Information Extraction with less error from Image of Document: a step towards Perfect OCR

Anonymous ACL-IJCNLP submission

Abstract

Optical Character Recognition or OCR is a great tool to replace manual data entry in many application. Accurate data extraction is main concern in OCR because perfect data extraction from image of text is still difficult. This paper is proposing a way to ensure less error in extracting data from an image of text and overcome challenges in OCR.

1 Introduction

Optical Character Recognition or OCR is electrical transformation of image of text to machine coded text.In this era of AI, ML, NLP, performing OCR with 100% accuracy is still difficult. There are some challenges in OCR to extract machine coded text with no error. Imbalance of light and shadow in image, background noise or pattern, text blended with background, intended data extraction from a stack of text, these are main challenges in OCR. This paper intends to overcome these challenges.The proposed process has been divided into two parts. At first, converting the input image into OCR perfect image. Secondly, extracting intended data from OCR text with possible less error as all data is not necessary.

2 Proposed idea

2.1 OCR Perfect Image

OCR Perfect Image refers to an image that is ready for OCR and any OCR engine can get machine coded text with minimum error possible. To get OCR perfect Image, some issues need to solve before performing OCR.

1. Imbalance of light and shadow can be solved with LAB color space. Altering L channel in this color space can equalize amount of lighting in the image. Moreover, balancing whiteness may also help.

2. Image of NID, Passport have background color and pattern which hinders text extraction. Image Enhancement can be applied to overcome this problem. By altering sharpness, contrast, brightness with predefined parameter can make the background less dominant. It also helps to make text more visible in the image.

2.2 Extract data from OCR text with less error

Not all text from OCR output is necessary. Sometimes selected data are needed from a whole document. It is another challenge to grab right one. 1. Data with some pattern like Date of Birth, NID no, telephone no etc, can be extracted with regular expression. 2. Data with no pattern like name, address, are difficult to distinguish them from stack of text. Error correction in these type of word is very difficult. For Bengali language, spelling error can be checked as it follows some rules.

Moreover, Noisy Channel Model can detect and correct spelling error without dictionary. It will solve the name error portion of OCR. By generating more than 1 version of input image and taking output with highest accuracy can lead to even more accurate OCR performance

3 Conclusion

The proposed model is all about increasing accuracy of OCR by overcoming different challenges in this field. This concept may lead to one step towards perfect OCR.

References

[1] https://en.wikipedia.org/wiki/Optical$_c haracter_r ecognition$

[2] https://www.pyimagesearch.com/category/optical-character-recognition-ocr

[3] https://guides.library.illinois.edu/OCR