ABSTRACT

Optical character recognition (OCR) method has been used in converting printed text into editable text. OCR is very useful and popular method in various applications. Accuracy of OCR can be dependent on text preprocessing and segmentation algorithms. Sometimes it is difficult to retrieve text from the image because of different size, style, orientation, complex background of image etc.The process of extracting information from a digital copy of invoice can be a tricky task. There are various tools that are available in the market that can be used to perform this task. However there are many factors due to which most of the people want to solve this problem using Open Source Libraries.The goal of text extraction from an image is to develop an Optical Capture Recognition (OCR) and Natural language processing (NLP). Content reports, pictures that have been captured by webcam and the recognized data is sent further for converting it into speech using google text to Speech for a blind person. The purpose of this system is to recognize text in scanned text documents, text images or any picture taken by a webcam to reuse it later. This system will permit its clients to perform numerous activities in a couple of moments minutes. This project is mainly related to image processing to recognize characters in an image. Considering text element, image can be classified as: (i) document image and (ii) scene text image. Currently available OCR algorithms have been developed to deal with document images. OCR algorithm expects the input to be black and white image and relatively clean and well structured. In these conditions, the direct application of an OCR on a scene text image would result in an almost certain failure.