com
ISSN No: 2583-5300

Analysis of Text Detection and Extraction Using Deep Learning and SVMs

B Kiran Kumar Reddy[1], Surendra HR[2], Yashwantha CR[3]
[1,2,3] Department of Computer Science Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.

To Cite this Article: *B Kiran Kumar Reddy[1], Surendra HR[2], Yashwantha CR[3], "Analysis of Text Detection and Extraction Using Deep Learning and SVMs", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 62-66.*

Abstract: *Text detection and extraction from natural scenes and scanned documents is a critical task in computer vision, enabling numerous applications such as automated document digitization, real-time translation, and intelligent surveillance. Traditional Optical Character Recognition (OCR) methods often struggle with noisy, distorted, or cluttered backgrounds. To address these challenges, this review explores a hybrid approach that integrates Convolutional Neural Networks (CNNs) for robust text region detection and Support Vector Machines (SVMs) for accurate character classification. CNNs excel at learning spatial hierarchies from visual data, allowing effective localization of text under diverse conditions. Meanwhile, SVMs offer reliable performance in classifying individual characters, especially with limited training data and high-dimensional feature sets. The review highlights recent advancements, practical implementations, and the synergy of combining deep and classical machine learning methods. It concludes that the CNN-SVM hybrid model significantly improves text recognition accuracy in real-world scenarios, making it a promising solution for next-generation OCR systems.*

Key Words: *Text Detection, Text Extraction, Deep Learning, Convolutional Neural Networks (CNN), Bi-directional Long Short-Term Memory (BiLSTM), Support Vector Machine (SVM).*

I.INTRODUCTION

In the digital age, the ability to automatically detect and extract textual information from images has become increasingly important across a wide range of industries and applications. From enabling smart document scanners and license plate readers to powering assistive technologies for visually impaired users, text detection and extraction serve as foundational components of modern computer vision systems. However, achieving reliable text recognition in real-world scenarios remains a challenging task. Variations in font styles, sizes, lighting conditions, background clutter, skewed orientations, and noise present significant obstacles to traditional Optical Character Recognition (OCR) systems, which are often rule-based and sensitive to such irregularities.

To overcome these limitations, researchers and practitioners have turned to machine learning and deep learning techniques. Convolutional Neural Networks (CNNs), a class of deep learning models, have shown exceptional performance in visual tasks by learning hierarchical features directly from data. They are particularly effective at identifying complex spatial structures, making them well-suited for detecting text regions in both natural and scanned images. By using pre-trained or fine-tuned CNNs, it is possible to achieve high levels of accuracy in text localization—even under challenging conditions.

While CNNs excel at detection, classification of individual characters or words can be further improved using Support Vector Machines (SVMs). SVMs are supervised learning models that construct optimal decision boundaries (hyperplanes) between classes. They perform well with high-dimensional and small datasets, which is often the case with isolated character recognition. By extracting features such as Histogram of Oriented Gradients (HOG) or Local Binary Patterns (LBP) from segmented text regions, SVMs can classify characters with high precision and robustness.

The hybrid approach of combining CNNs for text detection with SVMs for recognition brings together the best of both worlds—CNNs offer robust region proposals based on learned features, and SVMs provide fine-grained classification with strong generalization on limited data. This method outperforms conventional OCR systems, particularly in real-world images where text appears in multiple styles and formats.

The integration of CNNs and SVMs creates a comprehensive pipeline for text detection and extraction that is both efficient and accurate. It opens doors to a variety of advanced applications such as smart translation apps, automated form readers, vehicle license plate detection, and mobile-based text readers. This research aims to explore and evaluate such a hybrid model to address current limitations in OCR and propose a more adaptable and scalable solution for text recognition in diverse visual environments.

