Conversion of form data to processed data in CSV format

Megha Chandwani,
Btech.
Computer Engineering,
Vishwakarma Institute of
Information Technology
Pune,Maharashtra.

Jay Paghdar,
Btech.
Computer Engineering,
Vishwakarma Institute of
Information Technology
Pune,Maharashtra.

Dr. Kirti Wanjale,
Asso. Prof. Computer
Engineering department,
Vishwakarma Institute of
Information Technology
Pune,Maharashtra.

Abstract

This paper consists of processing data from various types of forms and organizing it in a tabular format for easy readability or analysis or data on excel sheets. This data on the forms can be in any digital format.It also comprises of the detailed working for processing the data using OCR and henceforth after the procedures of inserting this extracted data into any tabular format (excel or google sheets)

1. Introduction

Occasionally we come across multiple types of forms either as hard copy or a digital form. Everywhere,this form data is used to analyze multiple fields of the form and hence there becomes a need to convert this data to a tabular format. You have a piece of paper which contains multiple fields and their values added. You want this data(fields,values) filled in an excel sheet for all the same types of forms. How will you do this ? Type in all the fields and their values? This becomes a tedious task and this is where we get the use of OCR- Optical Character Recognition.

Generalising the system described it will consist of Optical character recognition, Pattern Recognition, Image Segmentation, Text Extraction using Tesseract. OCR is capable of recognising handwritten as well as printed text and its performance can be judged based on the quality of the documents and the camera being used to capture the raw image.Now a character recognition system based on android devices is being proposed.

What is OCR?

Optical character recognition, also referred to as optical character reader (OCR) is the electronic conversion of images of typed/handwritten/printed text into machine-encoded text, whether from a scanned document, a photograph of a document.

It is widely used as a method of data entry from printed paper data records – whether invoices, passport documents, printouts of static-data, bank statements, computerized receipts, business cards, invoices, mail, or any suitable documentation. OCR is a field of research in pattern recognition, AI and computer vision. It is a common method of digitizing printed texts so that they can be electronically edited, stored more compactly, displayed on-line, searched, and used in machine processes such as cognitive computing, text-to-speech, machine translation, text mining and key data.

Types Of OCR

There are different types of OCR:

1. Intelligent Word Recognition – It captures cursive text or handwritten texts. The algorithm recognizes an entire unconstrained handwritten word rather than picking up individual characters.
2. Intelligent Character Recognition – It captures handwritten or cursive text. The engine identifies a single character at a time and evolves with its embedded machine learning.

3. Optical Word Recognition – It targets typewritten text wordwise and is usually mentioned as OCR.
4. Optical Character Recognition – It captures typewritten text and goes one character at a time.
5. Optical Mark Recognition – It is a method of gathering human input data by identifying marks or patterns on a document.

How does OCR Work?

1. Pre-Processing

Pre-Processing of the images helps improve the OCR results. Following are some common techniques used based on the quality of the image which needs to be processed for data extraction.

a) De-skew: This helps to align the scanned images.
b) Binarisation: It converts the color of an image to black and white. It helps in separating text from the background and which helps make data recognition easier.
c) Despeckle: It smooths the edges by removing any spots whatsoever.
d) Line removal: It helps clean all the extra spaces and lines to optimize data.
e) Zoning: It separates different zones such as columns, captions, etc.
f) Script recognition: It helps to identify different scripts in the document. It is necessary so that the right script is invoked by the OCR during data capture.
g) Segmentation: It helps segment every character before OCR runs on it.

2. Character Recognition

1. Matrix matching: It works by comparing a character image with the glyph stored. This type of character recognition works best when fonts in the document are not that fancy.
2. Feature Extraction: This feature recognizes features like lines, direction, intersections, and loops which makes the entire character recognition an efficient system.

