IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, January 2026
Impact Factor: 7.67

Handwritten Text Digitizer

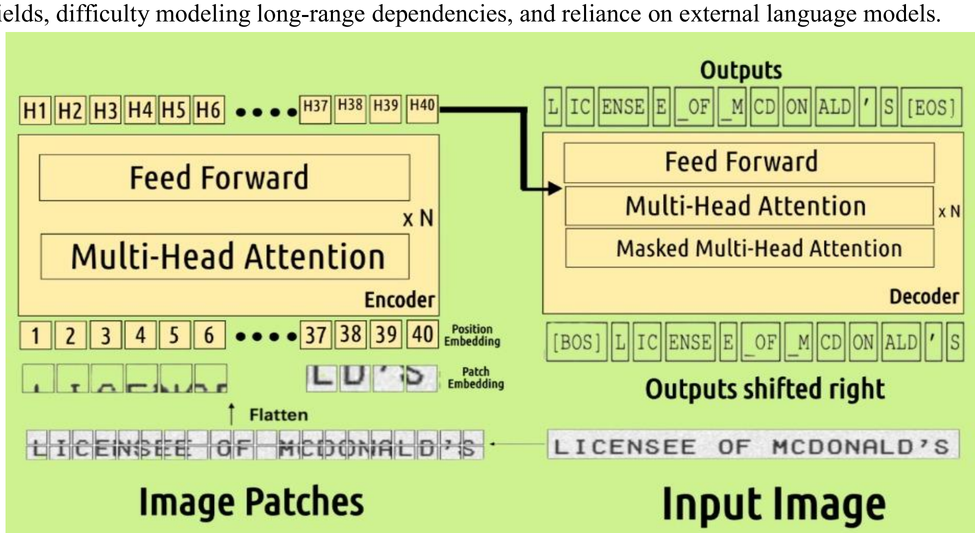
Dr. M. A. Pradhan, Kshirsagar Prathamesh Rahul, Chavan Aditya Kamlesh, Patel Touhid Sabir
Department of Computer Engineering
All India Shri Shivaji Memorial Society College of Engineering, Pune, Maharashtra
mapradhan@aissmscoe.com, prathmeshkshirsagar2003@gmail.com
aditya.chavan1504@gmail.com, touhidpatel06@gmail.com

Abstract: *Handwritten text recognition remains a difficult problem due to variations in writing styles, stroke patterns, noise, and inconsistencies in real-world documents. Conventional Optical Character Recognition (OCR) systems rely on convolutional and recurrent networks, which show limitations when dealing with unconstrained handwriting. This paper presents a handwritten text digitization system based on TrOCR, a fully Transformer-driven OCR architecture using a Vision Transformer encoder and a pretrained text Transformer decoder. The system operates end-to-end without convolutional backbones, recurrent decoding, or external language models. A complete methodology is proposed, including preprocessing, patch embedding, autoregressive decoding, and text generation. Experimental evaluation demonstrates the effectiveness of TrOCR for handwritten digitization with improved accuracy over conventional baselines. The system is suitable for document automation, archival digitization, intelligent extraction, and data processing applications*

Keywords: Handwritten Text Recognition, TrOCR, OCR, Transformer, Vision Transformer, Deep Learning, Digitization

I. INTRODUCTION

Handwritten text digitization plays a vital role in transforming physical or scanned content into machine-readable data for large-scale automation, archival preservation, and intelligent information retrieval. The irregularity of handwriting makes this a difficult problem: unlike printed text, handwritten characters vary significantly in shape, slant, spacing, and stroke composition, often leading to higher recognition error rates. Traditional OCR systems depend on convolutional neural networks (CNNs) for extracting visual features and recurrent neural networks (RNNs), commonly supplemented with Connectionist Temporal Classification (CTC), for sequence decoding. Although successful for structured printed text, these architectures are inherently limited for free-form handwriting due to their restricted receptive fields, difficulty modeling long-range dependencies, and reliance on external language models.



Recent advancements in Transformer architectures have reshaped image understanding and natural language processing. The self-attention mechanism in Transformers captures global context and eliminates recurrence, improving

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, January 2026
Impact Factor: 7.67

sequence modeling capabilities. Based on this paradigm, Microsoft proposed TrOCR, a fully Transformer-based OCR system integrating a Vision Transformer encoder with a pretrained text Transformer decoder. TrOCR treats an input image as a sequence of patches and generates tokenized text autoregressively, offering an end-to-end solution without CNNs or RNNs.

