Abstract

The images from the African articles dataset presented challenges to the Optical Character Recognition (OCR) tool. Despite successful binerisation in the Image Processing step of the pipeline, noise remained in the foreground of the images. This noise caused the OCR tool to misinterpret the text from the images and thus needed removal from the foreground. The technique involved the application of the Maximally Stable Extremal Region (MSER) algorithm, borrowed from Scene-Text Detection, and supervised machine learning classifiers. The algorithm creates regions from the foreground elements. Regions are classifiable into noise and characters based on the characteristics of their shapes. Classifiers were trained to recognise noise and characters. The technique is useful for a researcher wanting to process and analyse the large dataset. They could semi-automate the foreground noise-removal process using this technique. This would allow for better quality OCR output, for use in the Text Analysis step of the pipeline. Better OCR quality means less compromises would be required at the Text Analysis step. These concessions can lead to false results when searching noisy text. Fewer compromises means simpler, less error-prone analysis and more trustworthy results. The technique was tested against specifically selected images from the dataset which exhibited noise. It involved a number of steps. Training regions were selected and manually classified. After training and running many classifiers, the highest performing classifier was selected. The classifier categorised regions from all images. New images were created by removing noise regions from the original images. To discover whether an improvement in the OCR output was achieved, a text comparison was conducted. OCR text was generated from both the original and processed images. The two outputs of each image were compared for similarity against the test text. The test text was a manually created version of the expected OCR output per image. The similarity test for both original and processed images produced a score. A change in the similarity score indicated whether the technique had successfully removed noise or not. The test results showed that blotches in the foreground could be removed, and OCR output improved. Bleed-through and page fold noise was not removable. For images affected by noise blotches, this technique can be applied and hence less concessions will be needed when processing the text generated from those images.