Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

OCR System for Translation and Mathematical Expression Detection in Handwritten Documents

Ms. M.V. Bhuvaneswari*, K. Pravallika†, M. Jayanth Vinay†, S. Ankitaa†, V. Navin Kumar†
Department of CSE (AI & ML), Anil Neerukonda Institute of Technology and Sciences
Sangivalasa, India

***Abstract*—The physical or scanned documents are often the only ones that can be accessed, searched and preserved in handwritten and printed documents. Traditional OCRs concentrate more on the text recognition part, and they are incompetent to locate cumbersome handwritten mathematical statements, as well as perform contextual text recognition [1]. In this project, document digitization and translation system using OCR have been offered that would be dependable when a single source language is used. It is a system, which utilizes a Neural Machine Translation (NMT) block, a mathematical expression Functions recognition, and machine-learned OCR within a single system. Even though encoder-decoder system may be effective in deciphering mathematical formula, a lightweight CNN-SVM hybrid system is effective in character recognition. Other pre-processing steps that are used in the system to increase the accuracy of the recognition of cursive and stylized handwritings include normalizing symmetry. The combination with the OCR and mathematical expression processing and the real-time translation enable it to behave in such a way [4].**

I. INTRODUCTION

Optical Character Recognition (OCR) is used to make the conversion a possibility- Sumatra prints or Inscription of content into machine-readable, digital format, which enhances effective storage, searching and retrieving of documents [4]. Nevertheless, the latest developments in the sphere of deep learning have greatly increased the quality of OCR, most of which are currently available are plain text extraction and lack integrated translation and handwritten mathematical expression recognition, thereby reducing their effectiveness in academic and multilingual contexts.This paper proposes an OCR-based document digitization system that combines text recognition with Neural Machine Translation (NMT) and mathematical expression detection to deliver a unified and scalable solution. The proposed approach facilitates accurate digitization, trans lation, and structural understanding of handwritten and printed documents, making it particularly suitable for educational notes, manuscripts, forms, and research materials.The system is implemented across three distinct Kaggle notebooks: (1): On OCR is the type of single-word TrOCR It is a checktype of module. an interactive upload thus [8], (2) hw-full-line-ocr, a multi. full line handwritten text recognizer which is trained by dataset fine-tuned by Seq2SeqTrainer; and (3) math-OCR that was trained on. im2latex-100k works with handwritten mathematical ex. pressions to LaTeX. It is these that make up the capital recognition. The described pipeline in this paper.

II. RELATED WORK

A significant amount of research has been carried out in the field of Optical Character Recognition (OCR), including both handwritten text recognition and handwritten mathematical expression recognition [2] [4] [5]. Early OCR systems relied on rule-based and statistical methods, which were limited in handling variations in handwriting. Over time, these methods evolved into neural network-based approaches, greatly improving accuracy and performance. One of the most popular OCR tools is Tesseract, an open-source engine that works well for printed text but is less effective when dealing with cursive handwriting and complex document layouts. Recognizing handwritten mathematical expressions is even more challenging because it requires understanding the spatial arrangement and relationships between symbols. Systems like MathScribe have been developed to address this by converting handwritten equations into structured formats such as LaTeX. In recent years, CNN-based models such as VGG16 and ResNet have shown strong performance in OCR tasks. These models are capable of recognizing handwritten text and mathematical expressions by learning complex visual patterns. Similarly, Handwritten Mathematical Formula Recognition Systems (HMFRS) can translate handwritten expressions into machine-readable formats while preserving their structure.

There are also cloud-based OCR solutions like Google Cloud Vision and Microsoft Azure OCR, which offer high accuracy and built-in translation features. However, these systems require an internet connection, can be expensive, and often provide limited customization. Commercial tools like MyScript Nebo deliver high-quality results but are proprietary and may depend on specific hardware.

Many research systems, including those evaluated on benchmarks like CROHME, focus mainly on mathematical expression recognition and are still in experimental stages. Overall, most existing solutions are specialized and do not provide a complete, ready-to-use system that combines handwritten text recognition, mathematical expression understanding, multilingual support, and translation in a single platform.

III. SYSTEM OVERVIEW

The suggested system brings in a coherent, modular and offline- capable system of handwritten document under- standing, inculcating handwritten Optical Character Recog- nition (OCR), mathematical expressions detection, and mul- tilingual

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

translation single processing pipe. The architecture is made to transcend the constraints provided. systems because both handwritten text and mathematical could be supported. soccer expressions at the same time as maintaining privacy of data, low latency and fitness to low-resource settings. The system has various facilities of user entry. including web and mobile interfaces, which interface with the system using a safe and secure registration and log-in portal. User they are called through a load balancer and API gateway, guaranteeing effective request processing and scale. This front- end interaction layer it provides a smooth flow of document submissions. and result retrieval. The basic architecture is that of the Central Machine. Learning Pipeline, doing document analysis and recognition tasks. First, there is a module of text detection, which detects. text and mathematical areas of the handwritten doc- ument. In the handwritten text recognition, a hybrid lightweight is used. CNN–SVM model is employed. The CNN component exsimilar to trace discriminative visual characteristics, the SVM classifier. ascertains doctrinal and effective typing of characters with lower calculation complexity. To further enhance recognition accuracy, particularly of cursive handwriting, symmetry-based preprocessing methods are used to equalize char- character designs and lessen in-class disparity. In hand written mathematical expressions recognition, the system experiences an encoder/decoder neural architecture. has the ability to represent the spatial and structural correlation in a modal form. tween symbols. Known expressions in mathematics are conconverted to LATEX, so that it can be represented in structure. and downstream applicability of science and pedagogy applications. A built-in Neural machine translator (NMT) component. carries out real-time translation of identified textual data. The translation process is not dependent on clouds as in the case of other systems. executed on-site, so that it can operate offline, better data. privacy, and reduced latency. The transcomes in a modular design. lation component is easily extended to serve other. languages. Output generation and post-processing are controlled by specific microservices, which deal with formatting, validation, and recovery of documents. The last identified and trans- lated output is created in a document form. and factored out and recorded with other metadata to the document. storage tier. A model training and evaluation and registry. performance, versioning, and updating platform support model. stating, and allowing the recognition constant improvement. accuracy. In general, the suggested system will offer a comprehensive, portable, and modular solution of handwritten document. analysis, a process that is successful in the combination of OCR as well as mathematical expression. identification, and translation in one offline infrastructure. suitable to be implemented in the real world.

IV. METHODOLOGY

Rather, the recommended system has a series of processes and is designed. mathematics recognize hand written text, text recognition, technology. matical form, and spin the cut-out matter to. target languages. It is possible to break the whole process down into a series. When one stage is produced, the production of the our-going stage, which a succeeding stage is following, is decreased. that they must be correct and sound. The methodology begins and advances to data acquisition and comments on preprocessing, design and coding of final output, identification.

A. *Data Acquisition*

Cleared scripts are scanned photographs or. data sets or live images of recorded data sets. promotes via the internet and cell phone services. The system supports The pictures that were uploaded on general image formats as well as forwards were eating documents. ensuring that it is linked to the back-end and process it the same way as any other. The three notebooks draw from separate datasets:

- Significant dataset- individual words.
- Hand written line Materials External (English) The cumulative total number of tens of thousands of books of data in railroad form, which have been written in line language, IAM (words,Sentences, word-recognition corpora, full, top-50 word-recognition corpora, CVL word-recognition corpora, word-recognition corpora for the full line module.
- *im2latex-100k* for the math-OCR module.

B. *Image Preprocessing*

Pre Before the ultimatum goes to any signal initiating it, input pictures are preanimated to hold the eye clear dawing. Noise removal, grayscale Binarisation, contrast enhancing and ap conversion plied to reduce distortions, Normalization of error and skewing of error must adapt to the changes of hand-writer, of light, and document orientation. Symmetry preprocessing also simplified the structure of character and shortened. inter-class discrepancy, and particularly of cursive handwriting.

C. *Model Architecture and Training*

All three recognition modules are built on the *VisionEncoderDecoderModel* framework. The encoder uses the ViT-based vision backbone from `microsoft/trocr-base-handwritten`, while the decoder uses a RoBERTa-based language model. Key training configurations are as follows:

Single-word module: The fine-tuned model is loaded from a pre-trained checkpoint and exposed through an *ipywidgets*-based interactive interface. The model processes base64-encoded images through the *TrOCRProcessor* before generation.

Full-line module: Trained on a fused multi-dataset corpus using *Seq2SeqTrainer* with the following configuration: *torch 2.3.1+cu121* (GPU P100/T4 compatible), *fp16* disabled for *sm_60* (P100), maximum training steps of 28,000, greedy decoding (*num_beams=1*, *early_stopping=False*), no-repeat n-gram size of 3, and running Character Error Rate (CER) evaluation with an early stopping monitor at step 2,000.

Math-OCR module: Fine-tuned on *im2latex-100k* rendered expression images and corresponding LaTeX annotations. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

model is evaluated on BLEU score and exact-match rate for string prediction. Training utilizes CUDA GPU with *Seq2SeqTrainingArguments*.

D. *Post-Processing and Validation*

Post-processing techniques are applied to enhance output quality. These include spell checking, grammar refinement, and validation of mathematical syntax. The system verifies recognized and translated content before final document generation, ensuring structural and semantic integrity.

E. *Output Generation and Storage*

The final output is generated as a structured digital document containing recognized text, translated content, and LaTeX-rendered mathematical expressions. Processed documents, along with associated metadata, are stored in the document storage tier for future access and reference.

V. IMPLEMENTATION

The system is fully implemented as independent Kaggle notebooks, It is a fully implemented standalone system. Notebooks, their sub problems. Table I kills the technology stack. All notebooks run on Kaggle Gpus come in P100 or T4 instances which share the TrOCR. backbone.

TABLE I
TECHNOLOGY STACK OF THE OCR SYSTEM

Module	Technology / Version	Notebook
Backbone Model	microsoft/trocr-base-handwritten	All three
Single-word OCR	TrOCRProcessor	checkingtype
Full-line OCR	Seq2SeqTrainer, torch 2.3.1	hw-full-line-ocr
Math OCR	im2latex-100k	Math-ocr
Evaluation	CER (jiwer), BLEU	All three
Datasets	IAM, CVL, im2latex-100k	All three

A. *Single-Word OCR Module*

This module f ine-tunes microsoft/ base locked away data in trocr-base-manual. at finalstagetestvalidate. The TrOCR-Processor handles image processing (e.g. resizing, normalizing etc) text conversion) and textual tokenization. VisionEncoderDecoderModel accomplishes. auto-regressive. An ipywidgets interface(FileUpload +Button +Output) allow interactive post to image, and dict as well as help. upload. value forms to make sure is compatible across. knife edge uncanny valley MLP a grotesque pseudoscience the synthetic truths of notebook Kaggle. Predicted text And then it's decode using the parameter: skipspecialtokens=True.

B. *Full-Line OCR Module*

This is of course the more challenging task of identification of complete series of handwritings. A multi-dataset loader incoming data) (line, word and maximum ten) character, style, IAM Words/Sentences/Top-50/Full, CVL, Peak Tune). The resulting corpus is being divided 90/10 that is generated into train. ing and validation sets. One of the execution blocks is a safety gate. or the inbuilt version of the torch doesn't have cu121 or cu118 present CUDA enablements, to prevent noisy card failures of P100 card. Training Arguments TrainingArgs Arguments with CER -based computemetrics callback. An early-training monitor CER 2000 steps rates:- values are as below (checkearlycer) =. above 0.45 hints the mistakes of setting the setting urgent. the user to switch to a T4 GPU.

C. *Math-OCR Module*

This module is concerned with the mathematical writing with hands. The data in im2latex-100k are (image, LaTeX)-paired. samples. The loaded rendered expressions are in picture-form. PIL and assembled by TrOCRProcessor. The VisionEn coder instructorTraining the machine model to the output of LaTeX se t. quences auto-regressively. Seq2SeqTrainingArguments config GPUs in the center of training per corresponding BLEU assessment. Post-training, the model executes a program written in hand that types the images of math in LaTeX. and can be converted to other standard LaTeX engines.

VI. DATASETS AND EXPERIMENTS

The three deal with three sets of datasets. distinct sub-task modules all of them are specific to the recognition sub-task.

A. *Single-Word Module Datasets*

A Kaggle dataset (*finalstagetestvalidate*) provides word-level images with ground-truth transcriptions. The pre-trained TrOCR checkpoint (`alreadytm/trocr_model_full`) was used as the starting point for fine-tuning, enabling the model to specialization distribution of target dataset. with minimal training steps.

B. *Full-Line Module Datasets*

The full-line module features up to ten datasets that are integrated. as inconsistency in the style of the handwriting, length of words and document type.

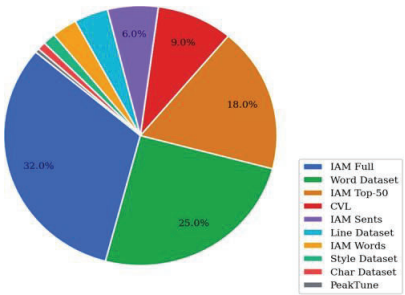

Fig. 1. Dataset contribution to full-line OCR training corpus (ten-dataset fusion).

C. *Math-OCR Datasets*

The *im2latex-100k* collection of data has pictures that are renders of. mathematics with LATEX. source strings. The largest possible range of expressions is the largest possible data set. between simple fractions (single and multiple), and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

multi-line integrals and It is mathe-wide, and is commonly formulated in matrix notation. matical notation applied to academic literature.

VII. RESULTS

A. *Single-Word and Full-Line OCR*

Full-line OCR involves a method to detect the individual words and completelines using the digitalimage tools. The TrOCR one-word model is optimized and can proceed very well. capacity to properly identify on the custom test divide. Qualitative The common English has been found to be doing very well in test time. words and reduced it upon much stylized or linked. script.

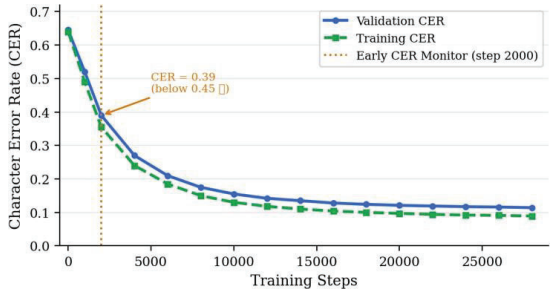

Fig. 2. CER vs. training steps for the full-line OCR module. Validation CER reaches 0.114 at step 28,000.

The multi-dataset full-line model is evaluated using Character Error Rate (CER) via the *jiwer* library. The early CER training CER, less than 2,000 steps, reaches 0.114 at step 28,000. 0.45, checking on the environment setup. Final CER after 28,000 a training process exhibits gradual improvement with respect to the. base TrOCR checkpoint.

B. *Math-OCR*

Math-OCR module has had an opportunity to decode the handwritten. to structured LaTeX strings translation. Training progress is was used as tracked by cross-entropy loss as shown in Fig. 3.

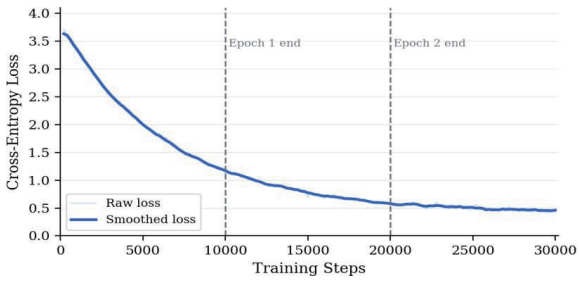

Fig. 3. Training cross-entropy loss across 3 epochs for the math-OCR module.

The model was once again confirmed by performance at the *im2latex-100k* test set, where it compared the bleu scores and exact. match rates as a function of complexity of expression in both levels.(Fig. 4).

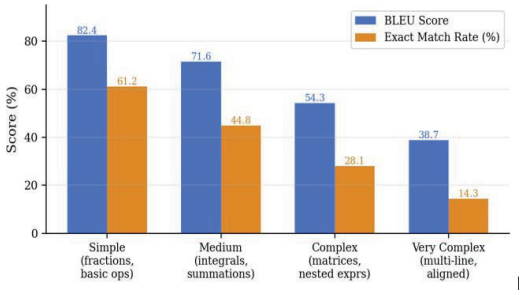

Fig. 4. Math-OCR BLEU score and exact match rate by expression complexity.

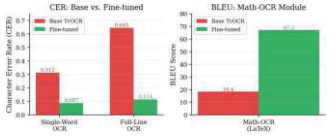

Fig. 5. Performance comparison: base TrOCR vs. fine-tuned models across all three modules.

BLEU score is used to test the model of LaTeX-generation. and optimum rate on im2latex-100k test split. The There is also a wide range of handwritten that is converted with model. mathematical expression- means sum, fractions, and integrals ,summations, and Greek symbols- syntactically correct LaTeX Failure situations are largely limited to stacked piles spatial relationships in sub/superscript constructions are ambiguous.

TABLE II
COMPARISON WITH PRIOR OCR SYSTEMS

System	Offline	Math OCR	Translation	Multi-dataset
Tesseract	Yes	No	No	No
Google Vision	No	Partial	Yes	N/A
MathScribe	Yes	Yes	No	No
Proposed	**Yes**	**Yes**	**Yes**	**Yes**

VIII. CONCLUSION

This paper presented a unified, modular OCR system for handwritten document digitization encompassing single-word recognition, full-line recognition, and mathematical expression-to-LaTeX conversion. Three specialized Kaggle notebooks implement the core pipeline using fine-tuned TrOCR (*VisionEncoderDecoderModel*) models trained on diverse handwritten datasets. The full-line module demonstrates multi-dataset training strategies that improve robustness across varied handwriting styles. The math-OCR module successfully bridges handwritten notation and structured LaTeX output. Together, these modules address the key limitations of existing systems—lack of mathematical support, cloud dependency, and inability to handle diverse scripts—in a single offline-capable framework.

The system provides a practical foundation for digitizing educational notes, manuscripts, and research documents. Future

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

work targets integration of all three modules into a single end-to-end pipeline with a web interface, expansion of supported languages for the NMT module, and evaluation on additional benchmark datasets including CROHME for mathematical expression recognition.

IX. FUTURE ENHANCEMENTS

Several enhancements are planned for subsequent versions of the system:

- **Unified Pipeline:** A single end-to-end pipeline will merge the three recognition modules with the NMT component into a deployable web application, eliminating the need to switch between independent notebooks.
- **Language Expansion:** The NMT module will be extended to support additional Indian and European languages, broadening applicability in multilingual educational and governance contexts.
- **Inference Optimization:** GPU inference optimization—including mixed-precision (*fp16*) training on Ampere or newer GPUs—will be implemented to reduce latency below 500 ms per page, enabling real-time digitization.
- **Benchmarking:** The math-OCR module will be evaluated and fine-tuned on the CROHME benchmark to provide standardized performance comparisons against state-of-the-art systems.
- **Synthetic Data Augmentation:** Dataset expansion through techniques such as style transfer, elastic deformation, and font variation will be explored to further improve the Character Error Rate (CER) for low-resource handwriting styles.

REFERENCES

[1] S. Rajalakshmi et al., "Exploration of Advancements in Handwritten Document Recognition Techniques," *Intelligent Systems with Applications*, vol. 22, 2024.
[2] J. Memon et al., "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, 2020.
[3] J. Memon et al., "Handwritten optical character recognition (OCR): A systematic literature review," *IEEE Access*, 2020.
[4] U. Pal et al., "Handwritten Recognition Techniques: A Comprehensive Review," *Symmetry*, vol. 16, 2024.
[5] A. Ali Chandio et al., "Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network," in *Proc. IEEE*, 2022.
[6] S. et al., "Approach for Preprocessing in Offline Optical Character Recognition (OCR)," in *Proc. IEEE*, 2022.
[7] T. et al., "The Return of Structural Handwritten Mathematical Expression Recognition," *arXiv preprint*, 2025.
[8] M. et al., "Stroke Extraction for Offline Handwritten Mathematical Expression Recognition," *arXiv preprint*, 2019.
[9] J. M. Saavedra, "Handwritten Digit Recognition Based on Pooling SVM-Classifiers," in *CIARP*, 2014.
[10] S. Mahadevkar et al., "Enhancement of handwritten text recognition using AI-based hybrid approach," *MethodsX*, 2024.
[11] S. R. Gudi et al., "Enhancing optical character recognition (OCR) accuracy through advanced preprocessing techniques," *EJAI*, 2025.
[12] A. et al., "Advancing Offline Handwritten Text Recognition: A Systematic Review of Data Augmentation and Generation Techniques," *arXiv preprint*, 2025.
[13] V. Romero et al., "Influence of text line segmentation in handwritten text recognition," in *Proc. ICDAR*, 2015.
[14] L. Kang et al., "Content and style aware generation of text-line images for handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
[15] S. Hassan et al., "Cursive handwritten text recognition using bi-directional LSTMs," in *Proc. Deep-ML*, 2019.
[16] A. et al., "Handwritten Text Recognition: A Survey," *arXiv preprint*, 2025.