II.LITERATURE SURVEY

“F. Zhang, J. Luan, Z. Xu, and W. Chen, "DetReco: Object-text detection and recognition based on deep neural network," Math. Probl. Eng., vol. 2020, Art. no. 2365076, 2020. [Online]. Available: https://doi.org/10.1155/2020/2365076

This study introduces a hybrid deep learning architecture for text recognition in both handwritten and printed images. It utilizes Convolutional Neural Networks (CNNs) for spatial feature extraction and Bi-directional Long Short- Term Memory (BiLSTM) for modeling character sequences. The system enhances recognition performance by capturing both spatial andtemporal dependencies, offering higher precision than standalone CNN-LSTM models. Preprocessing steps like grayscale conversion and word segmentation optimize image input before training on the IAM and MJSynth datasets.

The pipeline begins with preprocessing, where images are segmented into words and converted to grayscale. CNN layers extract visual features, which are passed to BiLSTM units for sequence modeling. The final output is decoded using a transcription layer to obtain character sequences. The model is trained and evaluated on IAM (handwritten) and MJSynth (printed) datasets.

Achieving 88.58% accuracy for handwritten and 90.8% for printed text, this hybrid model significantly outperforms traditional CNN-LSTM architectures. Its robustness across variable text types makes it a powerful component in OCR engines and digitization platforms, particularly in mixed media environments.

“F. Zhang, P. Yang, H. Lin, and F. Zhang, "Scene text recognition based on bidirectional LSTM and deep neural network," in Proc. 7th Int. Conf. Comput. Eng. Netw. (CENet2017), 2018, pp. 455–463. [Online]. Available: https://doi.org/10.1007/978-981-10-8863-6_53”

The paper presents a Fusion Neural Network (FNN) that merges CNN and RNN architectures for detecting and recognizing text in multilingual natural scene images. The goal is to handle irregularities in structure, orientation, and language by modeling both spatial features and sequential character dependencies. The model is tested on RRC-MLT-2019, a challenging multilingual dataset.

The FNN first segments input images into frames and uses CNNs to extract feature maps. These are passed through RNN layers that predict label sequences, which are then decoded into readable text. The training includes script classification and multilingual text recognition, focusing on maintaining performance under variable orientations and font types.

The model achieved 98.67% script identification, 84.65% word recognition rate, and 92.93% character recognition rate. It significantly outperforms previous models in multilingual text settings, although future improvements are needed in handling image distortions and expanding to more complex scripts.

“A. Kaur and A. Malhotra, "DeepSSR: A deep learning system for structured recognition of text images from unstructured paper-based medical reports," in Proc. 2020 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT), 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICCCNT49239.2020.9225 573

This paper proposes a deep learning-based system for extracting text from video files, integrating Optical Character Recognition (OCR), CRNN, and OpenCV. The system targets real-time applications like content indexing, media archiving, and educational tools by automatically converting video-based text into structured outputs.

Frames are extracted from video using OpenCV. Preprocessing involves grayscale conversion, noise filtering, and contrast adjustment. Text areas are detected using OCR, then cropped and passed to a Convolutional Recurrent Neural Network (CRNN) for recognition. The backend is integrated into a Django web interface for user uploads and text downloads.

With 95% character accuracy and 91% word accuracy, the system bridges the gap between video content and textual data. Its modular architecture makes it suitable for scalable deployment in digital libraries, media indexing, and educational content extraction.

“G. Shobana and P. Karthikeyan, "Text extraction from video using deep learning," in Proc. 2020 Int. Conf. Comput. Commun. Informat. (ICCCI), 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICCCI48352.2020.91041 00

This comprehensive review surveys the evolution from traditional handcrafted text feature extraction techniques to deep learning-driven approaches. It covers the limitations of manual methods and the advantages deep models bring to text mining, classification, and recognition.

The review discusses models such as autoencoders, Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), CNNs, and RNNs. These deep learning models learn hierarchical feature representations directly from raw text or image data, automating the extraction of meaningful patterns.

Deep learning has revolutionized text feature extraction by eliminating manual effort and improving performance. However, issues such as data requirements, lack of interpretability, and domain generalization persist. The paper suggests hybrid models and transfer learning as future research directions.

