World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:19, No:04, 2025

Optical Character Recognition of Handwritten Hebrew Documents

Authors : Tomer Kakou, Tal BoAhron, Natalia Vanetik

Abstract : As digital transformation accelerates, the demand for processing handwritten text images has significantly increased. The ability to convert handwritten text into a computer-readable format is crucial for enabling efficient searching, storage, editing, and interpretation, even for challenging handwriting. Organizations that need to accurately and efficiently digitize handwritten records, like educational institutions, would find this capacity very useful. Even while optical character recognition (OCR) for printed text has advanced, handwritten writing has additional difficulties that are especially challenging in low-resource languages like Hebrew. To bridge this gap, are developing an innovative method for Hebrew handwritten OCR that leverages both traditional and cutting-edge techniques. it approach integrates a newly curated dataset of handwritten Hebrew text images with an existing dataset of Hebrew texts called HDD for more precise image classification. The core of our methodology involves a multi-step process that first enhances image resolution to improve overall quality, followed by the extraction of individual character images using advanced image processing tools like OpenCV. Each character image is then classified into one of 27 classes, corresponding to the letters of the Hebrew alphabet. This step is crucial, as it enables the system to recognize individual characters, which are then reassembled into coherent text sequences. To achieve accurate recognition, it utilize deep learning models, including Vision Transformer (ViT) and ResNet-50, for multi-class image classification. These models have shown promising results in the domain of visual recognition tasks, and their adaptation to handwritten Hebrew text offers significant potential for improving OCR performance. Context-based word recognition will be used in the next stage, when large language models (LLMs) are used to provide contextual corrections. This increases the output's overall accuracy by resolving errors and ambiguities that occur during the character recognition process. For model evaluation, we employ several performance metrics, including Character Error Rate (CER), Word Error Rate (WER), and Normalized Levenshtein Distance (NLD). NLD has proven to be the most reliable metric in our case, as it accounts for small errors and typographical variations, making it particularly suited for evaluating OCR. This project's ultimate objective is to create a reliable, comprehensive end-to-end solution for handwritten Hebrew text digitization that may be used in a variety of contexts. Additionally, our method strives to achieve high accuracy even in situations with handwriting errors or deteriorated text by incorporating context-based adjustments, which makes it a useful tool for real-world applications.

Keywords : hebrew, image classification, low-resource languages, OCR
Conference Title : ICPRCV 2025 : International Conference on Pattern Recognition and Computer Vision
Conference Location : Jerusalem, Israel
Conference Dates : April 24-25, 2025