Optical Character Recognition for Information Extraction from Handwritten Assault and Violence Reports

Mohammad Sabik Irbaz, Md. Mushfiqur Rahman, Katherine Scafide, David Lattanzi, Janusz Wojtusiak, Kevin Lybarger
George Mason University, Fairfax, VA, USA

Introduction

Despite the widespread adoption of electronic health records (EHR), many reports of forensic nursing evaluations continue to be documented in handwritten form due to the need to separate sensitive forensic reports from standard EHR data. Additionally, specific patient examination procedures often require the creation of documentation while collecting samples and photographic evidence. These reports contain vital information, such as injury mechanisms, anatomical injury locations, and contextual details regarding the abuse.

Optical Character Recognition (OCR) technologies offer a means to transform these handwritten documents into machine-readable text, facilitating the extraction of critical information in a manner similar to that employed with clinical documents like physician notes and radiology reports.

Methods

We evaluated seven off-the-shelf OCR tools, including commercial platforms such as the Google Cloud Vision API and multimodal large language models (LLMs) like Meta's open-access Llama-3.2-11B-Vision. Our evaluation focused on three types of synthetic case reports: domestic violence, physical abuse, and sexual assault. Transcription performance was assessed using Word Error Rate (WER), while the accuracy of preserving key injury-related information—specifically injury mechanisms and anatomical locations—was also analyzed.

Results

Google Cloud Vision API demonstrated the lowest WER, averaging 0.17 across all report types. Llama-3.2-11BVision performed comparably in a zero-shot setting, achieving an average WER of 0.22, suggesting room for improvement with fine-tuning. Analysis of key information preservation revealed high accuracy, with location details preserved at 100% and injury mechanism preservation rates ranging from 71% to 100%.

Conclusion

Our findings support the use of OCR for converting handwritten abuse reports into text, with Llama-3.2-11B-Vision emerging as a promising, locally-deployable alternative to commercial solutions. The high accuracy in preserving critical injury information highlights the potential for automated data extraction from these reports, paving the way for enhanced documentation, analysis, and patient care in assault and violence cases.