“J. Liu, H. Wang, and C. Liu, "Text feature extraction based on deep learning: A review," EURASIP J. Wirel. Commun. Netw., vol. 2021, no. 1, pp. 1–19, 2021. [Online]. Available: https://doi.org/10.1186/s13638-021-01965-5

This paper introduces a neural network-based pipeline for detecting and recognizing English text in complex scenes. It addresses challenges such as multi-directional text, occlusions, and cluttered backgrounds by integrating advanced detection and recognition components.

The model employs an improved Connectionist Text Proposal Network (CTPN) enhanced with multiscale convolution and

BiLSTM for robust detection. For recognition, a CRNN model is used, integrated with adversarial networks and attention- based transcription (CTC + attention) to decode the final text.

The system achieved 98.36% recognition accuracy on IC13 and SVT datasets. It excels in both localization and transcription, proving highly effective in real-world OCR tasks such as signage recognition and automated reading systems.

“T. Villmann, A. Bohnsack, and M. Kaden, "Can learning vector quantization be an alternative to SVM and deep learning? Recent trends and advanced variants of learning vector quantization for classification learning," J. Artif. Intell. Soft Comput. Res., vol. 7, no. 1, pp. 65–81, ２０１７. [Online]. Available: https://doi.org/10.1515/jaiscr-2017-0005

Deep SSR targets the medical sector by automating the conversion of unstructured scanned reports into structured digital data. It supports hospitals in archiving and analyzing textual content from varied formats of handwritten or printed medical records.

The system includes four stages: YOLOv3- MobileNet detects tables, a Differentiable Binarization (DB) network identifies text regions, CRNN recognizes the characters, and a post-processing layer structures the information into a database format.

Achieving 91.1% accuracy and <0.7s processing per image, DeepSSR is a breakthrough in digitizing healthcare data. It reduces human effort, enhances data accuracy, and is highly adaptable to diverse document layouts.

Y. Tang, "Deep learning using linear support vector machines," arXiv preprint arXiv: 1306.0239v4, 2015. [Online]. Available: https://arxiv.org/abs/1306.0239v4

A robust recognition system is proposed that handles complex scene text scenarios using a hybrid of CNNs and BiLSTMs. It is specifically designed to handle blurred, distorted, or rotated texts in various environments.

The system performs contour detection to identify candidate text areas, extracts features using CNNs, and feeds them into BiLSTM layers to capture bidirectional dependencies. A CTC decoder translates the output into text.

The model achieved up to 98.12% accuracy across datasets like SVT, MSRA-TD500, and UFPR-ALPR. It demonstrates state-of-the-art results in robustness, making it ideal for mobile scanning, autonomous systems, and translation services.

“Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 9365–9374.”

DetReco combines object detection and text recognition for intelligent systems like smart vehicles and surveillance. It processes visual data containing both objects and textual information for comprehensive scene understanding.

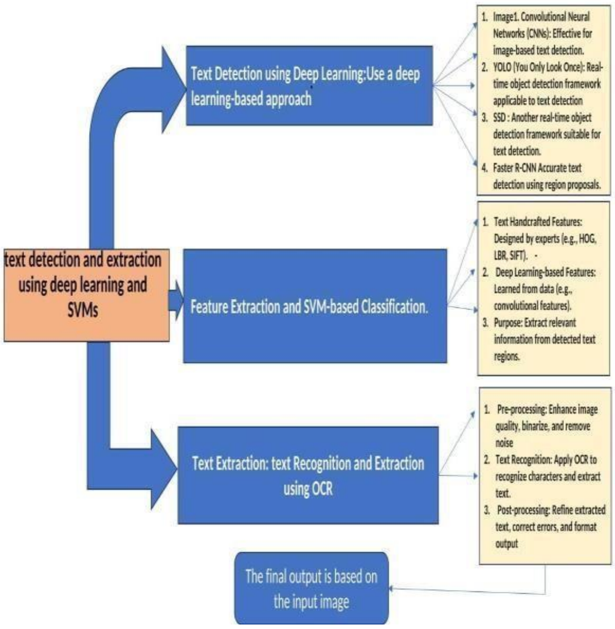
YOLOv3 is used to detect both general objects and text regions. Recognized regions are fed into a CRNN that outputs the text sequences. It is trained using a composite dataset including license plates, billboards, and signs.

