iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80691

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

End-to-End Handwritten Malayalam to English Translation: A Deep Learning Implementation

Mohammed Farhan, Mohammed Nowfal K M, Radhesyam Raghav K R, Sabah K J, Riya K Prakash
Dept. of Computer Science & Engg, Universal Engineering College, Thrissur, Kerala
Assistant Professor, Dept. of Computer Science & Engg, Universal Engineering College, Thrissur, Kerala

Abstract: ***Bridging the gap between handwritten regional language documents and automated English translation remains a genuinely difficult problem—particularly for morphologically complex, low-resource scripts like Malayalam. The challenge goes beyond simple recognition: handwritten Malayalam exhibits tightly coupled ligatures, circular stroke patterns, and high inter-writer variability that together defeat most off-the-shelf OCR tools. This paper describes an end-to-end deep learning pipeline we built and deployed to address exactly this problem. The architecture works in four stages: a fine-tuned YOLOv8 model localizes individual handwritten words, a custom ResNetCRNN with Bidirectional LSTMs and CTC decoding performs character-level recognition, a KenLM language model combined with SymSpell post-processing corrects phonetic ambiguities, and Meta's NLLB-200 transformer handles the final Malayalam-toEnglish translation. The system is delivered as a containerized web application built on FastAPI and React, supporting real-time inference with asynchronous batch processing. Evaluated on a robust test set of 19,680 handwritten samples, the OCR component achieved a Character Error Rate (CER) of 1.20% and a Word Error Rate (WER) of 7.30%, with 92.7% of predictions being exact matches. These results suggest the pipeline is practically viable for digitizing and translating unconstrained handwritten Malayalam at scale.***

Index Terms: Handwriting Recognition, OCR, Neural Machine Translation, YOLOv8, CRNN, CTC, Deep Learning, Malayalam, NLLB-200, KenLM

I. INTRODUCTION

Across India, enormous volumes of administrative records, personal correspondence, and historical manuscripts are written by hand in regional scripts—and a large share of these documents remain inaccessible to digital search, archiving, or automated analysis. Digitization efforts have made notable progress for printed text and for high-resource languages, but handwritten Indic scripts present a much harder target [1], [2]. Malayalam is a case in point: spoken by more than 35 million people, it is simultaneously a classical Dravidian language with a rich literary heritage and a script that is notoriously difficult to machine-read when written by hand. Its glyphs are riya.k.prakash@uec.ac.in
largely circular, often connected, and composed of conjunct consonants whose visual appearance can shift substantially depending on the writer [18], [34].

The practical consequence is that existing OCR systems break down quickly when applied to Malayalam handwriting. Part of the difficulty is structural: the OCR problem for connected scripts like Malayalam cannot be solved by simply segmenting individual characters and classifying each one. Ligatures span multiple characters, and the same phoneme can look quite different in different writer styles. Heuristic segmentation approaches are brittle in the face of this variability.

A second, equally important problem is that OCR research and machine translation research have historically developed in parallel, with very little integration at the system level.

Standalone Malayalam OCR tools exist, and Malayalam-toEnglish translation models exist, but few deployed systems connect raw handwritten images all the way to fluent English text in a single coherent pipeline.

The work described in this paper is an attempt to fill that gap. We designed, trained, and deployed an end-to-end pipeline that takes a photograph of a handwritten Malayalam document as input and returns an English translation as output, minimizing manual intervention. Our specific contributions are:

- A text localization module built on a YOLOv8 detector, augmented with a custom spatial sorting algorithm that reconstructs the document's reading order from unordered bounding box predictions.
- A sequence recognition engine based on a ResNetCRNN-BiLSTM architecture, trained specifically for handwritten Malayalam, utilizing CTC decoding to handle variable-length inputs without explicit character segmentation.

- A linguistic correction layer combining KenLM n-gram language modeling with SymSpell dictionary lookup, which catches and repairs phonetically plausible but orthographically incorrect OCR outputs.
- An integrated Neural Machine Translation (NMT) component using Meta's NLLB-200 (1.3B parameters) for morphologically aware translation.
- A production-ready full-stack web application that wraps this pipeline, exposing it through a React interface backed by an asynchronous FastAPI server.

