IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154273

A Review of Optical Character Recognition for Handwritten Hindi Text

Divya D[1], Bhuvaneshwari V[2], Shiva Kumar Swamy J[3], Siddu[4], Muhibur Rahman T.R[5], Anita Patil[6], Dadapeer[7]

6th Sem B.E. (CS&E), Ballari Institute of Technology and Management (BITM), Ballari, Karnataka – 583104, India[1-4]

Associate Professor**,** Ballari Institute of Technology and Management (BITM), Ballari, Karnataka – 583104, India[5]

Professor, Ballari Institute of Technology and Management (BITM), Ballari, Karnataka – 583104, India[6]

Asst Professor, Ballari Institute of Technology and Management (BITM), Ballari, Karnataka – 583104, India[7]

Abstract: Handwritten text recognition has long been a challenging area within the field of pattern recognition, especially for complex scripts such as Hindi written in the Devanagari script. Unlike printed text, handwritten Hindi exhibits significant variability in writing styles, stroke order, character shapes, and spacing, making Optical Character Recognition (OCR) a difficult problem. Over the past decade, advancements in machine learning, deep learning, and image processing techniques have significantly improved the performance of OCR systems for handwritten scripts.

This paper presents a detailed review of existing approaches for handwritten Hindi text recognition. It explores traditional methods based on feature extraction and classification, as well as modern deep learning approaches such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architectures. The study also examines preprocessing techniques, segmentation challenges, and benchmark datasets used in the domain. A structured classification of OCR systems is proposed based on methodology and level of automation.

Further, a comparative analysis is provided considering recognition accuracy, robustness, dataset dependency, and computational complexity. The review highlights that while deep learning models have achieved promising results, challenges such as lack of large annotated datasets, variability in handwriting, and segmentation errors still persist. The paper concludes by identifying research gaps and suggesting future directions for building more accurate, scalable, and real-time OCR systems for handwritten Hindi text.

Keywords: Optical Character Recognition, Handwritten Hindi, Devanagari Script, Deep Learning, CNN, RNN, Image Processing, Pattern Recognition, NLP

I. INTRODUCTION

Optical Character Recognition (OCR) refers to the process of converting text present in images into machine-readable form. While OCR for printed text has reached a high level of accuracy, handwritten text recognition continues to be a complex problem due to the inherent variability in human writing.

Hindi, written in the Devanagari script, is one of the most widely used languages in India. The script consists of vowels, consonants, modifiers, and compound characters, many of which are connected by a horizontal line known as the *shirorekha*. This structural complexity introduces additional challenges in segmentation and recognition.

Traditional OCR systems relied on handcrafted features and rule-based methods. However, these systems struggled with diverse handwriting styles. With the emergence of machine learning and deep learning, OCR systems have become more adaptive and capable of learning complex patterns directly from data.

This paper aims to review the evolution of handwritten Hindi OCR systems, analyze existing techniques, and identify the limitations that need to be addressed for practical deployment**.**

II. THEORETICAL BACKGROUND

A. Image Preprocessing

Preprocessing improves input image quality by applying techniques such as noise removal, binarization, skew correction, and normalization. This ensures clearer text for accurate recognition.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154273

B. Segmentation
Segmentation divides the text into lines, words, or characters. In handwritten Hindi, this is difficult due to connected characters and modifiers, making it a critical challenge.
C. Feature Extraction
This step converts images into meaningful features like strokes, shapes, and pixel patterns. Techniques such as structural features and HOG are commonly used.
D. Classification Techniques
Machine learning models like SVM and KNN classify characters based on extracted features. Their performance depends on feature quality.
E. Deep Learning Models
Deep learning methods such as CNN and LSTM automatically learn features and improve recognition accuracy, especially for complex handwriting.
F. Post-Processing
Language models and dictionaries are used to correct recognition errors and improve output accuracy.
G. Performance Metrics
System performance is evaluated using Accuracy, Character Error Rate (CER), and Word Error Rate (WER).

III. SYSTEM CLASSIFICATION

Handwritten Hindi OCR systems can be categorized into four levels:
Tier 1: Basic Image Processing Systems
These systems focus only on preprocessing and segmentation without intelligent recognition.
Tier 2: Feature-Based OCR Systems
These rely on handcrafted features and traditional classifiers such as SVM or KNN.
Tier 3: Machine Learning-Based Systems
These systems use trained models for classification but still depend on manual feature extraction.
Tier 4: Deep Learning-Based OCR Systems
These systems integrate end-to-end learning, eliminating the need for manual feature engineering and achieving higher accuracy.

