Saudi Electricity Company Digital Meter Reading Project (Phase I)

Progress Report

Character Recognition of electronic component

OCR

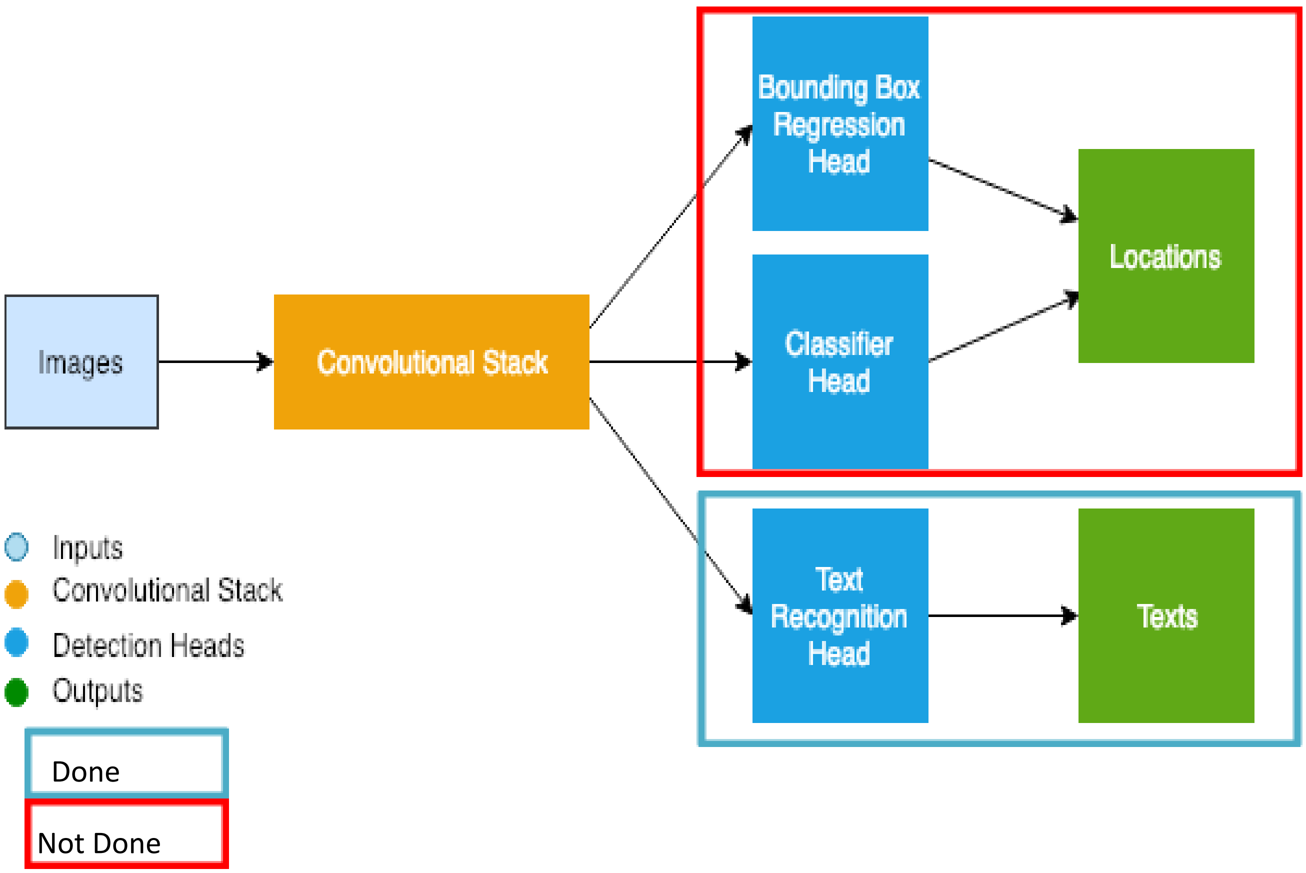
The challenge of extracting text from images of documents has traditionally been referred to as Optical Character Recognition (OCR) and has been the focus of much research. When documents are clearly laid out and have global structure (for example, a business letter), existing tools for OCR can perform quite well. A popular open source tool for OCR is the Tesseract Project, which was originally developed by Hewlett-Packard but has been under the care and feeding of Google in recent years. Tesseract provides an easy-to-use interface as well as an accompanying Python client library, and tends to be a go-to tool for OCR-related projects. More recently, cloud service providers are rolling out text detection capabilities alongside their various computer vision offerings. These include GoogleVision, AWS Textract, Azure OCR, and Dropbox, among others. It is an exciting time in the field, as computer vision techniques are becoming widely available to empower many use cases.There are, however, many use cases in what we might call non-traditional OCR where these existing generic solutions are not quite the right fit. An example might be in detecting arbitrary text from images of natural scenes. Problems of this nature are formalized in the COCO-Text challenge, where the goal is to extract text that might be included in road signs, house numbers, advertisements, and so on. Another area that poses similar challenges is in text extraction from images of complex documents. In contrast to documents with a global layout (such as a letter, a page from a book, a column from a newspaper), many types of documents are relatively unstructured in their layout and have text elements scattered throughout (such as receipts, forms, and invoices). Problems like this have been recently formalized in the ICDAR DeTEXT Text Extraction From Biomedical Literature Figures challenge. These images are characterized by complex arrangements of text bodies scattered throughout a document and surrounded by many "distraction" objects. In these images, a primary challenge lies in properly segmenting objects in an image to identify reasonable text blocks. Example images from COCO-Text and ICDAR-DeTEXT are shown below. These regimes of non-traditional OCR pose unique challenges, including background/object separation, multiple scales of object detection, coloration, text orientation, text length diversity, font diversity, distraction objects, and occlusions. The problems posed in non-traditional OCR can be addressed with recent advances in computer vision, particularly within the field of object detection. As we discuss below, powerful methods from the object detection community can be easily adapted to the special case of OCR.