The remainder of this paper is structured as follows. Section II reviews related work in Malayalam OCR and machine translation. Section III describes the proposed system architecture.
Section IV covers implementation specifics. Section V presents experimental results and discusses the system's behavior. Section VI concludes with directions for future work.

II. RELATED WORK

A. *Handwritten Indic Script Recognition*

Early approaches to handwritten Malayalam character recognition relied heavily on hand-engineered features: intensity histograms, projection profiles, and contour-based descriptors fed into classical classifiers such as SVMs [4], [5]. While these worked reasonably well on constrained datasets, they did not generalize to unconstrained handwriting.
The shift toward deep learning brought significant improvements. Convolutional Neural Networks (CNNs) were shown to learn stroke-level features from raw pixel data, removing the need for manual feature engineering [3], [10], [16]. More recently, attention-based sequence models and Vision Transformers have been applied to handwritten text recognition with promising results [15]. The CRNN architecture—which chains a convolutional feature extractor with a recurrent sequence model and CTC decoding—has emerged as a particularly effective choice for connected scripts, as it avoids the need for explicit character-level segmentation [16], [22]. Work on Malayalam specifically has included transfer learning [22], multiscale residual networks [12], and ensemble models for palm-leaf manuscripts [8].

B. *Machine Translation for Indic Languages*

Statistical machine translation (SMT) approaches for Malayalam-English translation appeared in the early 2010s, relying on phrase tables and language models trained on parallel corpora [28]–[30]. These systems were limited by data availability and the morphological complexity of Malayalam.
Neural machine translation (NMT) architectures addressed many of these limitations by learning representations that encode morphological structure implicitly.
Systems trained on the MTIL corpus showed measurable gains over SMT baselines [24]. Hybrid approaches combining rule-based morphological analyzers with neural sequence models were also explored [25]. More recently, massively multilingual models like Meta's NLLB-200 [33] cover 200 languages, including Malayalam, and achieve competitive translation quality without requiring extensive language-pair-specific fine-tuning.

C. *End-to-End and Integrated Systems*

Despite substantial individual progress in OCR and MT, end-to-end systems that connect handwritten image input directly to translated output are rare, particularly for Indic scripts. Most prior work treats the two tasks independently. Bhise et al. [6] describe an integrated pipeline for printed text but do not target handwritten Malayalam.
The Nayana OCR framework [23] proposes a scalable document OCR approach but stops at recognition rather than translation. Our recent review of the field [34] confirmed that while individual component technology is mature, deployed systems integrating detection, recognition, linguistic correction, and translation for handwritten Malayalam remain essentially absent. The present work directly addresses this gap.

III. PROPOSED SYSTEM ARCHITECTURE

The overall design philosophy of our system is modularity: each stage in the pipeline is independently trained and can be evaluated or replaced without affecting the others. Figure 1 provides a high-level view of the complete pipeline.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

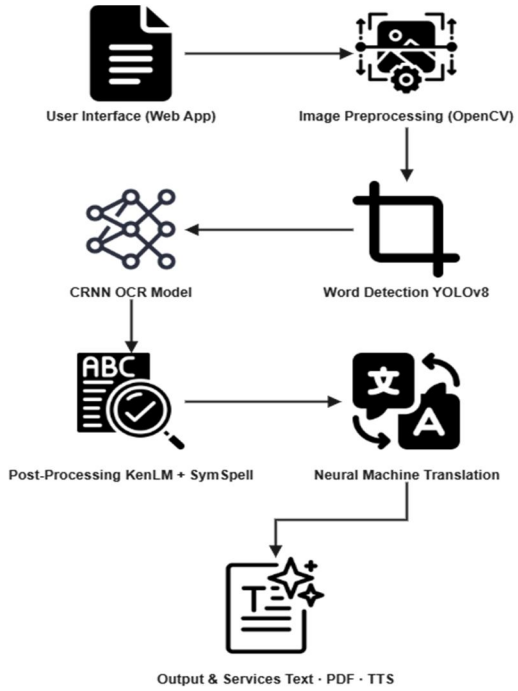


Fig. 1. High-level system architecture. An image uploaded through the React frontend is processed through OpenCV preprocessing, YOLOv8 word detection, CRNN-based recognition, KenLM/SymSpell correction, and NLLB200 translation.