IV. LITERATURE REVIEW

Research in handwritten Hindi OCR has evolved from traditional segmentation and feature-based methods to advanced deep learning and transformer-based approaches. Early works focused on segmentation and manual feature extraction but faced limitations with complex handwriting. Machine learning techniques improved classification accuracy, while CNN-based deep learning models further enhanced performance by automatically learning features. Recent transformer-based models, such as TrOCR, provide better context understanding and higher accuracy. Additionally, post-processing techniques using language models help correct recognition errors. Overall, the trend shows a shift toward more accurate and intelligent OCR systems.

Table I: Literature Review Summary

Sl. No.	Author(s)	Year & Title	Method / Technique	Key Findings	Venue
1	Upreti & Bag	2016 – Segmentation of Handwritten Hindi Words	Polygonal approximation	Effective segmentation of unconstrained Hindi words	ICFHR
2	Singh et al.	2021 – Hindi Character Recognition Review	Survey of OCR techniques	Identified challenges in handwritten Hindi OCR	GCAT
3	Tyagi et al.	2024 – ML for Devanagari Recognition	ML classification models	Improved recognition accuracy using ML	ICCSC
4	Mahajan & Ganpati	2025 – CNN Optimization	Deep learning (CNN)	Enhanced accuracy using optimized CNN models	OTCON
5	Rakib Hasan et al.	2023 – OCR for Nepali & Bengali	Transformer-based OCR	Better context-aware recognition	IEEE

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154273

6	Nguyen Van et al.	2025 – LVM-OCR	Transformer architecture	Improved document understanding	ICETISI
7	Dipu et al.	2021 – Bangla OCR	Deep learning classification	High accuracy in handwritten OCR	ICCIT
8	Rexi & Jacob	2022 – OCR with Segmentation	Projection + DL	Combined segmentation and DL improves results	ICAC3N
9	Kirana et al.	2025 – OCR Comparison	Tesseract, EasyOCR, Transformer	Transformer models outperform others	ICEEIE
10	Sharma et al.	2025 – TrOCR Model	Vision-language model	High OCR accuracy using pre-trained models	IEEE
11	Wang et al.	2024 – CNN + Transformer	Hybrid deep learning	Improved sequence recognition	ICEMCE
12	Li et al.	2022 – TrOCR	Transformer OCR	State-of-the-art OCR performance	arXiv
13	Meoded	2025 – Historical OCR	Transformer-based models	Effective for manuscript recognition	arXiv
14	Pavan Kumar et al.	2011 – Telugu OCR	OCR improvement techniques	Accuracy improvement in Indic OCR	ICDAR
15	Naeem et al.	2017 – Urdu OCR	Ligature-based analysis	Ligature coverage impacts accuracy	ICDAR
16	Lund & Ringger	2011 – OCR Correction	Error correction models	Improved OCR outputs using training	—
17	Saluja et al.	2017 – Indic OCR Correction	Error detection framework	Improved correction in OCR outputs	ICDAR
18	Valizadeh et al.	2025 – OCR + LLM	LLM-based correction	Enhanced post-processing accuracy	CSICC
19	Wemhoener et al.	2013 – OCR Improvement	Multi-edition OCR	Improved noisy OCR results	ICDAR
20	Jeevidha et al.	2026 – OCR + YOLO	OCR + DL + correction	Applied OCR in healthcare systems	ICCIDS

Note: AI = Artificial Intelligence. ML = Machine Learning. DL = Deep Learning. NLP = Natural Language Processing. RF = Random Forest. SVM = Support Vector Machine. KNN = K-Nearest Neighbors. XAI = Explainable Artificial Intelligence. EHR = Electronic Health Records. IVR = Interactive Voice Response.