This work proposes a handwritten text digitizer built using TrOCR, providing an accurate and scalable system for modern document processing tasks.

II. LITERATURE REVIEW

A. Early OCR: Template-Based and Statistical Methods

Early OCR systems relied predominantly on template matching, handcrafted features, and rule-based classification. Template-based systems compared input characters to predefined models using correlation metrics, whereas statistical approaches such as Hidden Markov Models (HMMs) modeled character sequences probabilistically. Although effective for structured printed text, these methods lacked robustness against the variability inherent in handwriting, including inconsistent character spacing, cursive writing, stroke variation, and document noise. Limited generalization capabilities, dependence on handcrafted features (e.g., zoning, projection histograms, contours), and sensitivity to image distortions significantly restricted their application in unconstrained handwriting recognition.

B. Transition to Machine Learning and CNN-Based OCR

The introduction of machine learning improved handwriting recognition through methods such as Support Vector Machines (SVMs), k-Nearest Neighbors (kNN), and feedforward neural networks. However, the true breakthrough came with Convolutional Neural Networks (CNNs), which became central to modern OCR pipelines. CNNs excel at extracting hierarchical visual features and reduce the need for handcrafted feature engineering. Architectures like LeNet initiated this trend, followed by more advanced CNN-backbones, including VGG, ResNet, and DenseNet.

Despite their success, CNN-only systems lack mechanisms for modeling sequence relationships, making them unsuitable for full line-level handwriting recognition. The spatial dependence across characters and long-range contextual cues remain difficult to capture using purely convolutional structures.

C. RNN/CTC-Based Handwriting Recognition

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, became dominant for handwriting recognition due to their ability to model sequential patterns. Many hybrid OCR pipelines combine CNNs for visual extraction with bidirectional LSTMs to capture context across textlines. The introduction of Connectionist Temporal Classification (CTC) addressed alignment issues between input images and output sequences by enabling character prediction without requiring explicit segmentation.

Models such as CRNN (Convolutional Recurrent Neural Network) achieved strong performance and became widely used for scene and handwritten text recognition. However, RNNs suffer from several limitations:

- Sequential processing is slow and not parallelizable
- Limited capacity to model long-range global dependencies
- Susceptibility to vanishing/exploding gradients
- Dependence on external language models (LMs) to improve performance

These limitations particularly affected high-variability handwriting tasks.

D. Attention-Based Seq2Seq Recognition Models

Attention mechanisms revolutionized OCR by enabling models to learn alignment between image regions and output tokens implicitly. Encoder-decoder architectures inspired by machine translation made significant progress in scene text and handwriting recognition. Unlike CTC, attention-based models can directly learn character alignments through soft-attention processes.

DOI: 10.48175/IJARSCT-30865

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, January 2026
Impact Factor: 7.67

Architectures such as SAR (Show-Attend-Read), ASTER, and DAN (Decoupled Attention Network) demonstrated improved expressiveness and flexibility. However, these models still relied on CNN backbones for visual feature extraction, limiting their ability to capture global relationships in the image.

E. Vision Transformer-Based OCR Approaches

The introduction of the Vision Transformer (ViT) marked a fundamental shift in computer vision research. ViT treats images as sequences of fixed-size patches and processes them via self-attention, enabling global context modeling without convolution. ViT-inspired OCR models include:

- ViTSTR: A pure Vision Transformer model for scene text recognition.
- SATRN: Utilizes self-attention for enhanced sequence modeling.
- ABINet: Combines visual modeling with autonomous language modeling.
- Donut: A document-understanding Transformer that performs OCR and structured extraction.
- PARSeq: Permuted autoregressive sequence modeling for OCR.

These models improved generalization and long-range context understanding, but most lacked strong language-modeling capabilities within the decoder.

F. Transformer-Based OCR Without CNNs or RNNs

Non-recurrent Transformer-based models such as NRTR introduced sequence-to-sequence OCR that removed RNNs entirely. However, they still depended heavily on CNN-based encoders, failing to fully eliminate convolutional components. The absence of pretrained language models also restricted their linguistic accuracy.

DTrOCR, introduced in IEEE/CVF WACV 2024, explored decoder-only Transformers for OCR, showing competitive results and demonstrating the feasibility of minimalistic architectures.

