Contribution ID: **324** Type: **not specified**

Exploring OCR: Combining Open-Source Engines for Improved Document Digitization

Thursday 18 September 2025 12:25 (15 minutes)

Document digitization involves converting physical documents into editable digital text, a process that offers significant benefits such as preserving archives, enabling remote access, and simplifying content modification. Optical Character Recognition (OCR) technologies facilitate this transformation by extracting text from scanned or photographed document images. However, OCR accuracy can be hindered by the wide variety of document layouts and conditions, including issues like faded text and uneven lighting. In this study, we investigate the potential of combining multiple open-source OCR engines to improve digitization accuracy, focusing on the Tesseract and EasyOCR engines. We developed a testing pipeline and conducted experiments targeting challenging scenarios for character recognition. Our results demonstrate that integrating outputs from both engines can enhance performance, highlighting their complementary strengths and the promise of ensemble approaches for more reliable document digitization.

Author: Mr PANDELICĂ, Mihai-Lucian (Universitatea Politehnica Bucuresti)

Co-authors: VLĂSCEANU, Giorgiana (University Politehnica of Bucharest); TURCANU, Mihai (Technical University of Moldova)

Presenter: Mr PANDELICĂ, Mihai-Lucian (Universitatea Politehnica Bucuresti)

Session Classification: Open Source Education and Research

Track Classification: Open Source and GNU in Education and Research