OCR (Optical Character Recognition)

17.11.2018

Shubham Kumar Sunwalka

Overview

This project deals with development of OCR Algorithms to retrieve meaningful information from Official Documents eg PAN Card,etc. In this module, I have built the python modules for information retrieval from PAN Card

TASKS

1. SKEW CORRECTION Module (works for -45degree to +45 degree range).
2. BINARIZATION of image from skimage
3. Input image >>Skew Correction >>Corrected image >>>>OCR output
4. Gain Familiarity with opencv,numpy,scifit,skimage,matplotlib,etc working libraries.
5. WORD DETECTION method line by line way using the concept of bounding boxes
6. Learnt to use EAST Text Detector
7. Learn the use of TESSERACT OCR Engine
8. Get hands on Dilation,erosion and other preprocessing methods
9. Retrieve meaningful information from PAN CARD i.e. NAME,FATHER's NAME,D.O.B,PAN number.
10. Study about Natural Language Toolkit

11. Generate DATASET of PAN CARD Document

Results

1. Created a SKEW CORRECTION named module in python
2. Word detection using pytesseract and EAST text detector in folder ocr-PAN card document
3. Retrieve meaningful information PAN CARD stored in folder named as 14NOVEMBER2018.

(SKEW Correction)

My first module was angle correction or the SKEW correction module in which i have used the image and have Globally thresholded the image after applying preprocessing steps .Here in this module i have used the concept of bounding boxes in order to find that whether the image is tilted or not.

This SKEW CORRECTION module runs for rotation which is in the range from **-45degree to +45degree.**Later in the coming modules I have used niblack and sauvola threshholding

.

(Binarisation)

In this I learnt the concept of binarisation as a preprocessing step.

(Word detection and pytesseract)

Here i used the EAST Text detector for the WORD Segmentation.EAST text detector is a pretrained model for word segmentation.

Provided below is the algorithmic approach adapted EAST (Efficient and Accurate scene Text Detector)

1. Read image
2. Resize it (multiples of 320)
3. Extract output feature maps of 2 Layers-
4. -The first layer is our output sigmoid activation which gives us the probability of a region containing text or not.

5. -The second layer is the output feature map that represents the “geometry” of the image — we’ll be able to use this geometry to derive the bounding box coordinates of the text in the input
6. Load EAST detector
7. Binarising and thresholding
8. Construct BLOB
9. Set BLOB as input
10. Supply Layer names to get list as required
11. Extract nRows,nColumns from Scores volume
12. Loop over nRows and nColumns and take text area whosr probability >0.5
13. Scale up feature map obtained so far
14. Calculate image
15. Calculate dimensions of bounding box
16. Add coordinates and associated probabilities to lists
17. Rescale image and assign bounding box
18. Use non max supression to remove multiple bounding boxes for same text (select one with max probability)
19. Run Pre trained model for detection(oem 3 and psm 11 mode tested)
20. OUTPUT

(TESSERACT-OCR engine)

Here in this module i used ocr engine which can extract texts from image.

(PAN CARD information retrieval)

1. Read the image
2. Skew correction(for skew module i have used global threshold but for later processes i have used niblack and sauvola thresholding in the code)
3. Thresholded image is then given to the pytesseract engine to extract texts from image.
4. Then the text is splitted up into array in form of strings at indexes.
5. Fetch the nameline usingg named.csv file and assume the next indexed value to be fathers name.
6. Using dparser function in difflib library we can identify the dobline from the extracted text
7. Still the logic to detect PAN no is still in progress

PS: I have started working on Contour based approach as it is possible that above algo might not properly on certain images.