Core Requirements

Some **attributes** of the OCR tasks:

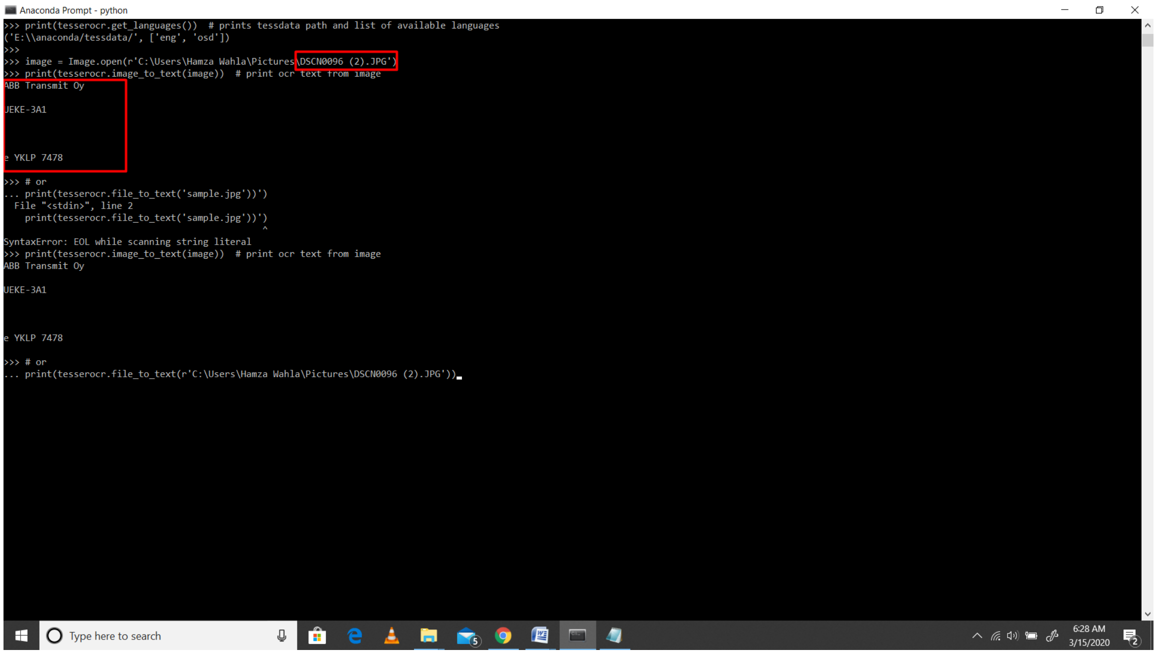
- **Text density**: on an image/written page, text is dense. However, given an image of a street with a single street sign, text is sparse.
- **Structure of text**: text on an image is structured, mostly in strict rows, while text in the wild may be sprinkled everywhere, in different rotations.
- **Fonts**: printed fonts are easier, since they are more structured then the noisy hand-written characters.
- **Character type:** text may come in different language which may be very different from each other. Additionally, the structure of text may be different from numbers, such as house numbers etc.
- **Artifacts**: clearly, outdoor pictures are much noisier than the comfortable scanner.
- **Location**: some tasks include cropped/centered text, while in others, text may be located in random locations in the image.

Methodology



Progress view

We were provided by the images of different electronic devices portraying the IDs Voltages consumed counts etc. We have chosen a data set (1012) out of that we have chosen DSCN0096.jpeg as shown above



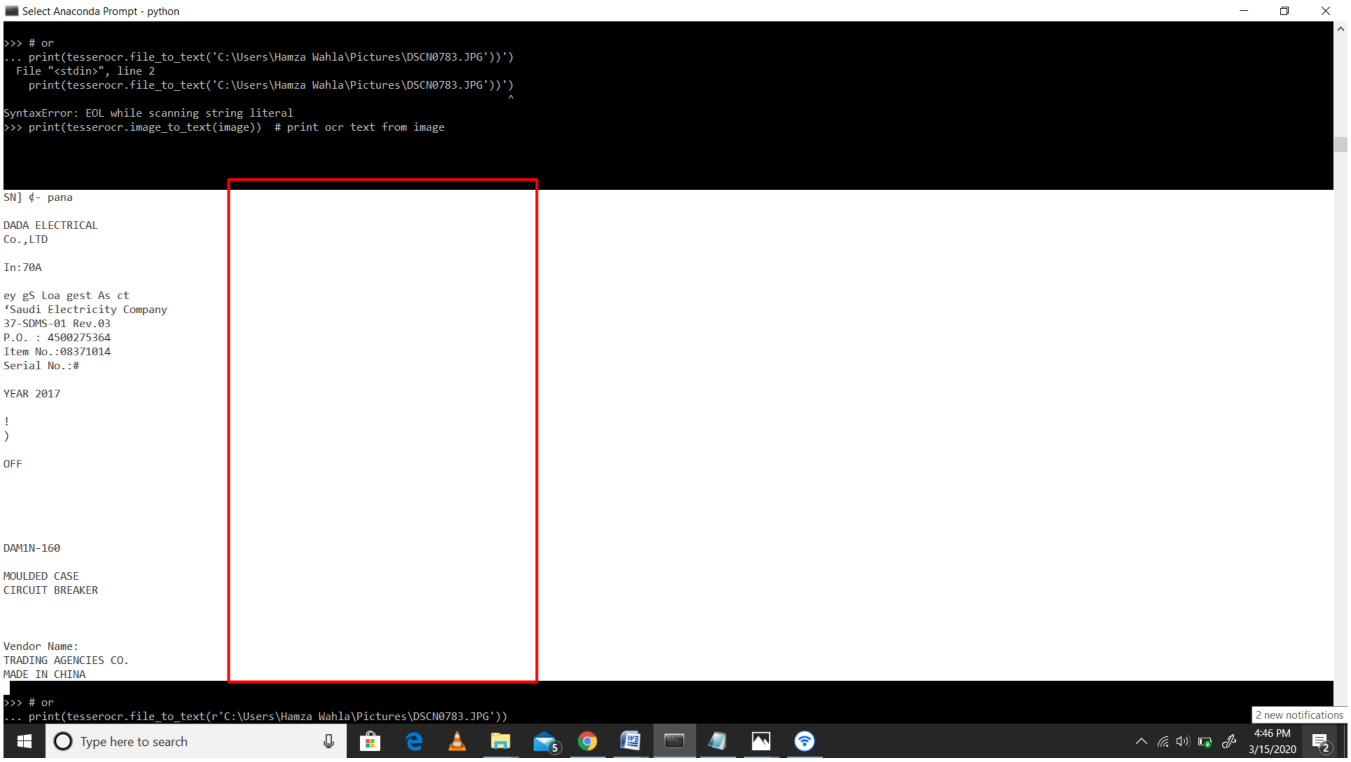
After the implementation of the OCR we have text output as follows in the 2nd screen shot

In this representation we have converted the image characters into that of texts but we haven't set the boundary box around them and we haven't allotted IDs to the word detector and this is why we can't classify the character attributes

Example 2 (circuit breaker)

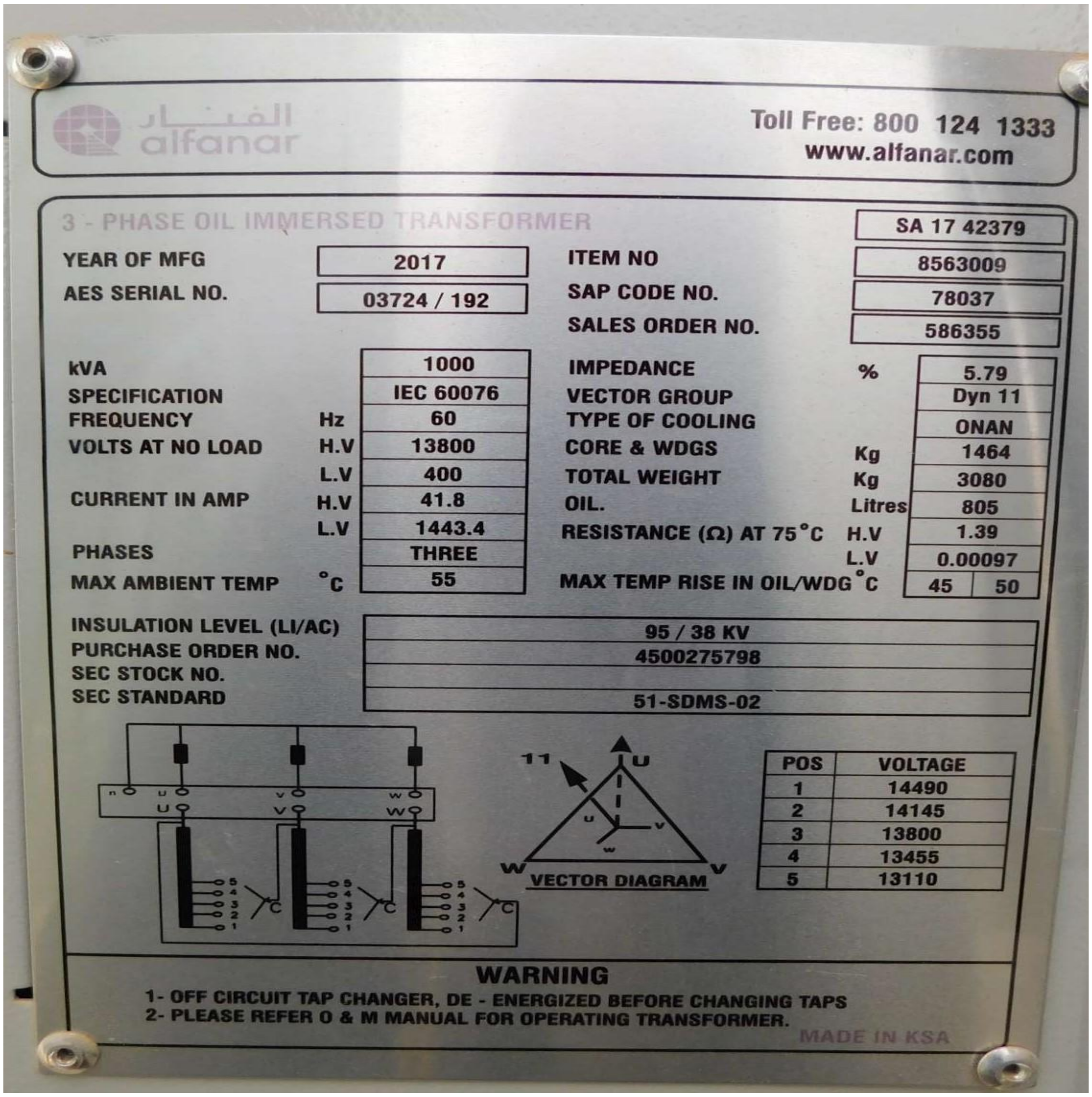

DAM1N-160
In
Ui
Uimp
ICU(KA)
Cat:A. T=55°C
50-60Hz
Standard
Ics=Icu
IEC60947-2
MCCB
70A
750V
8KV
20KA/400V
25KA/230V
Ii=10In
C·DADA
DAM1N-160
70
DADA ELECTRICAL
CO.,LTD
In:70A
الشركة السعودية للكهرباء
Saudi Electricity Company
37-SDMS-01 Rev.03
P.O. : 4500275364
Item No.:08371014
Serial No.:#
YEAR 2017
OFF
MOULDED CASE
CIRCUIT BREAKER
Vendor Name:
TRADING AGENCIES CO.,
MADE IN CHINA

Output

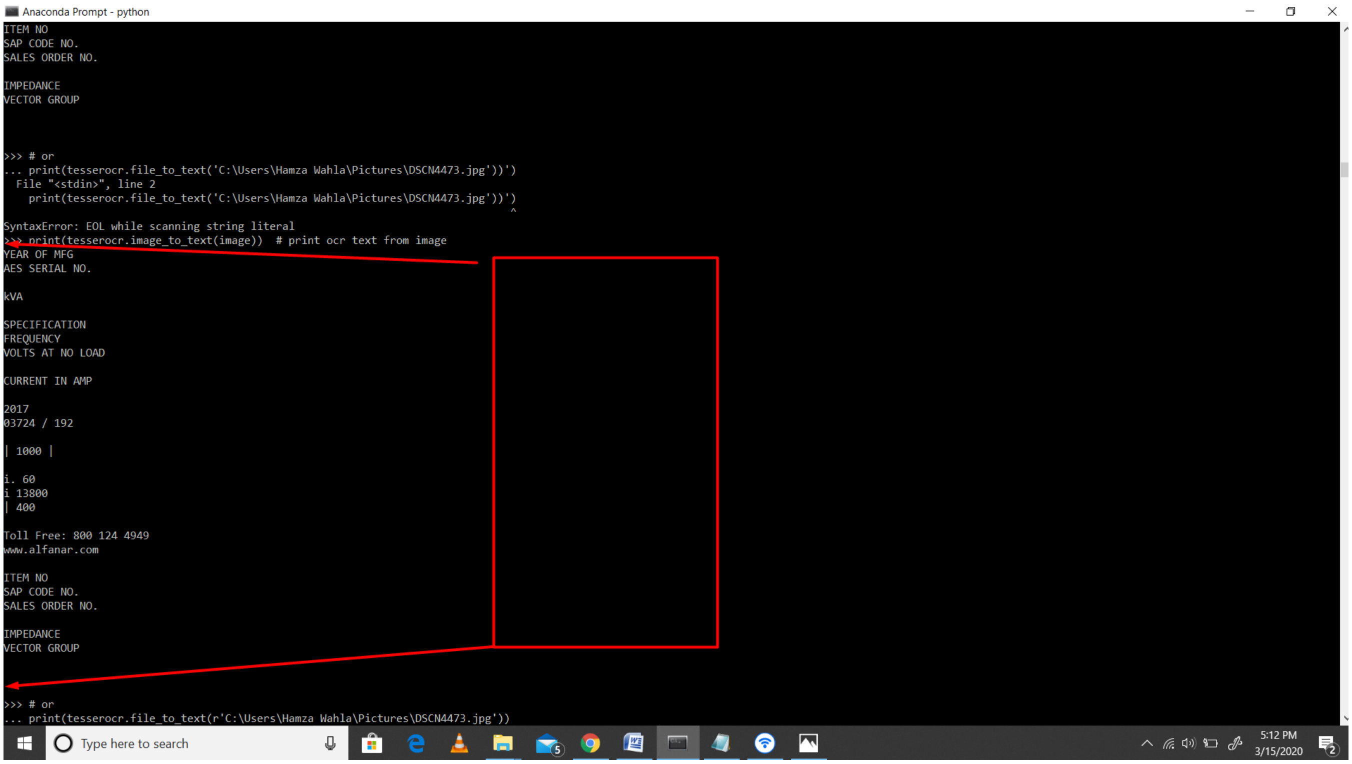


Select Anaconda Prompt - python
>>> # or
... print(tesserocr.file_to_text('C:\Users\Hamza Wahla\Pictures\DSCN0783.JPG'))')
File "<stdin>", line 2
print(tesserocr.file_to_text('C:\Users\Hamza Wahla\Pictures\DSCN0783.JPG'))')
SyntaxError: EOL while scanning string literal
>>> print(tesserocr.image_to_text(image)) # print ocr text from image
SN] ¢- pana
DADA ELECTRICAL
Co.,LTD
In:70A
ey gS Loa gest As ct
'Saudi Electricity Company
37-SDMS-01 Rev.03
P.O. : 4500275364
Item No.:08371014
Serial No.:#
YEAR 2017
!
)
OFF
DAM1N-160
MOULDED CASE
CIRCUIT BREAKER
Vendor Name:
TRADING AGENCIES CO.
MADE IN CHINA
>>> # or
... print(tesserocr.file_to_text(r'C:\Users\Hamza Wahla\Pictures\DSCN0783.JPG'))
Type here to search
2 new notifications
4:46 PM
3/15/2020

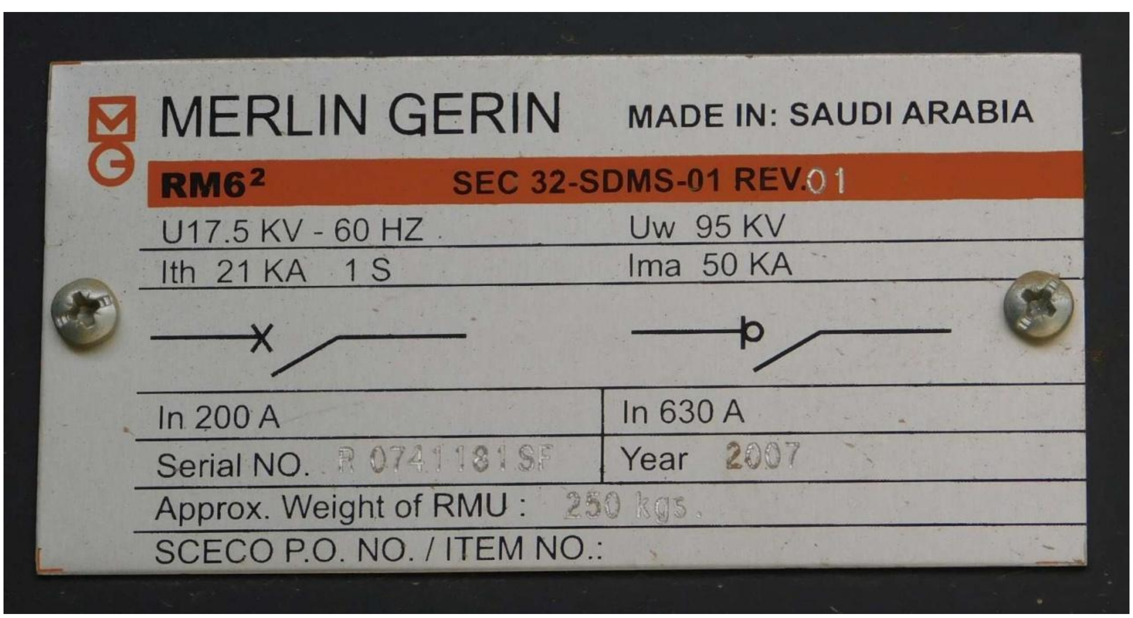
Example 3 (Transformer)



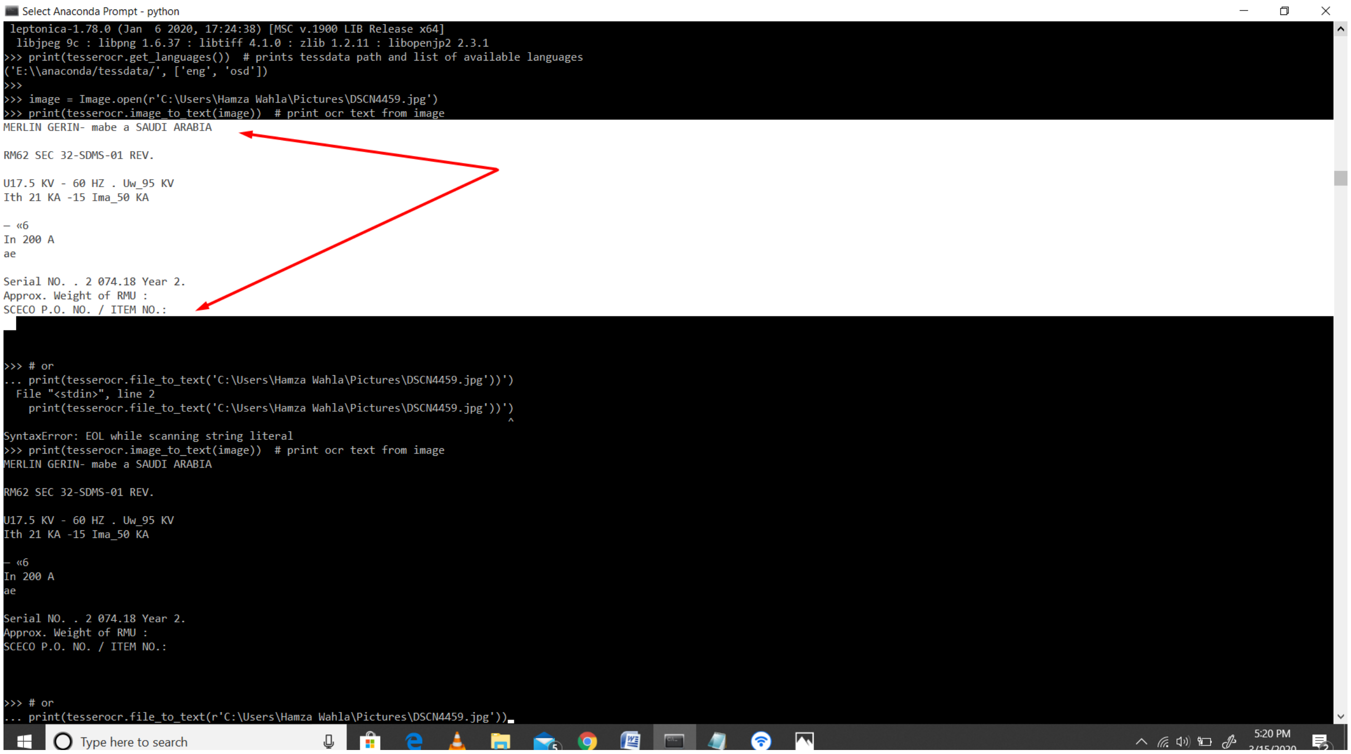
output



