ABSTRACT

Printed documents are still the choice of several industries to store company data such as invoices, receipts, and other printed documents. This creates problems when it is necessary to digitize the data. Therefore, we need a system that can convert the image of a printed document into a string so that the data does not need to be entered into the computer manually. Currently, the technology that can identify letters in an image is OCR engine which has been programmed to perform segmentation, feature extraction, classification, training, and recognition. One of the OCR engines that has the highest accuracy (96.38%) with the fastest processing time (4.60 seconds) is Tesseract. However, Tesseract's accuracy depends on image quality and noise, so additional image processing is required. Therefore, in this project, document scanner was designed using OCR Tesseract with image processing stages: grayscaling, unsharp masking, Otsu thresholding, and dilation with the OpenCV library. After doing some testing, the scanner can recognize writing on documents with 2.58% error for recognizing words, 3.5% error for recognizing words in a sentence, 10.5% error for recognizing words in paragraphs, and 9.5% error for recognize words in receipt documents. These results are for font size 16 with Arial, Calibri, Times New Roman, Dot Matrix, and Fake Receipt fonts.

Keywords: Optical Character Recognition, Tesseract, image processing, OpenCV, scanner