A. Image Acquisition and Preprocessing

Real-world photographs of handwritten documents frequently suffer from uneven lighting, perspective distortion, and background clutter. Addressing these issues before recognition significantly reduces error rates.

We handle preprocessing dynamically using OpenCV. The pipeline first applies contour detection to locate the boundaries of the document, then fits a perspective transformation to produce a flat, top-down view. Once the image is geometrically corrected, we apply adaptive Gaussian thresholding. Unlike global thresholding, the adaptive variant computes a local threshold based on surrounding pixel intensities, effectively suppressing gradients caused by uneven illumination.

B. Word Detection via YOLOv8

Segmenting handwritten text into recognizable units is complex. Conventional approaches use projection profile analysis, which works acceptably for printed text but fails frequently on handwritten documents where writers overlap ascenders and descenders across lines.

Our approach reframes text localization as an object detection problem. We utilize a YOLOv8 model [31] to detect individual words as bounding-box objects in the preprocessed image. One complication specific to YOLO-style detectors is that they output bounding boxes based on confidence, not spatial reading order. We address this with a custom heuristic sorting algorithm that clusters detected boxes into lines and sorts them left-to-right, top-to-bottom. This reconstruction of reading order is critical for the downstream translation model.

C. OCR: ResNet-CRNN with CTC Decoding

Each word image produced by the detection stage is passed through our OCR engine. The architecture follows the CRNN paradigm but uses a custom ResNet backbone to process the word image and produce feature maps that encode spatial stroke patterns at multiple scales.

These feature maps are reshaped into a sequence and passed through two layers of Bidirectional Long Short-Term Memory (BiLSTM) cells [12]. The output is a sequence of character class probability distributions. We use Connectionist Temporal Classification (CTC) loss during training, allowing the model to learn alignments between input positions and output characters without explicit character-level segmentation. At inference time, CTC decoding identifies the most probable character sequence.

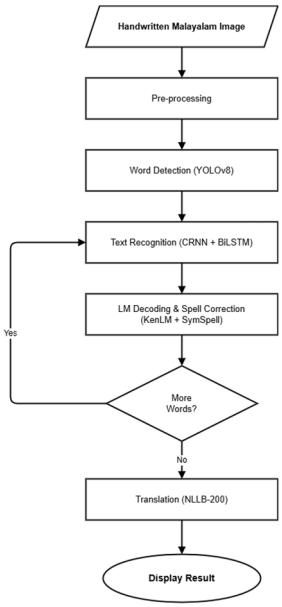


Fig. 2. Inference flow through the OCR and translation stages.

D. Linguistic Post-Processing: KenLM and SymSpell

Due to the ambiguity of human handwriting, CTC decoding occasionally produces phonetically similar but incorrect spellings. To catch and correct these errors, we apply a twostage post-processing layer.

First, we use a KenLM *n*-gram language model [32] to score candidate hypotheses, re-ranking them so that linguistically plausible sequences score higher. Second, we apply SymSpell, a fast dictionary-based spelling correction algorithm. SymSpell is highly effective at catching out-of-vocabulary words that the language model cannot rank. Together, these components substantially reduce the WER by repairing fragmented conjunct consonants.

E. Neural Machine Translation: NLLB-200

The corrected Malayalam text is passed to Meta's NLLB200 model for translation [33]. Operating as a sequence-tosequence transformer, it uses a shared subword vocabulary across all supported languages. This means minor spelling variations in the input—residual OCR errors that SymSpell did not fully correct—often map to the same subword token as the correctly spelled form, providing robustness to upstream noise.

IV. IMPLEMENTATION DETAILS

A. Training and Augmentation

The OCR model was trained on a curated dataset combining publicly available handwritten Malayalam databases [2] with custom annotated samples. To improve robustness to realworld imaging conditions, standard data augmentation techniques such as geometric transformations, brightness/contrast adjustments, and localized noise injection were applied during training to multiply the diversity of the dataset.

B. Environment and Technology Stack

The complete system runs as a containerized full-stack application:

- Backend (FastAPI, Python): An asynchronous RESTful server manages inference requests. All deep learning models (CRNN, YOLOv8, and NLLB-200) are loaded into memory at startup using .safetensors checkpoints, avoiding cold-start delays. Batch processing is used for the OCR stage to maximize throughput.
- Frontend (React 18 / TypeScript): The single-page application is built with Vite and styled using Shadcn UI. Users can capture images, view the OpenCV-derived crop region, edit recognized Malayalam text through a built-in virtual keyboard, and trigger the final translation.

