International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 05 | May 2025 www.irjet.net p-ISSN: 2395-0072

AI-POWERED OCR FOR DIGITIZING HANDWRITTEN HISTORICAL DOCUMENTS IN TAMIL

Dr. L. Rasikannan[1], S. Pratheek [2], N. Mohammed Riyas[3], R. Bharathkumar[4]

[1]Associate Professor, Government College of Engineering Srirangam, Tamil Nadu, India
[2,3,4,]UG Student, Department of CSE, Government College of Engineering Srirangam, Tamil Nadu, India

---

ABSTRACT

Sophisticated optical character recognition (OCR) technology enables conversion of handwritten documents with archived material to become editable Tamil versions which are easy to read. A regional script and handwriting style-specific OCR model enables the extraction of text from scanned images affected by time effects and faded ink as well as irregular document formatting. The preprocessing methods that include contrast modification and noise reduction and image binarization facilitate better recognition accuracy and enhanced image quality. The foundation supports important information conservation during document digitization especially when dealing with threatened historical records and court papers and cultural texts.

KEYWORDS: Optical Character Recognition (OCR), Handwritten Text Recognition, Document Digitization, Preprocessing Techniques

1.INTRODUCTION

The present digital era makes data management and preservation more effective through converting physical documents into machine-readable formats. Users require optical character recognition (OCR) technology due to its ability to extract text from different document sources including handwritten and printed as well as scanned materials. The system offers valuable benefits when transforming legal papers and historical archives and cultural documents that exist solely in physical form. OCR technology needs powerful adaptability to address handwriting diversity and paper deterioration and complex writing patterns that are common in Tamil scripts. Accurate and structured digitization is achieved by using advanced preprocessing solutions and specialized text recognition algorithms with post-processing methods in the proposed system to handle digitization obstacles. Digitized preservation of data together with enhanced document readability constitutes the main advantage of modern document processing and accessibility.

1.1 Optical Character Recognition

Digital conversion of text-based documents ranging from printed books to scan paper documents and handwritten notes becomes feasible through optical character recognition technology known as OCR. The recognition algorithms examine light and dark pixels to interpret detected words and text through image analysis. Digital archives get value from OCR to manage historical records and establish automatic data entry methods while allowing users to search text across entire archives. The most sophisticated OCR platforms accept various handwriting styles and script complexities to keep intact the original document layout.

1.2 Handwritten Text Recognition

Handwritten Text Recognition operates as a specific part of Optical Character Recognition (OCR) to convert hand-written content to digital machine-readable text. Natural text recognition proves harder than machine-printed text recognition mainly because of different writing styles together with alignment inconsistencies and diverse stroke patterns. Specialized algorithms of machine learning and deep learning enable accurate interpretation of handwritten digital data received through digital surfaces or scanned images. This method helps digitize both documents written in regional languages as well as historical manuscripts as well as personal notes and filling forms. Handwritten data extraction with precision allows better organization of data while ensuring easy access and extended digital preservation capabilities.

1.3 Document Digitization

The method of converting physical paperwork consisting of printed files and handwritten data alongside old manuscripts into digital storage formats for electronic handling and retrieval defines document digitization. When extracting and organizing textual material the scanning of physical documents follows with optical character recognition technology (OCR). The digitalizing process enables efficient document archiving and document sharing and seamless integration of systems across the board. Modern record-keeping depends on document digitization because it ensures vital information remains accessible even in the long term for academic institutions, healthcare facilities, government departments and cultural conservation programs.

1.4 Preprocessing Techniques

Before optical character recognition (OCR) or handwritten text recognition can be applied correctly to raw document images it is essential to use preprocessing techniques. The methods enhance text visibility through noise removal to fix distortions which aim to deliver superior quality by improving input image readability. The main preprocessing methods for document images include three techniques: binarization for converting greyscale to black-and-white, contrast adjustment for background-text clarity and noise reduction to remove unwanted markings. The text alignment requires skew correction together with scaling and sharpening processes to improve document sections. Standardized and optimized input data processing through preprocessing techniques results in superior recognition algorithm performance alongside better recognition accuracy which produces dependable and consistent text extraction from various document types.

2. LITERATURE REVIEW

Solene Tarrid[1] et al. have recommended setting up a novel database to obtain information from historical handwritten documents. The study corpus contains 5393 finding aids that originated from six distinct series across the 18th and 20th centuries. Unwritten papers which serve to identify and explain historical archives prior to them are denoted as finding aids. Archive materials stored at the French National Archives are located and identified by archivists through these documents. The information retrieval fields and annotation occur at a page level for all documents while these documents contain seven accessible recoverable fields. Research on segmentation-free information extraction algorithms cannot develop from this dataset because each field position remains unreachable to users. The presented Transformer-based model facilitates end-to-end information extraction while serving as the basis for our three training, validation, and testing sets which enable fair comparison with follow-up work. A free access system exists for this database. Machine learning and deep learning methods have transformed web information retrieval so they have become the standard process employed by archives today. Physical documents make up the majority of materials stored in archives while distinguishing them from naturally digital data that appears on the internet or standard information systems. The massive digitization of numerous archival documents during recent times has digitalized a substantial portion of the entire collection yet complete digitization seems unlikely due to the immense number of accessible documents. The decision about which documents to digitize should be determined according to their worth.

Denis Coquenet[2] and his team reveal through their research that current Computer vision systems encounter problems recognizing unbound handwritten text. Standards in text recognition of paragraphs utilize two operational models that identify text lines and perform line segmentation. The proposed solution provides an integrated end-to-end system which utilizes hybrid attention for resolving this issue. The FFLINE technology uses scanned paper pictures to determine their handwritten text content. The system processes images of words, lines or paragraphs or complete documents so that it produces the sequence of letters in the text. The research investigates a complete neural network system that processes paragraph text without line segmentation. Competing technologies for handwriting recognition and line segmentation reside in the research field for many years without achieving remarkable progress despite continued scientific effort. Such components only appear exceptionally throughout the literature and have rarely been educated collectively in one trainable structure. Historical approaches to this problem started with character segmentation then performed classification on each character individually. The word level segmentation technique was applied first followed by line segmentation. The problem retains the same nature of entity identification regardless of which segmentation threshold is applied. This research proceeds to analyse an entire textual paragraph without applying explicit segmentation at the training stage or during the decoding phase. Modern page recognition systems follow a two-step operation to detect full page pictures. The initial stage splits the document into text lines while the second step identifies each of those lines. Because OCR stands for Optical Character Recognition technology that applies optical models for recognition purposes it earned its name.

In this system, Wenqi Zhao [3] et al. have suggested Recently, there has been significant improvement in handwritten mathematical expression recognition using encoder-decoder models. However, appropriately allocating attention to visual characteristics remains a difficulty for present approaches. Furthermore, the encoder-decoder models often use RNN-based models for their decoder portion, rendering them ineffective for handling lengthy LATEX sequences. This research uses a transformer-based decoder instead of an RNN-based one, which results in a highly succinct model architecture overall. In addition, a new training approach is presented to fully use the transformer's potential in bidirectional language modeling. Experimental results show that our approach enhances the ExpRate of existing state-of-the-art techniques on CROHME 2014 by 2.23% when compared to numerous methods that do not employ data augmentation. In a similar vein, we increase the Exp Rate by 1.92% and 2.28% on CROHME 2016 and 2019, respectively. The encoder-decoder models have

shown impressive performance on a range of applications, including picture captioning and text recognition in scenes. The goal of Handwritten Mathematical Expression Recognition (HMER) is to use the handwritten math expression picture to construct the math expression LATEX sequence. In recent years, various encoder-decoder models have been suggested for HMER, as it is also an image to text modelling challenge. But to varied degrees, the lack of coverage issue plagues current approaches. Over-parsing and under-parsing are the two probable expressions of this issue. Under-parsing indicates that some areas of the HME picture are left untranslated, whilst over-parsing indicates that other areas are redundantly translated many times.

The authors of this system propose utilizing Transformer architecture according to Alexey Dozois [4] et al. The operation of attention in vision takes place alongside convolutional networks or through the replacement of specific convolutional network parts without changing the overall network structure. The application of a pure transformer to picture patch sequences produces exemplary outcomes in image classification tasks which makes specific reliance on CNNs redundant. The computational requirements for training the Vision Transformer (ViT) are much lower than those of current convolutional networks allowing it to execute effectively when trained on extensive data and then transferred to various image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.). We investigate the direct application of basic Transformers on photographs while making minimal modifications due to the Transformer scaling success in NLP. The method starts with dividing an image into sections before feeding Transformer with their linearly encoded segments. An NLP application approaches image patches in a method analogous to term processing which equals word handling. The model receives supervised image classification training signals as part of its educational process. The research team investigated applying Transformers as a direct solution for visual image recognition. We exclude all image particular inductive biases in our design except from the initial phase of extracting image patches while avoiding methods that use self-attention in computer vision. Our system treats pictures as series of patches and utilizes a standard NLP Transformer encoder to process them.

Sanket Biswas et al developed this system to address Although there has been great advancement in the state-of-the-art picture generating models, creating document images with intricate and multi-object layouts remains a difficult issue. The research delivers an automatic approach to produce document visuals by following certain layouts under the name DocSynth. Through our suggested DocSynth framework a collection of realistic document images emerges from user-supplied spatial layout (item category-marked bounding boxes) to produce layouts which match the predefined specifications. The framework exists in two variants for this study because it produces better artificial document images that usable alongside real data during document layout analysis training. The model performance has been improved through multiple sets of learning goals. Standard measurement techniques help us analyze the actual outcomes from our model alongside numerical data evaluation. Our model exhibits the ability to generate realistic and diversified document pictures with multiple items during its output process. The method details a complete qualitative breakdown of all the processes required for synthetic image synthesis. This marks the debut of a system of this nature which we are aware of. Document Analysis and Recognition sets understanding documents automatically as its top and essential goal among all subfields. Digital documents along with scanned paper documentation operates together during business operations at present. Different fields present unique papers in their real-world documentation (including forms and invoices along with letters etc.) that show remarkable variation. Modern Robotic Process Automation (RPO) systems have gained strong market demand to run automatic document workflow information handling in paperless workplaces that combine reading capabilities with understanding abilities.

3. EXISTING SYSTEM

Handwritten text recognition stands as a demanding task in computer vision because of its unrestricted nature. Handwritten text recognition techniques were traditionally split into two sequential steps which involve segmentation followed by text line identification. The Document Attention Network represents an end-to-end segmentation-free architecture which provides first-time solutions for handwritten document recognition. Training occurs by tagging textual elements using start and end markings in an XML structure as well as carrying out text recognition. The proposed model contains transformer decoder layers stacked together to produce token-by-token predictions alongside an FCN encoder available for feature extraction process. The system processes complete text documents as an entry before creating sequential logical layout tokens and characters. The system operates without requiring segmentation labels during training despite standard techniques using these labels. The competitive outcome produced 3.43% and 3.70% CER at the page and double-page analysis stages on READ 2016. The RIMES 2009 dataset shows a page-level CER measurement of 4.54% during the analysis.

4. PROPOSED SYSTEM

The proposed method converts documents of various types into editable data through OCR processing that recognizes both handwritten and printed content. The preprocessing steps featuring noise reduction as well as contrast enhancement and binarization produce the highest possible input quality thus enabling improved text visibility and recognition outcomes. A special OCR model setup enabled by configuration automatically extracts texts from multiple sources in an efficient and reliable manner across different handwriting styles and scripts. Post-processing techniques apply to recognized text to fix all types of distortion thereby improving both accuracy and readability of the output. The systematic documentation method protects valuable data while it enables better document accessibility through enhanced organization systems. Future access to digital content becomes possible through this method which enables better record management practices.

A. Image Acquisition

The photo ingestion module enables employees to take record images and import data from various sources such as printed documents and scanned handwritten documents and photographic material. The collection of input photographs uses this module to receive digital images in specific formats which maintain their necessary quality for further processing. The recorded image quality and consistency at this level strongly affects identification accuracy so it establishes all the fundamentals for the OCR system.

B. Preprocessing

The preprocessing software performs efficient text recognition on images that have been obtained. The image clarity improvement and text part separation process uses techniques such as noise reduction and contrast enhancement and scaling and greyscale conversion and binarization. Through this preprocessing step the OCR algorithm receives optimal inputs that lead to better character recognition because background noise gets removed while text becomes more visible.

C. Text Recognition

OCR methods extract text from preprocessed images as the foundational module's main operation. The module supports multiple languages as well as several handwriting styles together with printed typefaces. The OCR engine searches image pixels one at a time to find patterns which it converts into machine-readable text. This solution's adaptability permits it to function on files with various design options and writing directions and levels of image quality.

D. post-processing

The post-processing tools enhance extract output reading quality while correcting potential text errors that occur during raw text extraction. System maintenance focuses on correcting common automated text recognition errors which include mistaken characters alongside inconsistent designs together with wrong line placements. An additional language-based grammar check together with spelling correction and text alignment exists to create a clear output that reflects both structure and content of the original document..

E. Document Management

The last module manages organized editable texts originated from previous phases. The system provides easy access to modified documents through searchable storage solutions which also enable users to update and classify such documents. Digital document integrity retention with future availability becomes possible through this module because it supports metadata tagging along with indexing and secure storage features. This component supports long-term archive needs together with large-scale data organization for maintaining accessibility.

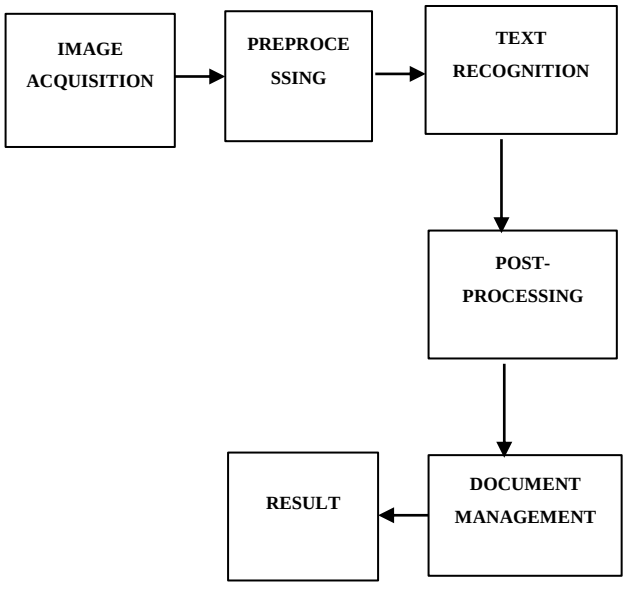

Figure 1 System Flow Diagram

5. RESULT ANALYSIS

The OCR system that was put into place performed well when it came to digitizing printed and handwritten documents, especially those in Tamil. Using preprocessing techniques greatly improved image clarity, which increased text recognition accuracy. The system extracted text with great reliability, handling handwriting style differences, deteriorated papers, and

uneven backdrops. Post-processing improved the output's overall readability and decreased recognition mistakes, further refining the results. With its searchable and editable formats and structural integrity preserved, the digitized information ensured improved document management and accessibility. This outcome attests to the system's resilience and usefulness in managing and conserving historical, legal, and cultural records for future digital usage.

The percentage of accurately detected positive events among all instances anticipated to be positive is known as precision.

Precision=True Positives (TP)+False Positives (FP)/Tru e Positives (TP)

Recall

The percentage of accurately identified positive events among all real positive instances is known as recall.