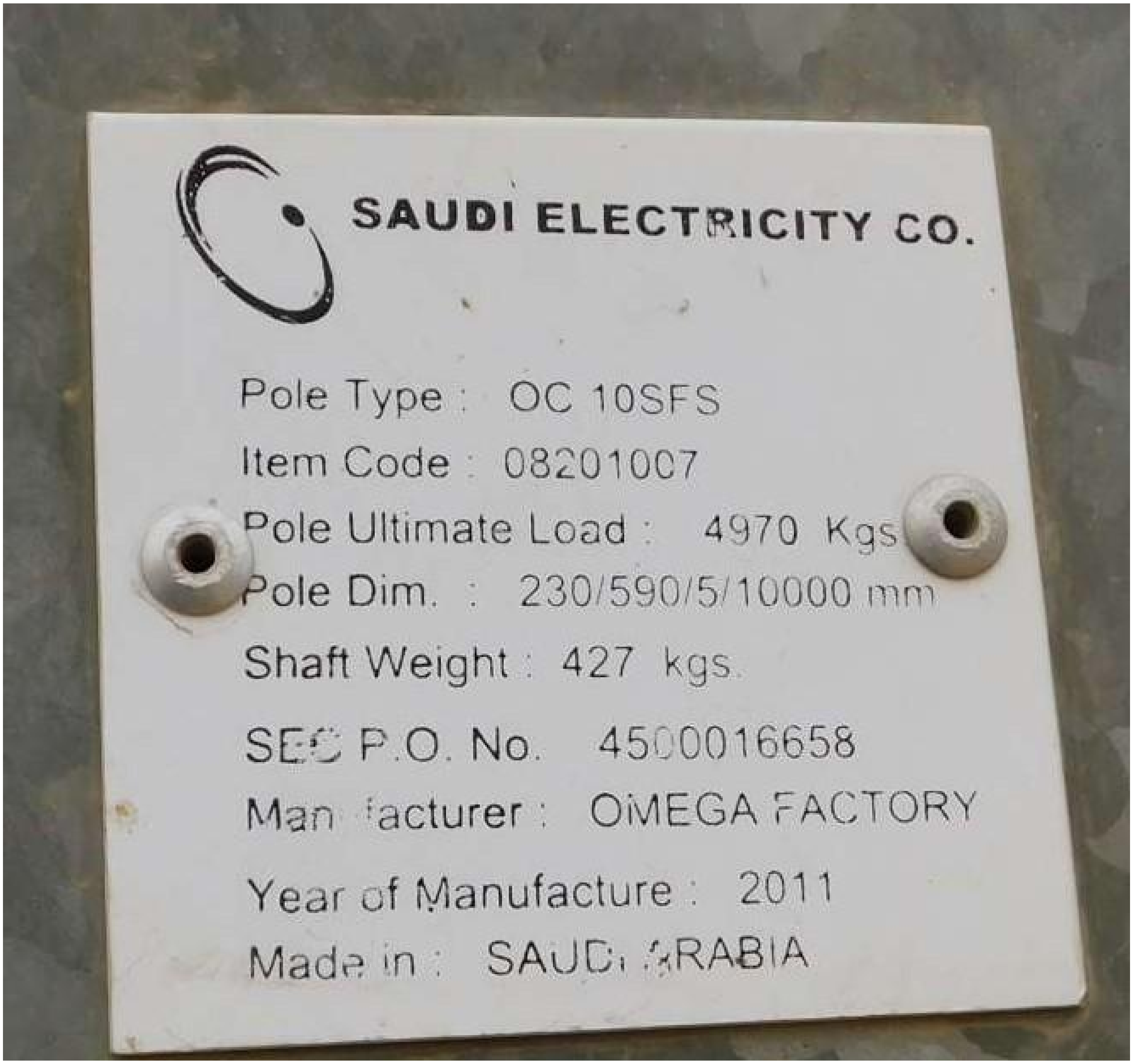
Example 4 (circuit breaker)