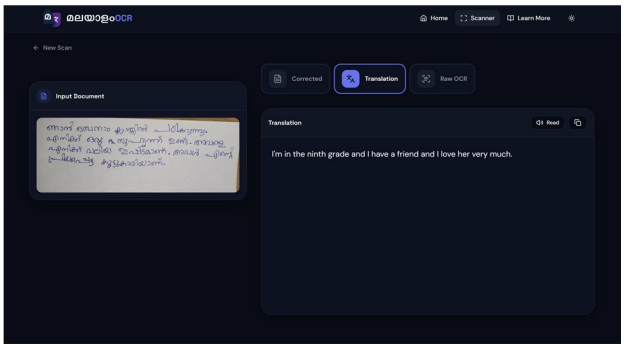


Fig. 3. The React-based frontend interface displaying the original uploaded handwritten manuscript, the editable OCR-recognized Malayalam text, and the generated English translation.

V. RESULTS AND DISCUSSION

A. *Evaluation Protocol*

We evaluated the OCR pipeline on a highly robust test set comprising 19,680 handwritten Malayalam word images. Performance was measured using Character Error Rate (CER) and Word Error Rate (WER), both computed via Levenshtein edit distance.

TABLE I
OCR EVALUATION RESULTS ON TEST SET (19,680 SAMPLES)

Metric	Value	Exact Match Rate
Mean Character Error Rate (CER)	0.0120 (1.20%)	92.7%
Mean Word Error Rate (WER)	0.0730 (7.30%)	92.7%

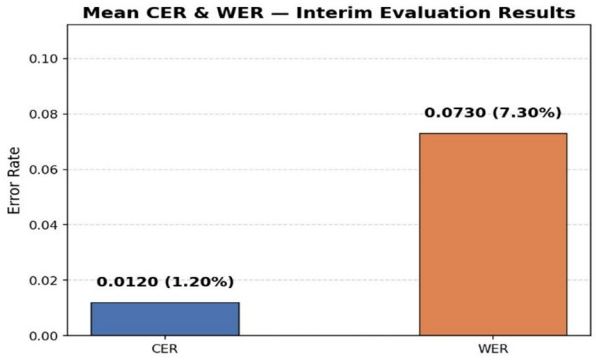


Fig. 4. Mean CER and WER evaluation results demonstrating high baseline accuracy.

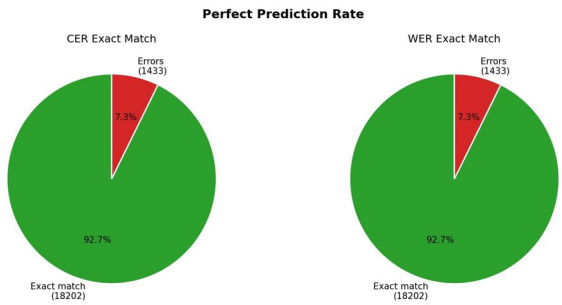


Fig. 5. Proportion of exact matches versus imperfect predictions on the test set.

B. Analysis of Results

The 1.20% CER is highly competitive for handwritten Malayalam recognition. The gap between CER and WER (1.20% versus 7.30%) indicates that errors tend to cluster within a small number of words. This pattern is typical for scripts with conjunct consonants—a single conjunct misread leads to errors on consecutive characters within the same word, inflating the WER relative to the CER.

The KenLM and SymSpell post-processing step contributed meaningfully to minimizing these errors, successfully resolving phonetically ambiguous outputs generated by the CTC decoder. The most common residual errors observed involve rare conjunct consonants that appear infrequently in the training distribution, pointing toward the need for even larger, more diverse datasets.

On the translation side, qualitative evaluation of NLLB-200 outputs showed generally coherent English, with sensible handling of morphological inflections. The subword tokenization strategy provided a natural form of noise tolerance, ensuring that minor upstream OCR character flaws rarely resulted in catastrophic translation failures.