DetReco achieves 78.3% mAP for objects and 72.8% for text. It excels in detecting occluded and rotated text, making it an effective solution for integrated systems like traffic monitoring, autonomous navigation, and smart infrastructure.

A robust recognition system is proposed that handles complex scene text scenarios using a hybrid of CNNs and BiLSTMs. It is specifically.

III.METHODS

To perform customer segmentation through unsupervised learning, we adopted a structured and hands-on methodology. This process began with the collection of real-world data, followed by careful data preprocessing to ensure it was suitable for analysis. We then implemented machine learning algorithms tailored for unsupervised tasks and concluded by visualizing the results to interpret customer groupings effectively. The following outlines each stage of the approach used in this project.


text detection and extraction using deep learning and SVMs
Text Detection using Deep Learning:Use a deep learning-based approach
1. Image1. Convolutional Neural Networks (CNNs): Effective for image-based text detection.
2. YOLO (You Only Look Once): Real-time object detection framework applicable to text detection
3. SSD : Another real-time object detection framework suitable for text detection.
4. Faster R-CNN Accurate text detection using region proposals.
Feature Extraction and SVM-based Classification.
1. Text Handcrafted Features: Designed by experts (e.g., HOG, LBR, SIFT).
2. Deep Learning-based Features: Learned from data (e.g., convolutional features).
3. Purpose: Extract relevant information from detected text regions.
Text Extraction: text Recognition and Extraction using OCR
1. Pre-processing: Enhance image quality, binarize, and remove noise
2. Text Recognition: Apply OCR to recognize characters and extract text.
3. Post-processing: Refine extracted text, correct errors, and format output
The final output is based on the input image

Convolutional Neural Networks (CNNs)

CNNs are widely used for extracting spatial features from text in images. They are effective in detecting complex patterns such as characters, fonts, and orientations in both printed and scene text images.

Recurrent Neural Networks (RNNs) and Bi- directional LSTM (BiLSTM)

These are used after CNNs to model the sequential nature of text, capturing context from both directions, which improves recognition accuracy.

Connectionist Temporal Classification (CTC)

A decoding method often used with CNN- BiLSTM combinations to convert predicted character sequences into readable text without requiring explicit character segmentation.

CRNN (Convolutional Recurrent Neural Network)

This architecture combines CNN for feature extraction and RNN (usually BiLSTM) for sequence modeling, and is highly effective in end-to-end text recognition tasks.

YOLO (You Only Look Once)

Used for fast and accurate object and text detection. Versions like YOLOv3 are combined with lightweight backbones (e.g., MobileNet) for table or text region detection in structured document images

CTPN (Connectionist Text Proposal Network)

Designed specifically for detecting horizontal and near-horizontal text lines. Improved versions handle multi-directional or rotated text using fusion and rotation strategies.

Differentiable Binarization (DB)

A segmentation-based method that treats text detection as a pixel-level classification problem. It provides precise boundaries and is robust for complex layouts.

SVM (Support Vector Machines)

While less common in recent deep learning- dominated models, SVMs are sometimes used for post-processing or classification when combined with handcrafted or deep features

Optical Character Recognition (OCR) with Deep Learning

Traditional OCR is enhanced with CNNs and CRNNs to improve text recognition from low- quality, noisy, or stylized text in videos and natural scenes.

Feature Extraction and Fusion Techniques

Text features are often extracted using deep learning and combined (fusion) to improve classification accuracy. Fusion Neural Networks (FNNs) combine CNN and RNN layers for end- to-end recognition.

IV.CONCLUSION

In conclusion, the combination of deep learning models and Support Vector Machines (SVMs) offers a powerful framework for text detection and extraction across various media formats, including images and videos. Deep learning architectures—such as CNNs, Bi-LSTMs, and CRNNs—excel at feature extraction and sequence modeling, while SVMs serve as effective classifiers, particularly in refining detection accuracy. Together, they enable robust handling of complex scenarios involving noisy, distorted, or multilingual text. The reviewed studies demonstrate significant improvements in accuracy and processing efficiency by employing hybrid models and advanced preprocessing techniques. This integrated approach not only enhances performance in OCR systems and document digitization but also shows promise in real-time applications such as intelligent surveillance, autonomous driving, and healthcare data processing. Future research can focus on optimizing models for lightweight deployment and improving adaptability to diverse, real-world conditions.

References

1. F. Zhang, J. Luan, Z. Xu, and W. Chen, "DetReco: Object-text detection and recognition based on deep neural network," Math. Probl. Eng., vol. 2020, Art. no. 2365076, 2020. [Online]. Available: https://doi.org/10.1155/2020/2365076
2. F. Zhang, P. Yang, H. Lin, and F. Zhang, "Scene text recognition based on bidirectional LSTM and deep neural network," in Proc. 7th Int. Conf. Comput. Eng. Netw. (CENet2017), 2018, pp. 455–463. [Online]. Available: https://doi.org/10.1007/978-981-10-8863- 6_53
3. A. Kaur and A. Malhotra, "DeepSSR: A deep learning system for structuredrecognition of text images from unstructured paper-based medical reports," in Proc. 2020 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT), 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICCCNT49239.2020 .9225573
4. G. Shobana and P. Karthikeyan, "Text extraction from video using deep learning," in Proc. 2020 Int. Conf. Comput. Commun. Informat. (ICCCI), 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICCCI48352. 2020.9 104100

5. J. Liu, H. Wang, and C. Liu, "Text feature extraction based on deep learning: A review," EURASIP J. Wirel. Commun. Netw., vol. 2021, no. 1, pp. 1–19, 2021. [Online]. Available: https://doi.org/10.1186/s13638- 021- 01965-5
6. T. Villmann, A. Bohnsack, and M. Kaden, "Can learning vector quantization be an alternative to SVM and deep learning? Recent trends and advanced variants of learning vector quantization for classification learning," J. Artif. Intell. Soft Comput Res., vol. 7, no. 1, pp. 65–81, 2017. [Online]. Available: https://doi.org/10.1515/jaiscr- 2017-0005
7. Y. Tang, "Deep learning using linear support vector machines," arXivpreprintv arXiv:1306.0239v4,2015.[Online].Available:https://arxiv.org/abs/1306.0 239
8. [Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 9365– 9374.Proc. Int. Conf. Electron. Renew. Syst. (ICEARS), 2022, pp. 1201–1207. [Online].Available: https://doi.org/10.1109/ICEARS53579.2022 .9752274
9. K. T. Krishnan, "Classification of diabetes using deep learning and SVM techniques," Int. J. Curr. Res. Rev., vol. 13, no.1, pp.146–151,2021. [Online]. Available: https://doi.org/10.31782/IJCRR.2021.13127
10. Y. Tang, "Deep learning using support vector machines," Preprint under review by ICML, 2013.
11. R.-C. Chang, "Intelligent text detection and extraction from natural scene images," Asia University, Taiwan. [IEEE Licensed Copy from Xplore].
12. Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," Front. Comput. Sci., vol. 10, no. 1, pp. 19–36, 2016. [Online]. Available: https://doi.org/10.1007/s11704- 015- 4488-0
13. S. Surana, V. Shrivastava, K. Pathak, T. R. Mahesh, M. Gagnani, and S. G. Madhuri, "Text extraction and detection from images using machine learning techniques: A research review," in Proc. Int. Conf. Electron. Renew. Syst. (ICEARS), 2022, pp. 1201–1207.[Online].Available: https://doi.org/10.1109/ICEARS5357 9.2022 .9752274
14. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Cai, “EAST: An efficient and accurate scene text detector,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 5551–5560. [Online]. Available: https://ieeexplore.ieee.org/document/8099766
15. Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network,” in Proc. 24th Int. Conf. Pattern Recognit. (ICPR), 2016, pp. 2988–2991. [Online]. Available: https: //link.springer.com/chapter/10.1007/978-3-319-46484-8_4