3. Post Processing

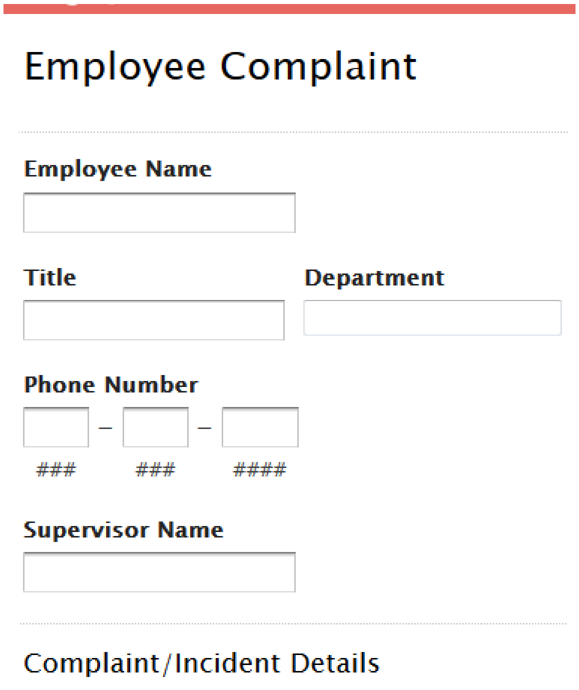
Once the Image data is processed, it's accuracy can be improved. Lexicon plays an important role in improving the quality of the extracted data. Lexicons are the list of words which can occur within the document. Data processing can get a bit tricky if the document doesn't contain Lexicons. There are other techniques like Natural Language Processing (NLP), Database Lookups which will further improve the accuracy of the image data extraction process.

2. Method:

A. Input Data

Input data can consist of any image which can have a form with fields or a word/pdf document which may consist of a form whose data needs to be accumulated in an excel sheet. While inputting an image we can select an existing excel sheet where we need data to accumulate or a fresh excel sheet.

Example-



Employee Complaint

Employee Name

Title

Department

Phone Number

– ### –

Supervisor Name

Complaint/Incident Details

B. TERSSERACT (A tool for identifying text from images)

Tesseract is an open source engine for optical character recognition. It is available on many operating systems. It is one of the most accurate OCR engine available. It can read and convert into over as many as 60 languages.It was developed at HP between years 1984 to 1994 but its first working copy was released only in 2005 as open source by HP. Tesseract is available at http://code.google.com/p/tesseract-ocr.
Architecture of Tesseract

Tesseract recognises a word in two passes, that is,it tries to recognize the words in the first pass. If the match is found, then the found word is passed on to the Adaptive Classifier, which recognizes the text more accurately. During the second pass, the words which were not at all recognized or were not well recognised in the first pass are recognized again through a run over through the page. Finally Tesseract resolves fuzzy spaces. To locate small and capital text, Tesseract checks alternative hypotheses for x-height.

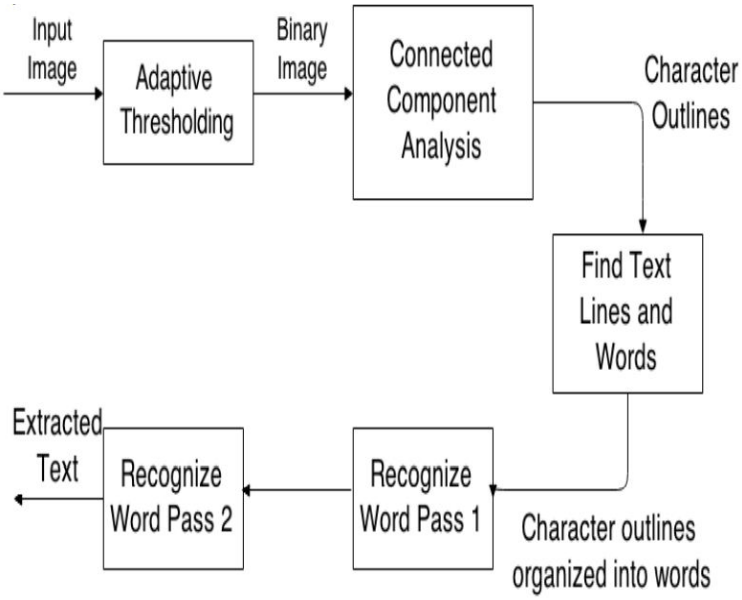


Figure 1 Architecture of Tesseract

Tesseract converts the input image into binary format using thresholding. Outlines of components are stored on connected Component Analysis. Nesting of outlines is done which gathers the outlines together to form a Blob. Text lines are analyzed for fixed pitch and proportional text. Then the lines are broken into words by analysis according to the character spacing. Fixed pitch is chopped in character cells and proportional text is broken into words by definite spaces and fuzzy spaces.