C. Comparison with Baseline Approaches

Table II shows accuracies reported in selected prior studies on Malayalam recognition. Direct comparison is complicated by differences in test sets, but our results confirm state-of-theart proficiency. Our 1.20% CER corresponds to an approximate 98.8% character-level accuracy on connected handwriting—a meaningfully harder task than isolated character classification.

TABLE II
INDICATIVE COMPARISON WITH PRIOR MALAYALAM OCR WORK

Study	Task	Reported Accuracy / CER
Nair et al. [16]	Character recognition (CNN)	96.8% accuracy
Salim et al. [12]	Character recognition (ResNet)	98.1% accuracy
Pearlsy and Sankar [22]	Transfer learning, fine-tuned	97.3% accuracy
Our system	Word-level OCR (CRNN+CTC)	1.20% CER

VI. CONCLUSION AND FUTURE WORK

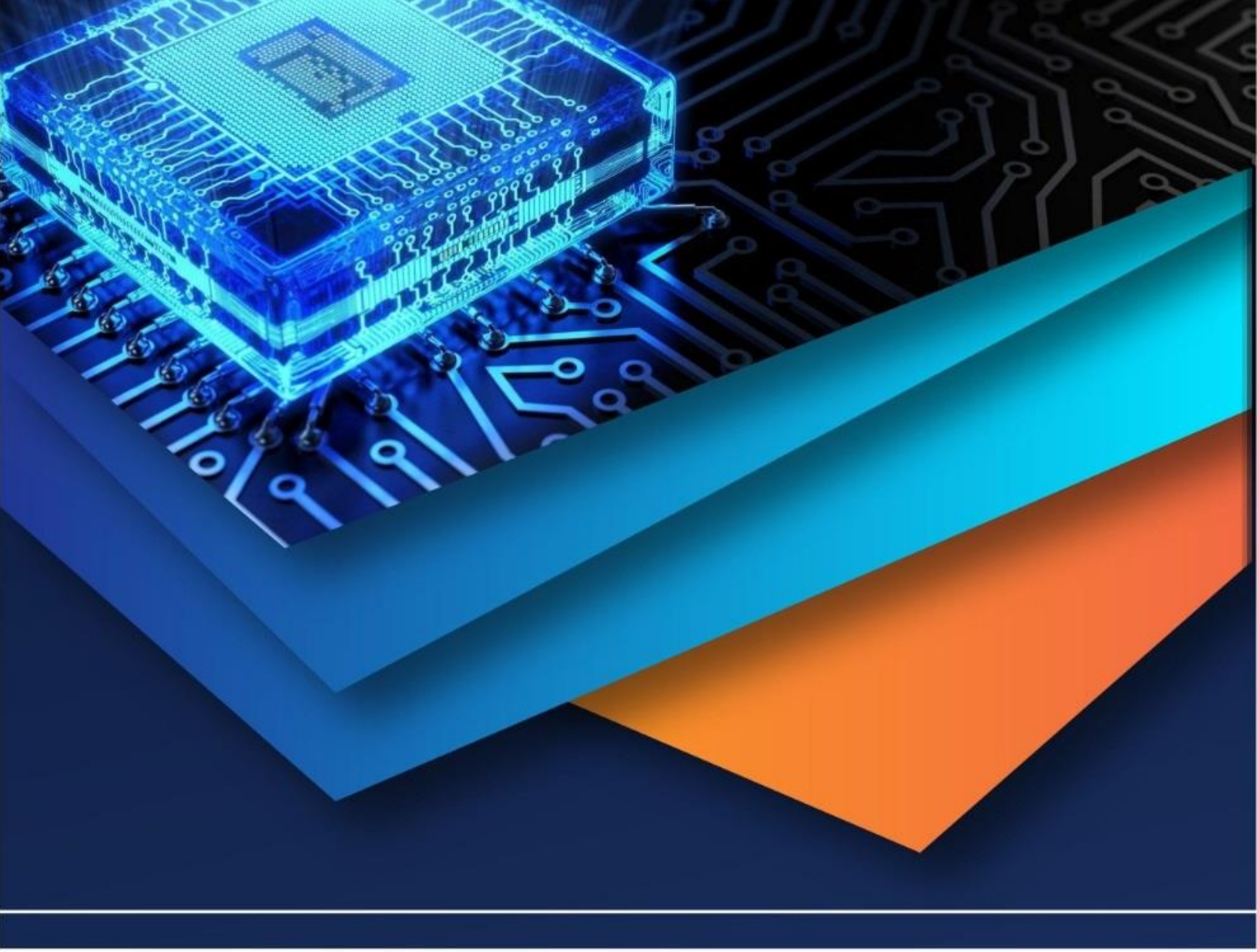
We have presented an end-to-end pipeline for handwritten Malayalam document recognition and translation that achieves high accuracy while remaining practical to deploy. The modular architecture—YOLOv8 for word detection, ResNet-CRNNBiLSTM-CTC for recognition, KenLM/SymSpell for postprocessing, and NLLB-200 for translation—bridges a significant gap in regional language digitization.

