A Study on OCR-Based Answer Sheet Evaluation Systems

Chris Mathew Joseph *, Devika Vinod *, Dheeraj Krishna *, Haritha P *, Josmi Jose , Sabeena K , Sulaja Sanal

Posted Date: 20 October 2025

doi: 10.20944/preprints202510.1533.v1

Keywords: ocr handwritten transformers trocr

 doi:10.20944/preprints202510.1533.v1

Article

A Study on OCR-Based Answer Sheet Evaluation Systems

Chris Mathew Joseph *, Devika Vinod *, Dheeraj Krishna *, Haritha P *, Josmi Jose, Sabeena K and Sulaja Sanal

Department of Computer Engineering, College of Engineering Chengannur, APJ Abdul Kalam Technological University Kerala, India

* Correspondence: devuvinod2004@gmail.com (D.V.); chrismathewjoseph@hotmail.com (C.M.J.); dheerukrish789@gmail.com (D.K.); harithap648@gmail.com (H.P.)

Abstract

This paper presents a comprehensive literature survey on Optical Character Recognition (OCR)-based systems for the automated evaluation of handwritten answer sheets, emphasizing the integration of modern deep learning and lan- guage understanding techniques. The review consolidates re- search spanning handwritten text recognition (HTR), post-OCR correction, writer-adaptive learning, and multimodal assessment that combines textual, mathematical, and diagrammatic inputs. Recent developments leveraging large vision–language models (VLMs) such as GPT-4V are analyzed for their potential to perform semantic comparison and rubric-aware grading of handwritten solutions. The survey also examines transformer- based architectures, meta-learning frameworks for unseen writer adaptation, and hybrid OCR pipelines integrating CNN, RNN, and attention mechanisms. Key datasets, benchmark results, and performance trends across diverse educational and language settings are discussed. In addition, the paper identifies major challenges such as OCR noise propagation, reasoning inconsis- tencies in large language models, and domain-specific calibration requirements for STEM assessments. By synthesizing current progress and limitations, this work aims to provide a structured foundation for developing future end-to-end, multimodal, and semantically aware AI-driven evaluation systems for scalable and reliable academic assessment. *Index Terms*—OCR, Handwritten Text Recognition, Answer Sheet Evaluation, Semantic Correction, Transformers, CTC, Meta-learning.

Keywords: ocr handwritten transformers trocr

I. INTRODUCTION

Handwritten Answer Sheet Evaluation using OCR has gained considerable attention in recent years as an auto- mated solution to traditional manual grading, which is time- consuming and error-prone. Manual evaluation often suffers from inconsistencies, subjectivity, and scalability issues. With the rapid adoption of deep learning in OCR, researchers have proposed advanced recognition models, error correction methods, and intelligent grading frameworks that aim to reduce human effort while improving fairness and accuracy in academic assessment [1,2,12].

A.Advances in OCR for Handwriting

Recent works show improved HTR by integrating deep encoders, transformers and task-aware losses. IMTLM-Net fuses ViT-like encoding, permutation language modeling and a localization head to handle messy English handwriting and hyphenation; experiments report clear gains on IAM. ViTSTR demonstrates ViT backbones for scene text recognition, re- moving RNN bottlenecks and showing improved throughput and competitive accuracy on STR benchmarks [1,9].

 doi:10.20944/preprints202510.1533.v1

B. Semantic Understanding and Intelligent Grading

Beyond recognition, accurate grading needs semantic com- parison. Recent studies investigate LLMs and multimodal LLMs (GPT-4/GPT-4o) to score handwritten math answers by combining OCR output and reasoning modules. Early results show promising partial automation with human-in-the-loop verification, but sensitivity to OCR errors and reasoning failure modes remains [2,3].

C. Error Correction, Augmentation and Script-specific Pipelines

Post-OCR correction using seq2seq and transformer lan- guage models (DeBERTa, etc.) reduces CER for historical and noisy text; adaptive augmentation or script-specific pipelines (YOLO + EfficientNet + Word2Vec correction for Bangla) have proven practical in improving recognition across low- resource scripts [12], [13].

D. Writer-Adaptation and Meta-Learning

Meta-learning approaches (MetaHTR) adapt a pretrained HTR to a new writer at inference via a few-shot support set and single-step gradient update. This yields notable accuracy gains on unseen writers, addressing per-writer style variation but adding inference-time computation [8].