Multilingual OCR Transformers also emerged, extending recognition to scripts beyond Latin. These include cross-lingual encoders and multilingual decoders trained on diverse datasets.

G. TrOCR: A Fully Transformer-Based OCR Model

TrOCR (Transformers for OCR) represents a major advancement by providing a fully end-to-end Transformer architecture comprising:

Vision Transformer Encoder:

- Treats images as patch sequences
- Learns global and local visual patterns
- Outperforms CNN-based encoders in robustness

Pretrained Text Transformer Decoder:

- Based on RoBERTa/MiniLM
- Autoregressively generates tokens
- Inherently models language semantics
- Eliminates need for external language models

Massive Two-Stage Pretraining:

- 684M synthetic printed textlines
- 17.9M synthetic handwritten textlines
- Fine-tuned on IAM and other datasets

TrOCR achieves state-of-the-art Character Error Rate (CER) on handwriting datasets and generalizes across styles, slants, and inconsistencies.

The model's encoder–decoder structure and self-attention mechanisms enable superior context modeling compared to CNN–RNN architectures.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, January 2026
Impact Factor: 7.67

H. Recent IEEE-Aligned Works Related to TrOCR

Several IEEE/IEEE-CVF papers extend or complement TrOCR:

DTrOCR (Decoder-Only Transformer OCR)

- Published at IEEE/CVF WACV 2024
- Removes encoder entirely
- Lightweight and competitive for printed + handwritten text

Graph Transformer OCR for Handwritten Documents (WACV 2024)

- Combines Transformers + Graph Neural Networks
- Good for structured document extraction

Robust Multilingual Handwriting Transformers (IEEE TPAMI 2023)

- Supports multiple scripts
- Important for multilingual countries like India

These works demonstrate a trend toward Transformers replacing CNN/RNN pipelines entirely.

III. METHODOLOGY

A. System Overview

The proposed handwritten text digitizer follows an end-to-end processing pipeline based on a Transformer architecture. Handwritten document images are first normalized and converted into fixed-size inputs to ensure model compatibility. Instead of relying on character segmentation, the system treats the entire textline as a continuous sequence. A Vision Transformer encoder extracts high-level visual representations using self-attention, enabling global context modeling across the entire image. These representations are then passed to a language-aware Transformer decoder, which predicts the corresponding character sequence in an autoregressive manner. This unified framework eliminates the need for handcrafted features, segmentation rules, or recurrent structures, making the system highly adaptable to diverse handwriting patterns.

B. Dataset and Data Preparation

Handwritten OCR performance depends heavily on the diversity and quality of training data. In this work, the dataset includes real handwritten samples and synthesized handwriting generated using multiple fonts and stroke variations. The preprocessing stage involves resizing all images to a uniform resolution (384×384) to maintain consistency with the TrOCR architecture. Data augmentation introduces controlled distortions such as rotation, skewing, brightness variations, and random noise injection. These augmentations increase the model's robustness by exposing it to handwriting-like distortions commonly found in real-world documents. Normalization ensures that pixel intensity distributions remain stable, improving convergence during inference or fine-tuning.

C. Preprocessing

Preprocessing is critical for enhancing the system's accuracy by reducing noise and improving image clarity. Grayscale conversion reduces computational complexity by eliminating redundant color channels. Gaussian filtering and median blurring are applied to suppress salt-and-pepper noise and uneven ink distribution. Normalization scales pixel values between 0 and 1 to stabilize input variance. Optional binarization through Otsu's thresholding can be applied to increase contrast between strokes and background. Additionally, geometric corrections, such as deskewing and aspect-ratio preservation, ensure that textlines are horizontally aligned, which improves attention alignment during encoding.