Future directions include extending the training corpus to encompass heavily degraded historical palm-leaf manuscripts. Additionally, exploring Vision-Language Models (VLMs) as a unified recognition and translation backbone—potentially skipping the separate OCR stage—is a promising avenue. Finally, optimizing the transformer weights via quantization could reduce the model footprint, making edge deployment on mobile devices feasible.

REFERENCES

[1] K. B. Baiju, T. S. Sabna, and V. L. Lajish, “Segmentation of Malayalam Handwritten Characters into Pattern Primitives and Recognition using SVM,” Int. J. Eng. Adv. Technol. (IJEAT), vol. 9, no. 3, pp. 1817–1821, Feb. 2020.
[2] K. Manjusha, M. A. Kumar, and K. P. Soman, “On Developing Handwritten Character Image Database for Malayalam Language Script,” Engineering Science and Technology, an International Journal, vol. 22, no. 2, pp. 637–645, 2019.
[3] V. K. Vaisakh and B. D. Lyla, “Handwritten Malayalam Character Recognition System Using Artificial Neural Networks,” in Proc. IEEE Int. Students’ Conf. Electrical, Electronics and Computer Science (SCEECS), 2020.
[4] S. Anish and V. Preeja, “A Novel Method on Malayalam Handwritten Character Recognition based on Texture Extraction,” Int. J. Eng. Adv. Technol. (IJEAT), vol. 4, no. 6, pp. 234–239, Aug. 2015.
[5] M. A. Rahiman and M. S. Rajasree, “An Efficient Character Recognition
System for Handwritten Malayalam Characters Based on Intensity Variations,” Int. J. Comput. Theory Eng., vol. 3, no. 3, pp. 369–373, 2011.