Recall=True Positives (TP)+False Negatives (FN)/True P ositives (TP)

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics.

F1-score=2×Precision+Recall/Precision× Recall

Accuracy

The percentage of accurate forecasts both positive and negative among all predictions is known as accuracy.

Accuracy=Total number of instances/True Positives (TP)+True Negatives (TN)

Algorithm	Accuracy
Existing	75
Proposed	88

Figure 1 comparison Table

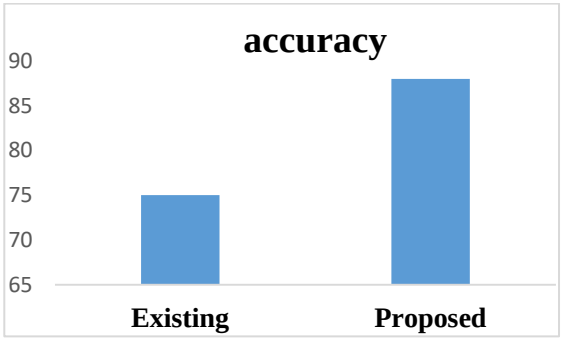

Figure 2 comparison Graph

6. CONCLUSION

The proposed Optical Character Recognition (OCR) approach successfully conquers digitization problems affecting multilingual (including Tamil) printed and handwritten documents. The conversion system transforms different documents to structured digital content with high precision through combined advanced processing with reliable recognition solutions and post-processing methods. This system can manage decayed original documents from various sources thus it ensures valuable historical legal and cultural information remains accessible for future generations. This OCR-based method safeguards linguistic and cultural history for future generations and enhances document management together with accessibility.

7. FUTURE WORK

Studies in the future will strengthen OCR system capabilities by expanding its capacity to process different kinds of documents written in multiple languages. The system's recognition performance may improve by applying the latest deep learning models including Transformers and Convolutional Neural Networks (CNNs) since they help with complex or heavily distorted handwriting. Making the system suitable for global use requires implementing real-time OCR function and support for different languages and scripts. Smoothing document management and communication using cloud storage integration would improve the system's usability to a higher degree. Future research should focus on simplifying the document post-processing process to develop automatic solutions for resolving challenging text-related errors.

8. REFERNECES

[1] "Vertical attention network for end-to-end handwritten paragraph text recognition," IEEE Transactions on Machine Intelligence and Pattern Analysis, 2022

[2] “Transformer-based technique for joint handwriting and named entity detection in historical documents,” Pattern detection Letters, vol. 155, pp. 128–134, 2022. 4, 5; C. Rouhou, M. Dhiaf, Y. Kessentini, and S. B. Salem

[3] “Handwritten mathematical expression recognition using bidirectionally trained transformer,” in International Conference on Document Analysis and Recognition, vol. 12822, 2021, pp. 570–58, W. Zhao, L. Gao, Z. Yan, S. Peng, L. Du, and Z. Zhang

[4] "An image is worth 16x16 words: Transformers for image recognition at scale," by A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit,

and N. Houlsby, at the 9th International Conference on Learning Representations, 2021.

[5] X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and T. Unterthiner, "Transformers for picture identification at scale: An image is worth 16x16 words,"

[6] "Docsynth: A layout driven technique for controllable document image synthesis," by S. Biswas, P. Riba, J. Lladós, and U. Pal, in International Conference on Document Analysis and Recognition, vol. 12823, 2021, pp. 555–568. 7

[7] "Trocr: Transformer-based optical character recognition using pre-trained models," 2021, M. Li, T. Lv, L. Cui, Y. Lu, D. A. F. Florêncio, C. Zhang, Z. Li, and F. Wei

[8] "Layoutlmv2: Multimodal pre-training for visually-rich document understanding," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference, by Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. A. F. Florêncio, C. Zhang, W. Che, M. Zhang, and L. Zhou

[9] The International Conference on Document Analysis and identification, 2021, pages. 319–334; R. Atienza, "Vision transformer for quick and efficient scene text identification."