V. COMPARATIVE SUMMARY OF REVIEWED LITERATURE

Traditional OCR methods are simple and require less computation but perform poorly on complex handwritten text. Machine learning approaches improve accuracy but depend on feature extraction. Deep learning models offer higher accuracy and robustness but require large datasets and computational resources. Transformer-based models achieve the best performance with improved context handling but are computationally expensive. Post-processing methods further enhance results. Overall, there is a trade-off between accuracy, complexity, and resource requirements.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154273

Table II: Comparative Summary of Reviewed Literature (2020–2025)

Sl. No.	Author(s)	Year & Title	Method / Technique	Key Findings	Venue
1	Upreti & Bag	2016 – Segmentation of Handwritten Hindi Words	Polygonal approximation	Effective segmentation of unconstrained Hindi words	ICFHR
2	Singh et al.	2021 – Hindi Character Recognition Review	Survey of OCR techniques	Identified challenges in handwritten Hindi OCR	GCAT
3	Tyagi et al.	2024 – ML for Devanagari Recognition	ML classification models	Improved recognition accuracy using ML	ICCSC
4	Mahajan & Ganpati	2025 – CNN Optimization	Deep learning (CNN)	Enhanced accuracy using optimized CNN models	OTCON
5	Rakib Hasan et al.	2023 – OCR for Nepali & Bengali	Transformer-based OCR	Better context-aware recognition	IEEE
6	Nguyen Van et al.	2025 – LVM-OCR	Transformer architecture	Improved document understanding	ICETISI
7	Dipu et al.	2021 – Bangla OCR	Deep learning classification	High accuracy in handwritten OCR	ICCIT
8	Rexi & Jacob	2022 – OCR with Segmentation	Projection + DL	Combined segmentation and DL improves results	ICAC3N
9	Kirana et al.	2025 – OCR Comparison	Tesseract, EasyOCR, Transformer	Transformer models outperform others	ICEEIE
10	Sharma et al.	2025 – TrOCR Model	Vision-language model	High OCR accuracy using pre-trained models	IEEE
11	Wang et al.	2024 – CNN + Transformer	Hybrid deep learning	Improved sequence recognition	ICEMCE
12	Li et al.	2022 – TrOCR	Transformer OCR	State-of-the-art OCR performance	arXiv
13	Meoded	2025 – Historical OCR	Transformer-based models	Effective for manuscript recognition	arXiv
14	Pavan Kumar et al.	2011 – Telugu OCR	OCR improvement techniques	Accuracy improvement in Indic OCR	ICDAR
15	Naeem et al.	2017 – Urdu OCR	Ligature-based analysis	Ligature coverage impacts accuracy	ICDAR
16	Lund & Ringger	2011 – OCR Correction	Error correction models	Improved OCR outputs using training	—
17	Saluja et al.	2017 – Indic OCR Correction	Error detection framework	Improved correction in OCR outputs	ICDAR
18	Valizadeh et al.	2025 – OCR + LLM	LLM-based correction	Enhanced post-processing accuracy	CSICC
19	Wemhoener et al.	2013 – OCR Improvement	Multi-edition OCR	Improved noisy OCR results	ICDAR
20	Jeevidha et al.	2026 – OCR + YOLO	OCR + DL + correction	Applied OCR in healthcare systems	ICCIDS

V. RESEARCH GAPS AND SYNTHESIS

Current handwritten Hindi OCR systems face several limitations. There is a lack of large and diverse datasets, which affects model performance. Variations in handwriting styles and complex Devanagari structures make recognition difficult. Segmentation of connected characters remains a major challenge. Many systems do not perform well on noisy, real-world data. Additionally, deep learning models require high computational resources, limiting their use in low-resource environments. Overall, there is a need for more robust, scalable, and efficient OCR solutions.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154273

VI. CONCLUSION

This review examined the progress of Optical Character Recognition systems for handwritten Hindi text, covering traditional methods, machine learning approaches, and recent deep learning techniques. It is evident that deep learning models, particularly CNN and hybrid architectures, have significantly improved recognition accuracy compared to earlier methods.

However, challenges such as variability in handwriting, segmentation difficulties, limited datasets, and high computational requirements still affect system performance. Most existing solutions also struggle with real-world deployment due to noise and resource constraints.

In conclusion, future work should focus on developing robust, scalable, and efficient OCR systems that can handle diverse handwriting styles and operate effectively in practical environments. Integrating advanced models with lightweight architectures and better datasets will be key to improving handwritten Hindi text recognition.