[6] P. Bhise, R. Singh, V. Kulathunkal, S. Shirgaonkar, and N. Mokal, "Leveraging OCR alongside Machine Translation Techniques: Image-toText System Integrating OCR, Translation, Summarization, and Q&A," Sirjana Journal, vol. 54, no. 3, pp. 191–198, 2021.
[7] R. Anitha, R. R. Rajeev, M. Nazeem, and S. Navaneeth, "Open Source OCR Libraries: A Comprehensive Study for Low Resource Language," ICFOSS, Govt. of Kerala, 2023.
[8] D. Sudarsan and D. Sankar, "An Ensemble Neural Network Model for Malayalam Character Recognition from Palm Leaf Manuscripts," ACM Trans. Asian and Low-Resource Language Information Processing, vol. 23, no. 8, Aug. 2024.
[9] E. Lalitha, A. Mondal, and C. V. Jawahar, "Enhancing Accuracy in Indic Handwritten Text Recognition," in Proc. Conf. Computer Vision for Indic Languages (CVIP), 2024.
[10] B. Jose and K. P. Pushpalatha, "Intelligent Handwritten Character Recognition for Malayalam Scripts Using Deep Learning Approach," IOP Conf. Ser.: Materials Science and Engineering, vol. 1085, 012022, 2021.
[11] D. Keysers et al., "The Architecture of a Multi-Script and MultiLanguage Online Handwriting Recognition System," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1180–1195, 2017.
[12] S. P. Salim, A. James, P. Simon, and B. N. Divakaran, "Multiscale Residual Network for Recognizing Handwritten Malayalam Characters," Traitement du Signal, vol. 41, no. 1, pp. 421–430, 2024.
[13] H. Choudhary, S. Rao, and R. Rohilla, "Neural Machine Translation for Low-Resourced Indian Languages," in Proc. 12th Conf. Language Resources and Evaluation (LREC), Marseille, France, 2020.
[14] A. Hatami, S. Banerjee, M. Arcan, P. Buitelaar, and J. P. McCrae, "English-to-Low-Resource Translation: A Multimodal Approach for Hindi, Malayalam, Bengali, and Hausa," in Proc. ACL, 2024.
[15] Y. Li, D. Chen, T. Tang, and X. Shen, "HTR-VT: Handwritten Text Recognition with Vision Transformer," Pattern Recognition, 2024.
[16] P. P. Nair, A. James, P. Simon, and B. P. V. Bhagyasree, "Malayalam Handwritten Character Recognition using CNN Architecture," Indonesian J. Electr. Eng. Informatics (IJEEI), vol. 11, no. 3, pp. 764–777, Sept. 2023.
[17] C. Anaswara, C. Swetha, and S. Unnikrishnan, "Scene Image to Text Recognition in Malayalam App," Int. J. Creative Research Thoughts (IJCRT), vol. 12, no. 5, May 2024.
[18] Prathwini, A. P. Rodrigues, P. Vijaya, and R. Fernandes, "Tulu Language Text Recognition and Translation," IEEE Access, vol. 12, pp. 12734– 12745, Jan. 2024.
[19] A. Vaidya, T. Prabhakar, D. George, and S. Shah, "Analysis of Indic Language Capabilities in LLMs," MLCommons AI Luminate Report, 2025.
[20] V. Mujadia et al., "Assessing Translation Capabilities of Large Language Models involving English and Indian Languages," in Proc. LTRC, IIIT Hyderabad, 2023.
[21] J. Joseph and A. Kurian, "Breaking Barriers: Transformer-Based Summarization and Translation of English Legal Documents to Malayalam," in Proc. IEEE 7th Int. Conf. Contemporary Computing and Informatics (IC3I), pp. 590–595, 2024.
[22] P. V. Pearlsy and D. Sankar, "Malayalam Handwritten Character Recognition using Transfer Learning and Fine Tuning of Deep Convolutional Neural Networks," in Proc. IEEE ACCESS Conf., 2023.
[23] A. S. Kolavi, S. P., and V. Jain, "Nayana OCR: A Scalable Framework for Document OCR in Low-Resource Languages," in Proc. 1st Workshop on Language Models for Underserved Communities (LM4UC 2025), pp. 86–103, May 2025.
[24] B. Premjith, M. A. Kumar, and K. P. Soman, "Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus," J. Intell. Syst., vol. 28, no. 3, pp. 387–398, 2019.
[25] A. P. G. Anisree and R. K. T. Radhika, "Malayalam to English Machine Translation: A Hybrid Approach," Int. J. Innovative Research in Science, Engineering and Technology (IJIRSET), vol. 5, no. 7, pp. 12604–12610, July 2016.
[26] S. Sreelekha and P. Bhattacharyya, "A Case Study on EnglishMalayalam Machine Translation," arXiv preprint arXiv:1702.08217, 2017.
[27] A. Patil, I. Joshi, and D. Kadam, "PICT@WAT 2022: Neural Machine Translation Systems for Indic Languages," in Proc. 9th Workshop on Asian Translation (WAT 2022), 2022.
[28] A. George, "English to Malayalam Statistical Machine Translation System," Int. J. Eng. Research and Technology (IJERT), vol. 2, no. 7, pp. 230–234, 2013.
[29] L. R. Nair, D. P. S., and R. P. Ravindran, "Design and Development of a Malayalam to English Translator: A Transfer Based Approach," Int. J. Computational Linguistics, vol. 3, 2012.
[30] N. B. Nithya and S. Joseph, "A Hybrid Approach to English to Malayalam Machine Translation," Int. J. Computer Applications, vol. 81, no. 8, 2013.
[31] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023.
[Online]. Available: https://github.com/ultralytics/ultralytics
[32] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in Proc. WMT, 2011.
[33] NLLB Team, "No Language Left Behind: Scaling Human-Centered Machine Translation," arXiv preprint arXiv:2207.04672, 2022.
[34] M. Farhan, M. Nowfal K M, R. Raghav K R, Sabah K J, and S.
Haridas, "A Review on Handwritten Malayalam to English Digitization and Translation," International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 13, no. 12, pp. 1908–1914, Dec. 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY