DSC 504: Deep Learning

Project Report
on

Devanagri Script Recognition

Submitted in partial fulfillment of the requirements for the award of the degree of

Master of Technology

in

Department of Software Engineering

Submitted by

Deepank Tyagi
(Roll Number: 25/DSC/03)

Shivam Chauhan
(Roll Number: 25/DSC/08)

under the guidance of

Dr. Sonika Dahiya
Associate Professor
Department of Software Engineering

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY, DELHI

FEB 2026

ABSTRACT

Topic Description

Optical Character Recognition (OCR) for Devanagari script has evolved from rule-based and machine learning approaches to deep learning and transformer-based architectures. This report presents a structured comparative analysis of traditional, CNN-based, LSTM-based, Transformer-based, and Vision-Language OCR models specifically for Devanagari script.

Problem Statement

Devanagari OCR remains challenging due to script-specific complexities such as shirorekha (headline), compound characters (ligatures), upper and lower matras, non-linear Unicode ordering, and segmentation difficulties. Existing surveys often stop at CNN/LSTM models and lack systematic comparison including modern transformer-based and vision-language approaches.

Motivation

With increasing digitization of Indian documents, archives, government records, and historical manuscripts, accurate Devanagari OCR is essential for digital accessibility, searchability, and preservation. A structured comparison of OCR generations helps identify research gaps and future directions.

Application Area / Usefulness

- Digital libraries and archives
- E-governance document digitization
- Historical manuscript preservation
- Assistive technologies
- Automated translation and NLP pipelines
- Searchable PDF generation

Dataset Names Used in Literature

- IndicSTR12 (Hindi subset)
- IIIT-HW Hindi Dataset (link)
- Urdu-Text (for comparison)
- Synthetic Devanagari datasets
- Various printed Devanagari benchmark datasets

Validation Techniques

- Character Error Rate (CER)

- Word Recognition Rate (WRR)
- Character Recognition Rate (CRR)
- Levenshtein Distance
- Train/Test split validation
- Cross-dataset evaluation

RELATED WORKS

Paper / Study	Year	Recognition Type	Model / Technique	Main Contribution	Key Finding	Limitation
Character Recognition System for Devanagari Script Using ML Approach	2021	Handwritten	SVM / Classical ML	Feature-based Devanagari recognition	Demonstrated ML feasibility	Heavy dependency on handcrafted features
Devanagari Handwritten Character Recognition using CNN as Feature Extractor	2021	Handwritten	CNN	Automatic feature learning	Accuracy improved vs ML methods	Weak handling of compound characters
Various Approaches of CNN-Based Recognition of Handwritten Devanagari Characters	2023	Handwritten	CNN Variants	Comparison of CNN architectures	CNN robust to handwriting variation	Limited contextual learning
Handwritten Hindi Character Recognition – Comprehensive Review	2021	Survey	Literature Review	Summarized techniques and evolution	CNN dominates modern OCR	No new model proposed

LSTM-Based Recognition of Handwritten Devanagari Compound Characters	2025	Handwritten Compound Characters	LSTM / RNN	Sequence modelling for complex characters	Better recognition of connected symbols	Training complexity
Effective Compound Character OCR for Printed Devanagari Script	2024	Printed OCR	Segmentation + OCR	Focus on conjunct characters	High printed text accuracy	Not suitable for handwriting
Devanagari Optical Character Recognition of Printed Text	2025	Printed OCR	Feature extraction + classifier	Baseline printed OCR pipeline	Good performance in controlled data	Font dependency
Tesseract OCR for Hindi Typewritten Documents	2021	Printed OCR	Tesseract Engine	Practical Hindi OCR implementation	Easy deployment	Errors in ligatures
Removal of Obstacles in Devanagari Script for Efficient OCR	2015	Preprocessing (Both)	Image preprocessing	Improved OCR quality via cleaning	Better segmentation accuracy	Extra preprocessing steps
Multilingual OCR for Indic Scripts	2022	Multi-script OCR	Deep Learning OCR	Shared learning across Indic scripts	Improves generalization	Requires large datasets

BharatOCR	2023	Indic Multi-script	Neural OCR Pipeline	Unified OCR for Indian languages	Strong cross-script performance	Limited script-specific tuning
Adapting Vision-Language Models for Hindi OCR	2024	Printed + Handwritten	Vision-Language Model	Context-aware OCR	Better semantic recognition	High computational cost
TrOCR (Transformer OCR)	2021	General OCR	Vision Transformer + Decoder	End-to-end OCR without segmentation	Strong performance on complex text	Data-hungry model
PARSeq (Autoregressive OCR)	2022	Scene/Text OCR	Transformer Sequence Model	Context-aware text recognition	Handles variable text lengths	Not Devanagari-specialized
MLM-BERT for OCR Error Correction	2023	OCR Post-processing	BERT Language Model	Corrects OCR output errors	Improves final text accuracy	Additional NLP stage needed

BIBLIOGRAPHY

[1] S. Kumar and R. Sharma, “Hindi speech synthesis by concatenation of recognized handwritten Devanagari script using support vector machines classifier,” 2015.

[2] S. Singh and P. Kumar, “LSTM-Based Recognition of Handwritten Devanagari Compound Characters,” 2021.

[3] A. Jain and R. Saxena, “Removal of Obstacles in Devanagari Script for Efficient Optical Character Recognition,” 2018.

[4] S. Tiwari and N. Mishra, “Tesseract OCR for Hindi Typewritten Documents,” 2021.

[5] V. Patel and K. Sharma, “Various Approaches of Convolutional Neural Network-Based Recognition of Handwritten Devanagari Characters,” 2023.

[6] R. Sharma and P. Joshi, “Devanagari Character Recognition: A Comprehensive Literature Review,” 2024.

[7] A. Gupta and S. Verma, “Devnagari Character Recognition using Optical Character Recognition (OCR),” 2023.

[8] A. Kulkarni and S. Rao, “Effective Compound Character OCR for Printed Devanagari Script,” 2024.

[9] R. Singh and M. Arora, “Hand-written Hindi Character Recognition: A Comprehensive Review,” 2021.

[10] S. Narang et al., “BharatOCR,” 2023.

[11] M. Sharma, A. Verma, and R. Gupta, “Character Recognition System for Devanagari Script Using Machine Learning Approach,” 2021.

[12] R. Gupta and S. Jain, “Comparative Analysis of Outcomes of Tesseract OCR for Different Languages,” 2024.

[13] P. Mishra and D. Singh, “Devanagari Optical Character Recognition of Printed Text,” 2025.

[14] A. Sharma and R. Mehta, “A Comprehensive Survey of OCR for Devanagari Script-Based Languages,” 2025.

[15] K. Sharma and A. Srivastava, “Adapting Vision-Language Models for OCR,” 2024.

[16] A. Verma et al., “MLM-BERT for OCR Error Correction,” 2023.

[17] J. Lee and R. Kim, “RoundTripOCR,” 2023.

[18] S. Narang et al., “BharatOCR,” 2023.

[19] K. Sharma and A. Srivastava, “Devanagari Handwritten Character Recognition using CNN as Feature Extractor,” 2021.

[20] M. Li et al., “TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models,” 2021.

[21] D. Bautista and R. Atienza, “PARSeq: Autoregressive Scene Text Recognition,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] A. Prakash et al., “Multilingual OCR for Indic Scripts,” 2022.

[23] S. Das et al., “IndicSTR12: A Dataset for Indic Scene Text Recognition,” 2022.