World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:11, 2020

A Conglomerate of Multiple Optical Character Recognition Table Detection and Extraction

Authors : Smita Pallavi, Raj Ratn Pranesh, Sumit Kumar

Abstract : Information representation as tables is compact and concise method that eases searching, indexing, and storage requirements. Extracting and cloning tables from parsable documents is easier and widely used; however, industry still faces challenges in detecting and extracting tables from OCR (Optical Character Recognition) documents or images. This paper proposes an algorithm that detects and extracts multiple tables from OCR document. The algorithm uses a combination of image processing techniques, text recognition, and procedural coding to identify distinct tables in the same image and map the text to appropriate the corresponding cell in dataframe, which can be stored as comma-separated values, database, excel, and multiple other usable formats.

Keywords : table extraction, optical character recognition, image processing, text extraction, morphological transformation
Conference Title : ICDAR 2020 : International Conference on Document Analysis and Recognition
Conference Location : Jerusalem, Israel
Conference Dates : November 26-27, 2020