Evaluating Optical Character Recognition Accuracy on Academic Documents

Alex Tan, Mei Ling Wong
Department of Computer Science, Universiti Malaya
Faculty of Information Technology, Monash University Malaysia

Abstract

Optical Character Recognition (OCR) plays a critical role in digitizing academic literature, enabling search, analysis, and knowledge extraction from scanned documents. However, academic papers present unique challenges to OCR systems due to complex layouts, multi-column formats, mathematical symbols, and varied font styles. This study evaluates the performance of a modern OCR pipeline on single-page academic documents. We analyze character-level and word-level accuracy using manually verified ground truth. The results show that while OCR achieves high accuracy on plain text sections, errors frequently occur in headings, references, and technical terms. These findings highlight the need for layout-aware post-processing techniques to improve OCR reliability for scholarly documents.

Keywords: Optical Character Recognition, Academic Documents, Text Extraction, Evaluation, Document Analysis

Introduction

The digitization of academic documents has become increasingly important as research outputs continue to grow in volume. OCR systems are widely used to convert scanned papers into machine-readable text, supporting downstream tasks such as indexing, summarization, and information retrieval. Despite recent advances in deep learning-based OCR models, accurately recognizing text in academic papers remains challenging.

Methodology

We evaluated OCR performance using a single-page academic paper scanned at 300 DPI. The document was processed using a layout-aware OCR system that detects text regions before recognition. Ground truth text was manually created for evaluation. Accuracy was measured using Character Error Rate (CER) and Word Error Rate (WER), which are standard metrics in OCR evaluation.

Results and Discussion

The OCR system achieved an average CER of 2.8% and a WER of 6.1% on the evaluated page. Most errors occurred in section titles and author affiliations, where spacing and capitalization were inconsistently recognized. Reference entries also showed higher error rates due to punctuation and abbreviated names. These results suggest that post-processing and language-aware correction could significantly improve OCR output quality.

Conclusion

This study demonstrates that modern OCR systems perform well on standard academic text but still struggle with structural elements and specialized formatting. Future work should focus on integrating document layout understanding and domain-specific correction models to enhance OCR accuracy for academic papers.

References

[1] Smith, R. “An Overview of the Tesseract OCR Engine.” Proceedings of ICDAR, 2007.
[2] Baek, J. et al. “Character Region Awareness for Text Detection.” CVPR, 2019.
[3] Deng, Y., and Liu, X. “Evaluation Metrics for OCR Systems.” Journal of Document Analysis, 2021.