C. ALGORITHM

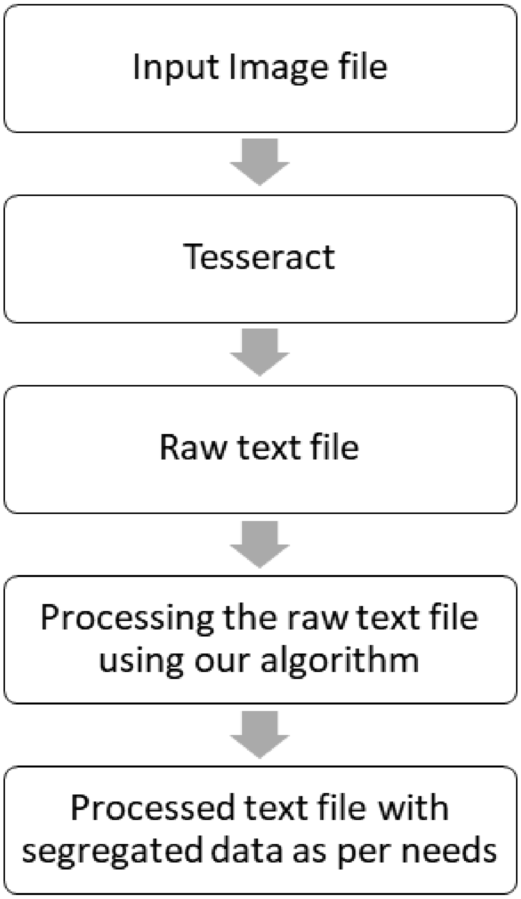
Algorithm for extracting text from image in required format

1. Use tesseract to get the image data in a text file. The text file will have data exactly in the horizontal order in which data is present in the image.
2. Process the text file which is obtained from output given by tesseract. . Text file has to be

processed in such a way that a delimiter will be added after each field and its value. Lets assume our delimiter is ‘,’ -- The processing will take place as follows-

a. We will search for field names that are required in excel and add a delimiter just before the field name. The field names will be taken as input from the user.
b. All fields that are not required to be present in the excel form will be removed from the text file.

3. Since we have processed the text file we now have all fields separated by a delimiter . In our case we have chosen a comma. Example - Name Megha Chandwani , Email ,meg@gamil.com , Age 18
4. We now segregate delimiter separated fields in a separate line


Input Image file
Tesseract
Raw text file
Processing the raw text file using our algorithm
Processed text file with segregated data as per needs

Our next task is to convert the text file into csv. We will be using python pandas library to achieve the same. For more details about pandas https://pandas.pydata.org/docs/

Algorithm for converting text file into excel (csv)

1. The method read_csv() will take in the text file,separator and header if any. With this we will get a pandas dataframe. (What is dataframe ? It is a simple 2-dimensional data structure with rows and columns.)

df = pd.read_csv("sample..txt", sep= “,”)

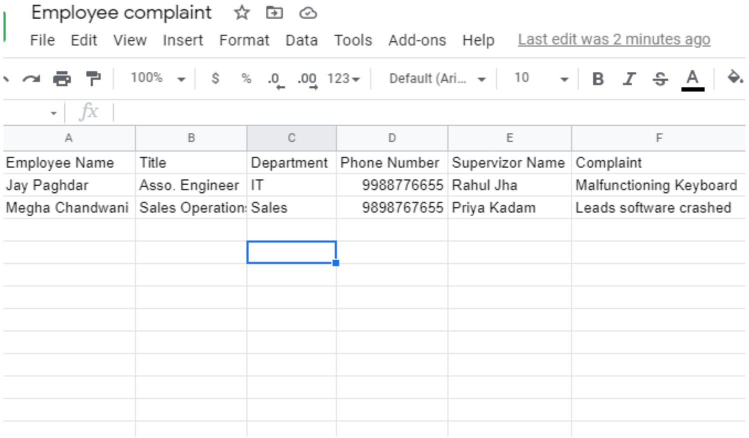
2. We will now take transpose of the data frame so we will get all field names in a row and multiple rows of field values