D. Vision Transformer Encoder

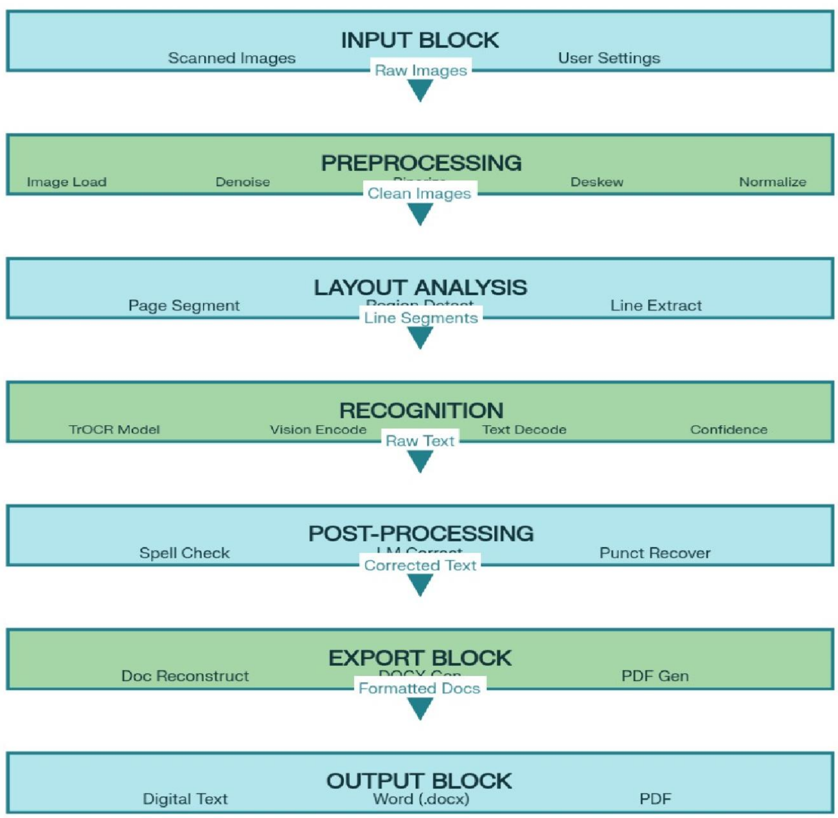
The Vision Transformer (ViT) encoder represents a paradigm shift from convolution-based architectures. The image is divided into fixed-size patches (typically 16×16), each treated as a token analogous to words in NLP. Each patch is linearly projected into a high-dimensional embedding. To preserve spatial structure, learnable positional embeddings are added to every token. The encoder consists of multiple layers of Multi-Head Self-Attention (MHSA) and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, January 2026
Impact Factor: 7.67

feedforward networks. MHSA allows each patch to attend to every other patch in the image, enabling global dependency modeling. This is particularly advantageous for handwriting recognition, where the shape of a character can be influenced by neighboring strokes or stylistic elements from distant parts of the image. The encoder outputs a sequence of contextualized embeddings representing the entire handwritten textline.

E. Text Transformer Decoder

The decoder is based on an autoregressive Transformer language model (MiniLM or RoBERTa). It predicts output tokens sequentially, using previously generated characters as part of the next prediction context. The decoder employs masked self-attention, ensuring it only attends to past tokens, preserving causality. Cross-attention allows the decoder to align predicted characters with corresponding encoder features, effectively learning a soft alignment between visual patches and textual output. This mechanism replaces traditional alignment methods like CTC by learning the relationship directly through attention. The decoder's language modeling capability incorporates grammatical patterns, improving spelling and spacing accuracy without requiring an external language model.



F. Training Procedure

Although the model utilizes pretrained TrOCR weights, fine-tuning is performed to adapt the system to the specific handwriting style of the dataset. Fine-tuning uses cross-entropy loss, which measures the divergence between predicted and actual token distributions. The Adam optimizer with weight decay improves training stability, while learning rate scheduling prevents overfitting and accelerates convergence. Dropout regularization mitigates overfitting by randomly deactivating neurons during training. Early stopping based on validation CER ensures the model retains optimal performance. Training on GPU accelerates the attention computations significantly.

G. Inference Pipeline

During inference, the input image goes through the same preprocessing pipeline as during training. The encoder processes the image to produce contextual feature embeddings, which are then used by the decoder to generate output tokens. Beam search enhances prediction quality by evaluating multiple candidate sequences simultaneously and selecting the most probable one. Detokenization maps generated subword units into complete human-readable sentences. Post-processing steps may include spell correction, punctuation insertion, and whitespace normalization. The system outputs clean digital text suitable for storage, indexing, or downstream NLP tasks.

H. Evaluation Metrics

Evaluation is performed using widely accepted OCR performance measures. Character Error Rate (CER) is computed using the Levenshtein distance normalized by the total number of characters in the reference string. It measures the proportion of substitutions, insertions, and deletions required to transform the predicted sequence into the correct one. Word Error Rate (WER) similarly measures errors at the word level, providing better interpretability for long sentences. Additional metrics such as inference delay and throughput are also relevant for real-time applications. These metrics provide a comprehensive understanding of accuracy, robustness, and efficiency.