E. Multimodal Grading (Text + Diagrams)

Answer sheets frequently include diagrams and equations. Work combining text detectors (CRAFT/TrOCR) with object detectors (YOLOv5) and LLM semantic comparators demon- strates end-to-end multimodal grading pipelines for STEM, with accuracy improvements but limited by dataset size and OCR-to-LLM error propagation [11].

F. Trends and Open Challenges

Transformers and permutation/autoregressive sequence models (PARSeq, ABINet) are trending for both scene and handwriting tasks due to global context and iterative correc- tion. However, remaining challenges include handwriting vari- ability, segmentation and layout understanding, multilingual coverage, reliable error-correction at scale, and inference cost for large transformer+LLM stacks [5–7].

II. LITERATURE SURVEY

In their work IMTLM-Net: Improved Multi-task Trans- former for Handwritten English Text Recognition, Zhang et al. [1] proposed an innovative framework that effectively ad- dresses the challenges of reading messy or irregular handwrit- ten English. Their approach combines a Vision Transformer (ViT) encoder to capture rich visual features with a Permu- tation Language Model (PLM) to understand the sequential context of characters, along with a localization mechanism that helps the system focus on individual characters even in crowded or overlapping handwriting. By training the network with hyphenation annotations and multiple learning objectives, including character recognition and sequence modeling, the model becomes more robust to variations in handwriting style, inconsistent spacing, and irregular word breaks. The authors validated their approach on the widely used IAM dataset as well as a custom dataset containing challenging handwrit- ten samples, achieving state-of-the-art recognition accuracy compared to previous CNN- and RNN-based methods. While the results demonstrate impressive performance, the model is computationally demanding and has so far been tested only on English scripts, leaving room for future exploration in multilingual handwritten text recognition. Overall, IMTLM- Net represents a significant step forward in combining vi- sual perception, contextual understanding, and attention-based localization for more accurate and reliable handwritten text recognition.

Caraeni et al. [2] conducted a systematic evaluation of GPT-4 Vision (GPT-4V) for automated grading of handwritten mathematics solutions. Their study explored GPT-4V's ca- pacity to semantically interpret and compare student answers against reference marking schemes. The authors created a benchmark dataset of university-level handwritten math exam scripts, including calculus, linear algebra, and differential equations problems. The evaluation pipeline first used GPT- 4V to extract textual and

 doi:10.20944/preprints202510.1533.v1

symbolic information directly from scanned responses, followed by semantic similarity analysis to match student reasoning with model solutions. Results demon- strated that GPT-4V achieved high alignment with human grading in well-structured problems and short derivations, with over 85% consistency on direct-answer questions. However, the model exhibited reasoning instability when faced with multi-step derivations or diagrams, leading to inconsistent par- tial credit allocation. Additionally, OCR preprocessing noise and handwriting variations significantly affected performance, suggesting that hybrid human–AI workflows remain necessary for high-stakes evaluation.

Liu et al. [3] introduced an AI-assisted framework for grading short handwritten mathematics answers by integrating optical character recognition (OCR) with transformer-based semantic scoring models. Their pipeline used Tesseract OCR to digitize handwriting from scanned exam sheets, and then fed the recognized text into a fine-tuned transformer model trained on annotated scoring rubrics. The system performed semantic comparison between student responses and standard solutions using contextual embeddings rather than simple keyword matching. Experiments on a dataset of 3,000 un- dergraduate math answers showed that the system achieved above 90% correlation with human graders in short-answer tasks. The model also reduced grading time by approximately 60% compared to manual evaluation. However, it struggled with ambiguous phrasing, mathematical symbols, and multi- line derivations. The authors noted that rubric-aware prompt engineering and domain-specific fine-tuning were critical to maintain reliability, highlighting the importance of human oversight in high-complexity problems.

Kortemeyer [4] explored AI-assisted grading of handwritten thermodynamics exams to reduce instructor workload in large engineering classes. The workflow began with OCR-based digitization of handwritten responses using Google Vision API, followed by automated content extraction and semantic scoring using transformer-based text analysis. Multiple grad- ing strategies were tested, including fully automated, semi- automated, and AI-assisted human verification modes. Ex- perimental results from 150 thermodynamics exams revealed that AI-assisted workflows achieved a 40–50% reduction in grading time while maintaining grading consistency compara- ble to human-only evaluation. However, the study highlighted OCR accuracy as a critical bottleneck, as even minor symbol misinterpretations (e.g., “T” vs “τ ”) propagated into incorrect scores. The most effective approach was a hybrid model where AI suggested initial grades and humans verified ambiguous cases. The paper concluded that while AI can significantly aid grading, full automation remains limited by OCR robustness and dataset variability.