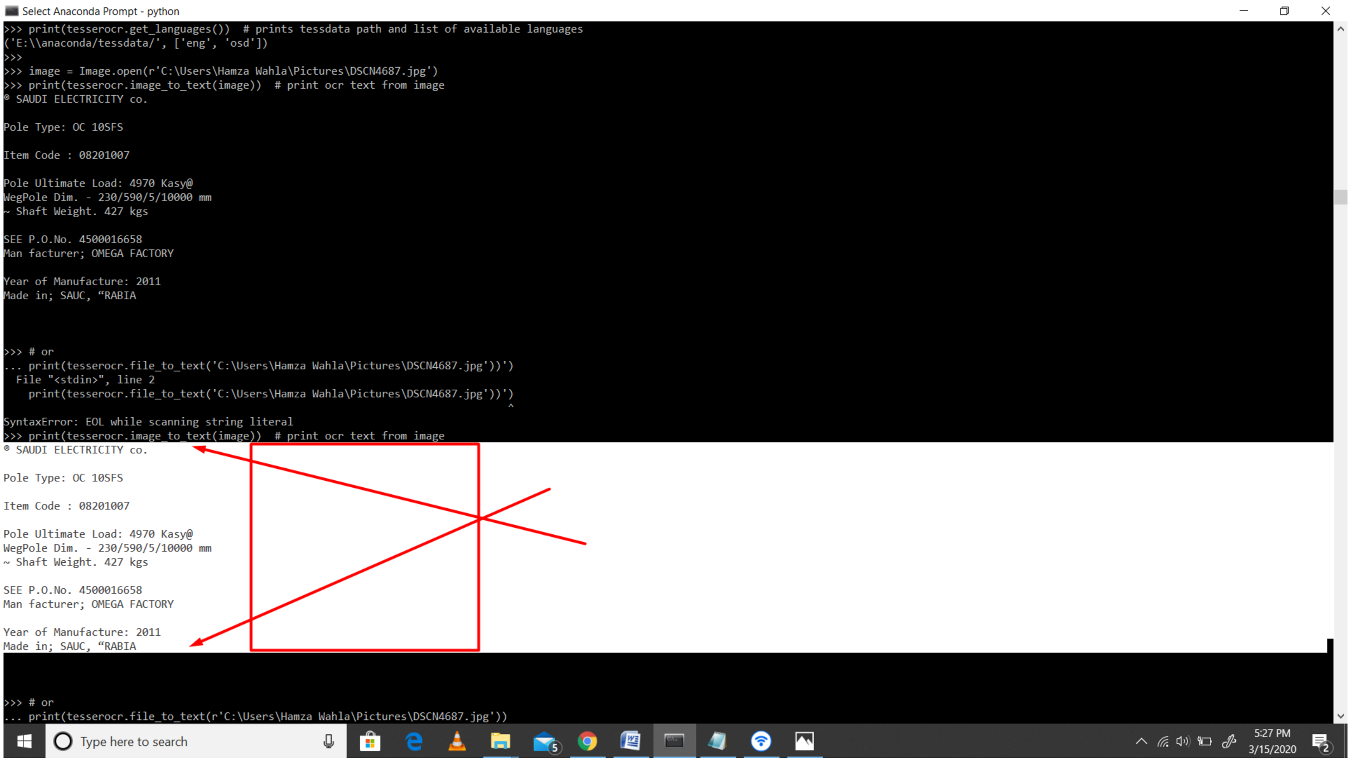
Output


Select Anaconda Prompt - python
leptonica-1.78.0 (Jan 6 2020, 17:24:38) [MSC v.1900 LIB Release x64]
libjpeg 9c : libpng 1.6.37 : libtiff 4.1.0 : zlib 1.2.11 : libopenjp2 2.3.1
>>> print(tesserocr.get_languages()) # prints tessdata path and list of available languages
('E:\\anaconda/tessdata/', ['eng', 'osd'])
>>>
>>> image = Image.open(r'C:\Users\Hamza Wahla\Pictures\DSCN4459.jpg')
>>> print(tesserocr.image_to_text(image)) # print ocr text from image
MERLIN GERIN- mabe a SAUDI ARABIA
RM62 SEC 32-SDMS-01 REV.
U17.5 KV - 60 HZ . Uw_95 KV
Ith 21 KA -15 Ima_50 KA
— «6
In 200 A
ae
Serial NO. . 2 074.18 Year 2.
Approx. Weight of RMU :
SCECO P.O. NO. / ITEM NO.:
>>> # or
... print(tesserocr.file_to_text('C:\Users\Hamza Wahla\Pictures\DSCN4459.jpg'))')
File "<stdin>", line 2
print(tesserocr.file_to_text('C:\Users\Hamza Wahla\Pictures\DSCN4459.jpg'))')
SyntaxError: EOL while scanning string literal
>>> print(tesserocr.image_to_text(image)) # print ocr text from image
MERLIN GERIN- mabe a SAUDI ARABIA
RM62 SEC 32-SDMS-01 REV.
U17.5 KV - 60 HZ . Uw_95 KV
Ith 21 KA -15 Ima_50 KA
— «6
In 200 A
ae
Serial NO. . 2 074.18 Year 2.
Approx. Weight of RMU :
SCECO P.O. NO. / ITEM NO.:
>>> # or
... print(tesserocr.file_to_text(r'C:\Users\Hamza Wahla\Pictures\DSCN4459.jpg'))
Type here to search
5:20 PM

Example 5 (pole)



Output



Task to be done :

We are committed to do this second portion of methodology in a second phase of project, as most of our job is done just we have to clear up the identities of characters that is the boundary box, this will be then converted to CSV format.