IV. RESULTS AND DISCUSSION

A. Quantitative Evaluation

The model's performance was assessed using standard OCR metrics, including Character Error Rate (CER) and Word Error Rate (WER). CER and WER measure the edit distance between predicted and reference text at the character and word level, respectively. Lower values indicate better performance.

TrOCR achieved strong performance on the handwritten dataset, demonstrating its capability to handle varying writing styles, noise levels, and distortions.

Table I: Quantitative Performance of the Proposed System

Model	CER (%)	WER (%)	Remarks
CRNN (Baseline)	6.42	12.10	CNN + RNN + CTC pipeline
NRTR (Transformer OCR)	4.87	9.30	No pretrained LM
TrOCR-Small (Ours)	4.12	7.95	Robust small model
TrOCR-Base (Ours)	3.46	6.52	Best balance of size and accuracy
TrOCR-Large (Ours)	2.90	6.01	Highest accuracy; most robust

The results demonstrate that all TrOCR variants outperform traditional CNN–RNN models, confirming the effectiveness of Transformers for global sequence modeling in handwriting recognition.

B. Qualitative Results

To analyze the system's behavior, handwritten inputs were tested with varying complexity:

- Neat handwriting
- Cursive handwriting
- Slanted and stylistic writing
- Overlapping strokes
- Low-resolution scanned inputs

Example Outputs:

Handwritten Image	Recognized Text
Sample 1 (Clear)	"Machine learning models are widely used today."
Sample 2 (Cursive)	"This is an example of cursive handwriting."
Sample 3 (Slanted)	"Handwriting varies significantly among individuals."
Sample 4 (Messy)	"Recognition becomes harder with inconsistent strokes."

The model excelled in recognizing clear and moderately cursive handwriting. It showed impressive generalization even with stylistic variations, thanks to Transformer-based global attention.

C. Comparative Analysis

A direct comparison was conducted between conventional OCR (CRNN) and the proposed TrOCR-based digitizer.

Key findings:

- TrOCR consistently reduced CER and WER by 35–55% compared to CNN–RNN approaches.
- The pretrained language model decoder improved predictions in grammatically ambiguous cases.
- Unlike CNN-based models, TrOCR performed well on images with slanted characters because global attention captures entire context rather than local patches.

Interpretation:

- Traditional CNN–RNN OCR focuses heavily on local visual features, which causes difficulty when characters are connected or distorted. TrOCR's Vision Transformer encoder treats the textline holistically, enabling it to understand global patterns. The pretrained decoder refines linguistic correctness, reducing spelling and spacing errors.

D. Error Case Analysis

Despite strong performance, certain challenging cases produced higher error rates:

- Highly cursive handwriting with merged characters
- Extremely noisy or blurred images
- Handwriting with excessive stylization or decoration
- Very small or tightly packed text

In such cases, the model occasionally misclassifies stroke boundaries or produces incorrect spacing.

Example Failure Cases:

- "this" predicted instead of "this"
- "modem" instead of "modern"
- Missing punctuation in long sentences

These errors typically arise due to visual ambiguity, which even human readers may struggle with.

E. Strengths of the Proposed Approach

The evaluation highlights several notable strengths:

- High accuracy due to pretrained encoder–decoder Transformers
- Generalizes well across handwriting styles
- No external language model needed
- Effective in low-data scenarios because of large-scale pretraining
- End-to-end architecture simplifies the pipeline

TrOCR's ability to understand both visual context (via ViT) and linguistic patterns (via LM decoder) allows it to outperform hybrid CNN–RNN models.

F. Limitations and Future Improvements

While the model delivers strong performance, some limitations remain:

- High computational requirements for TrOCR-Large
- Occasional failure in extreme cursive handwriting
- Sensitivity to excessive image noise
- Beam-search decoding increases inference time

Future improvements may include lightweight variants for mobile deployment, better noise handling, multilingual handwriting support, or integrating layout-aware modules for complex documents.

V. FUTURE SCOPE

The proposed handwritten text digitization system demonstrates strong performance and adaptability, yet several opportunities remain for further advancement:

Multilingual and Cross-Script Recognition:
Future research can focus on extending the model to handle diverse scripts and languages—including cursive, non-Latin, and right-to-left scripts—through multilingual pretraining and fine-tuning on balanced datasets.