df1 = df.T

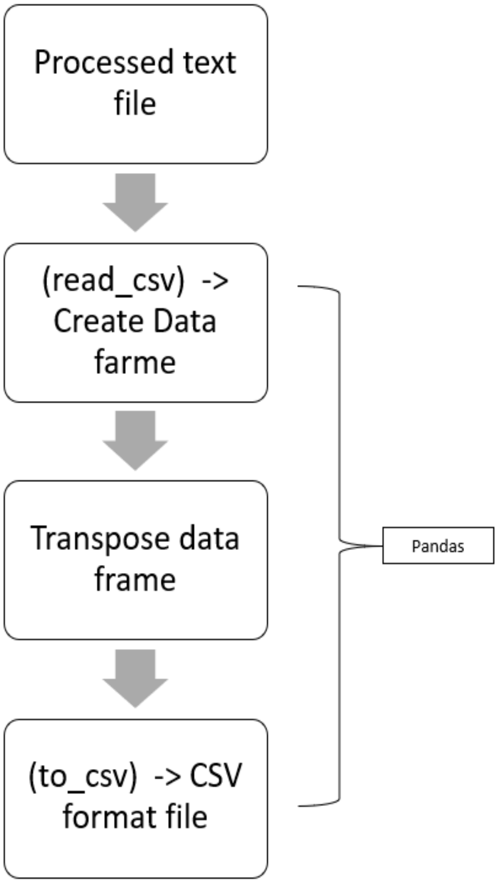
3. Once this dataframe is created we will store this dataframe in csv file format (excel). This will be done using dataframe.to_csv() function which will take in name of csv file as parameter

df1.to_csv(‘sample.csv’,index=None)

After this step we will get a csv file with required data. We can use any separators.


Employee complaint
File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago
100% $ % .0 .00 123 Default (Ari... 10 B I S A
fx

A	B	C	D	E	F
Employee Name	Title	Department	Phone Number	Supervizor Name	Complaint
Jay Paghdar	Asso. Engineer	IT	9988776655	Rahul Jha	Malfunctioning Keyboard
Megha Chandwani	Sales Operation:	Sales	9898767655	Priya Kadam	Leads software crashed



D. Output Data:

3. Issues/Limitations of this system

This system highly depends on the input data and results may vary for some complex forms. The forms whose data needs to be extracted in excel should be kept as simple as possible to get accurate results.

If Tesseract is used for identifying image data,it is not capable of recognizing handwriting and hence other alternatives for identifying hand written forms must be used. It works best with digital data

4. Accuracy

Quality of input image highly determines accuracy of OCR systems. Manier types the output is noisy and post processing has to be performed. We can note that for 10 words to be identified the system can take about 160ms. Factors that can affect quality of input image consist of - Focus of the camera used to take the image, resolution of the picture, amount of noise present etc. Tesseract tool for OCR as detected gives an average accuracy of 93%.

5. Conclusion

This paper tells us about converting any kind of form data in any format particularly image data into a csv. It gives the user an ease of having its own customized form and getting that data in excel.

Many people tend to use google forms which do give us data in excel format but it has its own limitations.

Simply our system can take in forms in any format ie users can take a picture or screenshot of their form give that as input to this system and get that data in excel sheets

6. Motivation

One fine day I came across my father typing data from multiple similar forms into excel. Post this ,it struck me if we had a system which could import form data directly into excel and this brought me to research about how can this be done.

7. Acknowledgements

The authors would like to thank Department of Computer Engineering, Vishwakarma Institute of Information Technology and indebted to our guide Prof.Kirti Wanjale for her guidance which helped us design this paper.

REFERENCES

[1] Ch, Sravan & Mahna, Shivanku & Kashyap, Nirbhay. (2015). Optical Character Recognition on Handheld Devices. International Journal of Computer Applications. 115. 10-13. 10.5120/20281-2833.

[2] The Tesseract open source OCR engine, International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 22, April 2015 13 http://code.google.com/p/tesseract-ocr .

[3] R. Smith. "An overview of the Tesseract OCR Engine."
Proc 9th Int.Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil,Sep 2007

[4] Doc Acquire (2019 August 28) "What is OCR?". https://www.docacquire.com/resources/blog/what-is-ocr/ .

[5] Yuvraj Singh "Extracting text from images with Tesseract OCR, OpenCV, and Python" May 2020, 21. https://www.opcito.com/blogs/extracting-text-from-images-with-tesseract-ocr-opencv-and-python

[6] Convert Text File to CSV using Python Pandas, September 2020, 02. https://www.geeksforgeeks.org/convert-text-file-to-csv-using-python-pandas/

[7] Pandas Documentation https://pandas.pydata.org/docs/

The extraction and recognition of text from multimedia document images Smith, R. W. (Author). 1987