REFERENCES

[1]. K. K. Upreti and S. Bag, "Segmentation of unconstrained handwritten Hindi words using polygonal approximation," in Proc. 15th Int. Conf. Frontiers in Handwriting Recognition (ICFHR), 2016.
[2]. A. K. Singh, B. Kadhiwala and R. Patel, "Hand-written Hindi character recognition – A comprehensive review," in Proc. 2nd Global Conf. Advancement in Technology (GCAT), 2021.
[3]. S. Tyagi, C. Dutta and M. Singh, "Machine learning for recognition of handwritten Devanagari character," in Proc. Int. Conf. Computing, Sciences and Communications (ICCSC), 2024.
[4]. B. Mahajan and A. Ganpati, "Optimizing deep learning based CNNs for effective handwritten character image recognition of Devanagari script using R language," in Proc. 4th OPJU Int. Technology Conf. (OTCON), 2025.
[5]. S. M. Rakib Hasan, M. H. K. Mehedi, A. Dhakal and A. A. Rasel, "Optical text recognition in Nepali and Bengali: A transformer-based approach," in Proc. IEEE 5th Int. Conf. Advances in Electronics, Computers and Communications (ICAECC), 2023.
[6]. B. Nguyen Van, N. Dinh, V. T. Hoang, H. H. Thien and K. T. Trung, "LVM-OCR: A transformer-based architecture for context-aware document understanding," in Proc. 1st Int. Conf. Emerging Trends in Information Systems and Informatics (ICETISI), 2025.
[7]. N. M. Dipu, S. A. Shohan and K. M. A. Salam, "Bangla optical character recognition (OCR) using deep learning based image classification algorithms," in Proc. 24th Int. Conf. Computer and Information Technology (ICCIT), 2021.
[8]. A. Rexi F and L. Jacob, "Optical character recognition system with projection profile based segmentation and deep learning techniques," in Proc. 4th Int. Conf. Advances in Computing, Communication Control and Networking (ICAC3N), 2022.
[9]. K. C. Kirana, A. Maqbullah, I. Kumalasari, A. F. Shobari and B. Hidayat, "Comparison of Tesseract OCR, EasyOCR, and transformer OCR on handwritten image," in Proc. 9th Int. Conf. Electrical, Electronics and Information Engineering (ICEEIE), 2025.
[10]. M. Sharma, S. Agarwal, R. K. Saxena, A. Saxena and A. Runthala, "TrOCR: Transformer-based OCR with pre-trained vision-language models," in Proc. IEEE Pune Section Int. Conf. (PuneCon), 2025.
[11]. Y. Wang, K. Chen, T. Yang, H. Cao and D. Zhao, "Handwritten English recognition based on CNN and CNN-head transformer," in Proc. 8th Int. Conf. Electrical, Mechanical and Computer Engineering (ICEMCE), 2024.
[12]. M. Li et al., "TrOCR: Transformer-based optical character recognition with pre-trained models," arXiv preprint arXiv:2109.10282, 2022.
[13]. E. Meoded, "Handwritten text recognition of historical manuscripts using transformer-based models," arXiv preprint, 2025.
[14]. P. Pavan Kumar, C. Bhagvati, A. Negi, A. Agarwal and B. L. Deekshatulu, "Towards improving the accuracy of Telugu OCR systems," in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2011.
[15]. M. F. Naeem et al., "Impact of ligature coverage on training practical Urdu OCR systems," in Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR), 2017.
[16]. W. B. Lund and E. K. Ringger, "Error correction with in-domain training across multiple OCR system outputs," 2011.
[17]. R. Saluja et al., "A framework for document specific error detection and corrections in Indic OCR," in Proc. 14th IAPR Int. Conf. Document Analysis and Recognition (ICDAR), 2017.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154273

[18]. F. Valizadeh et al., "Comparative analysis of large language models for OCR post-processing in Persian," in Proc. 29th Int. Computer Conf. of Iran (CSICC), 2025.
[19]. D. Wemhoener, I. Z. Yalniz and R. Manmatha, "Creating an improved version using noisy OCR from multiple editions," in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2013.
[20]. S. Jeevidha, A. Prem Anand, P. Yogeshwaran and K. Kannathasan, "An advanced electronic health record for prescription analysis using YOLO and OCR with spell correction," in Proc. 9th Int. Conf. Computational Intelligence in Data Science (ICCIDS), 2026.