AlKendi et al. [5] presented a comprehensive survey of re- cent progress in handwritten text recognition (HTR), covering both offline and online recognition paradigms. The review cat- egorized research developments across preprocessing, feature extraction, augmentation, and deep learning architectures such as CNNs, RNNs, and transformers. The authors emphasized the shift from handcrafted feature-based systems to end-to- end deep learning approaches, particularly transformer-based sequence models. The survey analyzed benchmark perfor- mance on standard datasets including IAM, RIMES, and Bentham, identifying trends such as the growing dominance of self-attention mechanisms and cross-lingual pretraining. Key challenges identified included noise sensitivity, handwriting style diversity, and the lack of multilingual datasets. Although

 doi:10.20944/preprints202510.1533.v1

informative, the review lacked original experimental validation and was primarily a synthesis of existing literature. Nonethe- less, it provided valuable insights into how deep architectures are redefining modern HTR research.

Bautista and Atienza [6] introduced PARSeq (Permuted Autoregressive Sequence Model), an innovative model for scene text recognition (STR). PARSeq reformulated text recog- nition as a permutation language modeling task, allowing the system to predict characters in multiple orders rather than a single left-to-right sequence. This permutation-based learning significantly improved robustness to irregular text layouts and partial occlusions. The architecture combined a transformer encoder–decoder with autoregressive training, leading to stronger generalization and reduced overfitting to sequential bias. Evaluations on standard STR benchmarks such as IIIT5K, SVT, and ICDAR2013 demonstrated state- of-the-art accuracy, surpassing prior models like CRNN and TRBA. However, the model was computationally intensive due to multi-order predictions and primarily targeted scene text, making it less suitable for cursive handwriting or educational answer scripts. Nevertheless, its sequence permutation strategy has influenced subsequent HTR transformer research.

Fang et al. [7] proposed ABINet (Augmented Bidirectional Iterative Network), a hybrid model that couples visual en- coding with language modeling for text recognition. The sys- tem performs iterative refinement, where the language model reinterprets visual outputs in multiple passes—essentially "re- reading" text to correct initial recognition errors. The visual encoder is based on a CNN backbone, while the language model employs bidirectional transformers that enhance con- text awareness. Evaluations across ICDAR2013, SVT, and IIIT5K datasets showed substantial improvements in accuracy, particularly under noisy or distorted conditions. The model outperformed existing CNN–RNN hybrids but incurred higher computational costs and slower inference times due to the multi-pass mechanism. Although primarily designed for scene text, its concept of iterative semantic correction inspired later adaptations for handwriting and document OCR systems.

Bhunia et al. [8] tackled the challenge of writer adaptation in HTR using a meta-learning framework called MetaHTR. The approach was designed to generalize recognition models to unseen handwriting styles using few-shot adaptation. By training the model to quickly learn new writing patterns from a small number of samples, MetaHTR achieved improved recog- nition across diverse writers. The system integrated a CNN encoder for feature extraction and a sequence-to-sequence de- coder for transcription. Evaluations on IAM and CVL datasets demonstrated a notable improvement (3–5% WER reduction) compared to standard HTR baselines. However, this adapt- ability came at the cost of higher inference computation and reduced scalability in large-scale datasets with thousands of unique writers. Despite these trade-offs, MetaHTR represented a major step toward personalized HTR systems capable of learning new handwriting styles dynamically.

Atienza [9] developed ViTSTR, one of the earliest Vision Transformer (ViT)-based architectures applied to scene text recognition. The model replaced both convolutional and recur- rent layers with pure transformer blocks, enabling parallelized training and faster inference while preserving high accuracy. ViTSTR's architecture divided input text images into patches, encoded them using positional embeddings, and processed them through transformer layers for sequence prediction. Evaluations on STR datasets including IIIT5K, ICDAR2013, and SVT demonstrated competitive accuracy and robustness, even without recurrent or convolutional modules. ViTSTR also exhibited better training efficiency due to full parallelization. However, its scope was limited to printed and scene text, and it did not address handwritten input variability, limiting its direct applicability to HTR tasks. Nonetheless, ViTSTR helped pave the way for transformer-based handwriting recognition models.