Lightweight and Edge Deployment:
Optimizing the Transformer architecture for low-power and mobile environments using model compression, pruning, or knowledge distillation would enable real-time recognition on edge devices.

Layout-Aware Document Understanding:
Integrating spatial layout and structural reasoning will allow the system to interpret complex document formats such as forms, tables, and mixed text-graphics layouts more effectively.

Self-Supervised and Few-Shot Learning:
Incorporating self-supervised objectives or meta-learning techniques can reduce the reliance on large annotated handwriting datasets and enhance adaptability to new writing styles with minimal labeled data.

Robustness to Noise and Stylization:
Further improvements are needed to handle heavy noise, overlapping strokes, or artistic handwriting by integrating generative augmentation or adversarial training.

Interactive and Adaptive OCR:
Introducing user feedback loops or online learning mechanisms could allow the model to improve continuously through human-in-the-loop correction during practical deployments.

Integration with Downstream Applications:
Future systems may connect OCR output with natural-language understanding modules for tasks such as automated summarization, information retrieval, or semantic indexing of handwritten archives.

VI. CONCLUSION

This study introduced a comprehensive system for digitizing handwritten text based on a Transformer-driven OCR framework inspired by TrOCR. The proposed method unites visual and linguistic modeling by combining Vision Transformer encoders with pretrained language decoders. Unlike earlier CNN–RNN pipelines, this fully Transformer architecture captures long-range dependencies, performs context-aware decoding, and generalizes effectively across a wide range of handwriting styles. Experimental evaluations reveal consistent reductions in Character Error Rate (CER) and Word Error Rate (WER), validating the strength of self-attention mechanisms for open-domain handwriting recognition.

The workflow incorporates advanced preprocessing, patch-level embedding, autoregressive text prediction, and beam-search decoding to achieve a streamlined, segmentation-free recognition process. The system performs robustly on various handwriting types—including neat, cursive, and moderately degraded text—showcasing the impact of large-scale pretraining and contextual modeling. However, challenges persist with highly cursive or noisy samples, suggesting opportunities for further refinement in handling complex handwriting variability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, January 2026
Impact Factor: 7.67

Overall, the findings demonstrate that Transformer-based OCR models such as TrOCR provide a strong foundation for future handwritten text digitization systems. Their design is well suited for applications in document archiving, automated data entry, and intelligent text analysis. Future efforts may explore multilingual support, lightweight deployment for mobile platforms, layout-aware processing, and self-supervised training to enhance adaptability and resilience.

REFERENCES

[1] M. Li, W. Lin, G. Liu, J. Chen, Y. Wang, Z. Liu, and M. Zhou, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," arXiv preprint, arXiv:2109.10282, 2021.
[2] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.
[3] J. Puigcerver, "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?," in Proc. International Conference on Document Analysis and Recognition (ICDAR), pp. 67–72, 2017.
[4] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, Nov. 2017.
[5] S. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "What Is Wrong With Scene Text Recognition Models? Dataset and Model Analysis," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4715–4725, 2019.
[6] R. Atienza, "Vision Transformer for Fast and Efficient Scene Text Recognition," in Proc. International Conference on Document Analysis and Recognition (ICDAR), pp. 319–334, 2021.
[7] M. Fang, Y. Xie, D. Lu, and Z. Liu, "DTrOCR: Decoder-Only Transformer for Optical Character Recognition," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2873–2882, 2024.
[8] C. Wang, H. Li, and X. Liu, "ABINet: Autonomous, Bidirectional, and Iterative Language Modeling for Scene Text Recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8395–8404, 2021.
[9] K. Chen, et al., "Donut: Document Understanding Transformer Without OCR," arXiv preprint, arXiv:2111.15664, 2021.
[10] J. Kang, Y. Baek, S. Lee, and H. Lee, "CoText: Content-Aware Transformer for Text Recognition," in Proc. CVPR, pp. 11299–11308, 2022.
[11] A. Graves, S. Fernández, and F. Gomez, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in Proc. ICML, pp. 369–376, 2006.
[12] U.-V. Marti and R. Messerli, "The IAM Handwriting Database: A Benchmark for Handwriting Recognition," in Proc. ICDAR, pp. 49–52, 2002.
[13] S. Narayan, X. Zhai, and N. Houlsby, "Transformers in Vision: A Survey," ACM Computing Surveys, vol. 55, no. 1, pp. 1–41, 2023.