Undergraduate Thesis Summary

Benchmarking OCR Engines and VLMs for Automated Evaluation of Handwritten Scanned Answer Scripts

Summary of Undergraduate Thesis under the guidance of Dr. Uttam Kumar Sarkar, Associate Professor, Computer Science and Engineering (Artificial Intelligence and Machine Learning) Department

Olivia Chatterjee
olivia.chatterjee@ieee.org

Abstract

Handwritten academic answer scripts pose persistent challenges for optical character recognition due to high variability in handwriting styles, inconsistent layouts, scan artifacts, and the frequent presence of non-textual elements such as mathematical expressions, diagrams, and tables. Despite recent advances in transformer-based OCR architectures and large-scale multimodal vision–language models (VLMs), their reliability on full-length handwritten examination scripts remains insufficiently studied.

This undergraduate research project investigates the performance of classical OCR engines (Tesseract, EasyOCR), transformer-based models (TrOCR), and contemporary VLMs on a curated dataset of real handwritten university answer sheets. A unified evaluation framework is developed using Word Error Rate, Character Error Rate, ROUGE-L, BLEU, and embedding-based semantic similarity measures to assess both transcription accuracy and contextual fidelity. Model performance is analyzed across diverse content types, including continuous prose, mathematical notation, labeled diagrams, and structured tabular layouts.

By systematically benchmarking these models under identical preprocessing and evaluation conditions, the study identifies key limitations of existing OCR and multimodal approaches for handwritten document understanding. The findings provide a reproducible foundation for downstream applications such as grader-assist systems and automated educational assessment, while highlighting the need for controlled and auditable NLP pipelines in high-stakes academic contexts.

Project Status and Outcomes

This work is an ongoing undergraduate thesis scheduled for completion in June 2026. The project has resulted in a curated dataset of handwritten examination scripts, a reproducible evaluation pipeline, and comparative analysis of OCR and vision–language models. The outcomes are intended to inform future work on automated academic assessment systems.

References

[BKL+19] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Donghyun Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4715–4723. IEEE, 2019. Available at https://openaccess.thecvf.com/content_ICCV_2019/papers/Baek_What_Is_Wrong_With_Scene_Text_Recognition_Model_Comparisons_Dataset_ICCV_2019_paper.pdf.

[KSS24] Takeshi Kojima, Shota Saito, and Kazunari Sugiyama. Document understanding with vision–language models: A survey. *ACM Computing Surveys*, 56(2), 2024. Comprehensive survey of vision–language models for document understanding. Available at https://dl.acm.org/doi/10.1145/3637891.

[Lin04] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization*, pages 74–81, 2004. Available at https://aclanthology.org/W04-1013.

[LLC+23] Minghao Li, Tengchao Lv, Lei Cui, Yijia Lu, and Furu Wei. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023. Available at https://ojs.aaai.org/index.php/AAAI/article/view/26623.

[LZCL24] Yang Liu, Haotian Zhang, Wei Chen, and Xiang Li. OCRBench: On the hidden mystery of OCR in large multimodal models, 2024. Available at https://arxiv.org/abs/2305.07895.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002. Available at https://aclanthology.org/P02-1040.

[Smi07] Ray Smith. An overview of the tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633. IEEE, 2007. Foundational work describing the architecture of the Tesseract OCR engine. Available at https://ieeexplore.ieee.org/document/4376991.

[ZKW+20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. Available at https://openreview.net/forum?id=SkeHuCVFDr.