Garrido-Mun˜oz et al. [10] produced a comprehensive sur- vey detailing the evolution of handwritten text recognition technologies. The review spanned traditional feature-based approaches, deep CNN–RNN hybrids, and transformer-based architectures, providing an end-to-end perspective on model evolution. It summarized major public datasets (IAM, RIMES, CVL, Bentham, and READ),

 doi:10.20944/preprints202510.1533.v1

evaluation metrics (CER, WER), and benchmarking trends across the HTR community. The survey identified a paradigm shift toward transformer-based models for end-to-end sequence recognition, emphasizing self- attention's effectiveness in capturing long-range dependen- cies. It also noted ongoing challenges such as multilingual recognition, data scarcity, and computational overhead in large transformer models. While the paper lacked experimental validation, it served as a key reference resource consolidating five years of progress in HTR research.

Patil et al. [11] presented a multimodal AI framework for automated grading of STEM answer sheets. Their approach integrated image analysis, OCR-based text extraction, equation recognition, and semantic reasoning using large language models. The system handled typed and handwritten content, including mathematical expressions, diagrams, and textual ex- planations. Through multimodal fusion, the framework could evaluate both correctness and partial understanding—for in- stance, identifying conceptual errors even when final answers were wrong. Experiments conducted on a dataset of over 5,000 STEM responses achieved human-grade agreement above 92% and reduced grading time by 70%. However, performance varied across subjects (physics, math, chemistry) and required domain-specific calibration. Computational demands were high during multimodal fusion, especially for equation-heavy scripts. The study demonstrated the transformative potential of multimodal LLMs in educational technology, bridging vision and language understanding for complex assessments.

Beshirov et al. [12] addressed OCR error correction for Bul- garian historical documents, which often suffer from degraded paper, inconsistent fonts, and archaic spelling. They proposed

a post-OCR correction pipeline combining a context-aware language model with a statistical error detector. The system identified likely OCR errors at character and word levels, then generated corrections using transformer-based contextual embeddings fine-tuned on historical Bulgarian corpora. Unlike conventional correction systems, it preserved original orthog- raphy and formatting, maintaining the authenticity of digitized texts. Testing on the Bulgarian Historical Text Corpus showed a 15–20% improvement in word-level accuracy over baseline OCR outputs. Despite its success, the system required domain- specific linguistic resources and was not directly transferable to other languages. This research underscores the significance of post-processing in improving OCR pipelines for cultural heritage preservation.

Maung et al.Maung et al. [13] developed a hybrid recogni- tion model for Bangla handwritten text, integrating YOLO for object detection and a deep CNN for character classification. The pipeline first localized text regions—individual characters or words—using YOLO's bounding box detection, which effectively handled overlapping strokes and non-linear layouts. The segmented text was then passed to a CNN classifier for precise recognition. Experiments on BanglaLekha-Isolated and Ekush datasets demonstrated superior character-level ac- curacy and generalization across handwriting styles compared to CNN-only baselines. The model achieved over 95% ac- curacy on clean samples and remained robust even under moderate noise. However, the two-stage design introduced computational overhead, making real-time processing chal- lenging for large document batches. Still, the hybrid YOLO– CNN approach offered a scalable and accurate framework for digitizing regional scripts, contributing to broader OCR inclusivity for low-resource languages.

III. CONCLUSION

This literature survey explored the recent advancements and emerging methodologies in Optical Character Recogni- tion (OCR)-based handwritten answer sheet evaluation. The review identified that modern Handwritten Text Recognition (HTR) systems have evolved from traditional CNN–RNN hybrids to transformer-driven architectures such as IMTLM- Net, PARSeq, ABINet, and ViTSTR, which leverage self- attention and permutation-based modeling for improved con- textual understanding and error robustness. These models have significantly enhanced recognition accuracy on benchmark datasets like IAM and RIMES, laying the foundation for reliable digitization of handwritten academic content.

 doi:10.20944/preprints202510.1533.v1

Beyond recognition, considerable progress has been made in the integration of semantic analysis and grading automation. Studies employing Large Language Models (LLMs) and Vi- sion–Language Models (VLMs) such as GPT-4V demonstrate that AI can interpret handwritten mathematical and textual re- sponses, perform semantic comparison with reference answers, and assign partial credit in rubric-based assessments. However, current systems still depend on high-quality OCR outputs and controlled input structures; errors at the recognition stage often propagate into semantic grading, underscoring the need for error-tolerant and reasoning-aware pipelines.

The survey also highlighted the role of meta-learning frame- works like MetaHTR, which enable writer-specific adaptation through few-shot learning, improving accuracy across diverse handwriting styles. Similarly, multimodal grading approaches combining text, diagrams, and mathematical equations using integrated OCR and object detection models (e.g., CRAFT, YOLO, TrOCR) have demonstrated promising results in STEM evaluations, achieving human-comparable consistency in pilot studies. Complementary advances in post-OCR text correction, such as language model–based contextual correc- tion and script-specific augmentation, further enhance recog- nition quality across low-resource and multilingual datasets.

Despite these achievements, several open challenges remain. Current transformer-based architectures are computationally intensive and often limited to specific scripts or datasets. Large language models exhibit reasoning instability, inconsistency in partial credit allocation, and limited domain calibration. Moreover, multimodal systems face difficulties in integrating diagrammatic understanding and maintaining alignment be- tween OCR text and visual context. The scarcity of standard- ized datasets for handwritten answer sheets, especially those involving formulas and diagrams, also restricts benchmarking and reproducibility.

Future research should focus on developing unified, end- to-end frameworks that integrate robust OCR, handwriting adaptation, semantic reasoning, and multimodal analysis. Ef- ficient transformer variants, reinforcement learning for rubric- based feedback, and self-supervised pretraining on diverse handwritten corpora could improve both generalization and efficiency. Human–AI collaborative grading workflows, where models provide interpretable justifications for assigned scores, are likely to represent the most practical near-term solution. Ultimately, the convergence of OCR, vision–language model- ing, and educational AI promises to make automated handwrit- ten evaluation systems scalable, transparent, and pedagogically reliable in real-world academic environments.

REFERENCES

1. Q. Zhang, F. Liu, W. Song, “IMTLM-Net: Im- proved Multi-task Transformer based on Localization Mechanism Network for Handwritten English Text Recognition,” *Complex & Intelligent Systems*, 2025. doi:10.1007/s40747-024-01713-8.
2. A. Caraeni, A. Scarlatos, A. Lan, “Evaluating GPT-
3. 4 at Grading Handwritten Solutions in Math Exams,” arXiv:2411.05231, 2024.
4. T. Liu et al., “AI-assisted Automated Short Answer Grading of Handwritten University Level Mathematics Exams,” arXiv:2408.11728, 2024.
5. G. Kortemeyer, “Grading Assistance for a Handwritten Thermodynamics Exam,” *Phys. Rev. Phys. Educ. Res.*, 2024.
6. W. AlKendi, “Advancements and Challenges in Hand- written Text Recognition,” *Journal of Imaging (MDPI)*, 2024.
7. D. Bautista, R. Atienza, “PARSeq: Permuted Autore- gressive Sequence Models for Scene Text Recognition,” ECCV, 2022.
8. S. Fang, H. Xie, Y. Wang, Z. Mao, Y. Zhang, “ABINet: Autonomous, Bidirectional and Iterative Language Mod- eling for Scene Text Recognition,” arXiv:2103.06495, 2021.
9. A. K. Bhunia, S. Ghose, A. Kumar, P. N. Chowdhury, A. Sain, Y.-Z. Song, “MetaHTR: Towards Writer-Adaptive Handwritten Text Recognition,” in *Proc. CVPR*, 2021.

 doi:10.20944/preprints202510.1533.v1

10. R. Atienza, "ViTSTR: Vision Transformer for Fast and Efficient Scene Text Recognition," arXiv:2105.08582, 2021.
11. C. Garrido-Mun˜oz, A. R´ıos-Vila, J. Calvo-Zaragoza, "Handwritten Text Recognition: A Survey," arXiv:2502.08417, 2025.
12. R. Patil, et al., "Automated Assessment of Multimodal Answer Sheets in STEM," arXiv preprint, 2024.
13. A. Beshirov, M. Dobreva, D. I. Dimitrov, M. Hardalov, Koychev, and P. Nakov, "Post-OCR text correction for Bulgarian historical documents," Int. J. Digit. Libr., vol. 26, no. 1, p. 4, 2025.
14. A. T. Maung, S. Salekin, and M. A. Haque, "A hybrid approach to Bangla handwritten OCR: combining YOLO and an advanced CNN," Discover Artificial Intelligence, vol. 5, art. no. 119, 2025.