Educational Training For Processing Invoice Of Vendor Identification And Payments Using Python-Tesseract

[1]Rekha M, [2]P Srividya devi, [3]U Vijaya Laxmi, [4]P Sirisha, [5]M Karthika,

[1]Assistant Professor, GRIET, Hyderabad
ORCiD:0000-0001-9090-0204
[2]Associate Professor, GRIET, Hyderabad
ORCiD: 0000-0001-6131-7421
[3]Assistant Professor, GRIET, Hyderabad
ORCiD:0000-0001-9870-6789
[4]Assistant Professor, GRIET, Hyderabad
ORCiD:0000-0002-8888-3017
[5]Assistant Professor, GRIET, Hyderabad
ORCiD:0000-0001-6591-3837

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract. The aim of the project is to recognize the invoices of receipts from various vendors, by using automated invoice processing using various learning educational tools. This automated invoice processing is far better than manual invoice processing, it saves a serious amount of time and money creating efficiencies and increasing the accuracy of captured data. Basically, the invoices were calculated from the scanned receipts by using python-tesseract software. Python- tesseract is an optical character recognition (OCR) tool for python. It will recognize and read the text embedded in images. So, this python-tesseract software extracts key information like bill or invoice number, amount etc.; from all receipts and imports the calculated invoices and total amount of all receipts which are given by vendors to the database.

Keywords: Invoice Number, Python, OpenCV, Tesseract

INTRODUCTION

Invoice processing is the process that is very useful for an organization to handle their accounts and to pay the invoices amount. An organization or a company needs to process their invoices when they receive the bills or invoices and it ends when the amount of those invoices are paid. This requires a whole process of steps to handle the invoices. In earlier, these invoices were processed by manpower, where it needs huge amount of time and labor. Also, by using this manual process the errors in the payments may cause and it has high risk of bad reputation to a company. In manual invoice processing the input that is bills or receipts were sent to organization by email or by paper etc.; by taking these receipts the company's accounts man must process the invoices and pay the amount to the vendor. This process includes some steps were the accounts men must go through every receipt. And then the account men must keep a record of the invoice number, amount and other details which are needed to the company. Then the amount is needed to be paid on time to the vendors as the above process include huge amount of manpower, man hours, this method is not appreciable. So now a days there are several software's that are used to calculate the invoice numbers, amount to be paid etc.; There software's needs the receipts as input, as these receipts carry all the information regarding the vendor as well as the amount to be paid. These receipts are sent to the companies through PDF's or images of paper or through email. So these receipts might contain dirt or noisy which enables the software to give faulty outputs. Because of this, the reputation of the company might get bad and losses the trade of the company. At sometimes the company might lose their supplies from the vendor or supplier. So one should be careful regarding these receipts thing. Receipts have everything that is needed to a company to handle the accounts. As in manual invoice processing, this method also requires receipts to handle the invoice processing. In automated invoice processing, processing of invoices to recognize invoice number and total amount, requires Optical Character recognition tool and some other software requirements. But to use the receipts in these software's the receipts are needed to be clean without any noise or any disturbances. But it is very difficult to get the receipts like that (noise free, dirt free) from the vendors. If the receipts are noise free and dirt free, then the process is very easy and will get the accurate outputs also. But in most of the cases the receipts have huge noises, disturbances. So, it is difficult to get the accurate output. In that case this method is not so good. In order to avoid these kind of mistakes, image processing and correction of image is needed. So there are several software's were the image processing and image correction is done easily. Also, there are several other things to get fault in results. They are handwritten text in receipts, noisy images, faded images, small font or different character images, water marking in the images due to any reason, faded text in the receipts, and wrinkles in the receipts etc; These things also cause errors or the output is not accurate with these things. Extracting the required information from the receipt is also a big task. If the receipt contains above mentioned noises then it is difficult to automate the invoice processing. By using image processing and correction software's the automation gets easier but extracting and converting the image to string is difficult. This job is done by Optical Character Recognition tool. Optical character Recognition tool is the most common thing used to extract text embedded in the image that is receipt. By using these tools the company can avoid the problems of printing and paper costs, huge amount of time and errors caused by humans. This invoice processing is not only used in companies, these are also used in taxation areas. But compared to manual invoice processing automated invoice processing has many advantages. So now automated invoice processing is used everywhere to avoid man hours, labor, transparency costs, paper costs, printing costs. Automated invoice processing enables a company to get more and more trades and reputation. It has several advantages, so even a small company is also shifting to automated invoice processing rather than using manual invoice processing. The receipts are mostly noisy and disturbance, so it is difficult to extract the information from them. So the Optical Character Recognition tools and the extraction tools work well to automate the invoice processing. Image processing is required for avoiding the noisy text, faded text or any other disturbances. This image processing is called as preprocess step in the automated invoice processing. There are several preprocess methods like noise removal, grey-scaling, thresholding etc.; these preprocessing methods are mainly used in removing the noisy text and disturbances. Image processing step in automation of invoice processing help to correct the image and contrast the image etc.; so that the automation of invoice processing can be done easily with accurate output and greater efficiencies.

Optical Character Recognition tool is used for extracting the text embedded in the image read the text embedded in the image such as scanned copy of images. This tool can convert any kind of data embedded in the scanned copy to text. This Optical Character Recognition can be achieved in two steps. First thing is detection of text in the scanned copy of the image. Second step is to read or recognize the text embedded in the scanned copy of the image. Now the invoices are processed by using python tesseract software. Python tesseract is an optical character recognition tool. This tesseract software is used to automate the invoice processing in this project.

AIM AND OBJECTIVE

Automated invoice processing is very useful to a company to process their invoices with greater efficiencies. First of all, a company or an organization receives the invoices or receipts from the vendors. And then these receipts are needed to be scanned without any margin mistakes or any other mistakes. These scanned receipts are given as input to the Optical Character Recognition tool that is python tesseract software. Then the invoice number and amount paid to the vendor is displayed. Hence the amount is paid to the vendor at correct time with more accurate data.

1. Reduced paper and printing costs.
2. Saving time manually entering financial data into the systems.
3. Producing timely payments to suppliers.
4. Improving the accuracy of the data.
5. Errors in supplier payments, such as over payments and duplicated payments.
6. Reducing time spent on the phone dealing with supplier queries.

By automating the invoice processing, errors that are caused due to human are minimized. Time taken to write all the details that are in the receipts gets reduced by automating invoices.

Improves the accuracy of data and eliminates the risk of missing invoices. Reduces the paper and printing costs. In earlier days all the accounting tasks were handled manually and in resulting, there were lots of errors. With today's available technology, there's no reason to be relying on email and spreadsheets to handle invoicing needs. Manual invoice processing has high risk of human errors and may not enable you to pay all your invoices by the payment date. It requires a huge amount of time, man-hours, paper and printing costs, also the accuracy of data is not good.

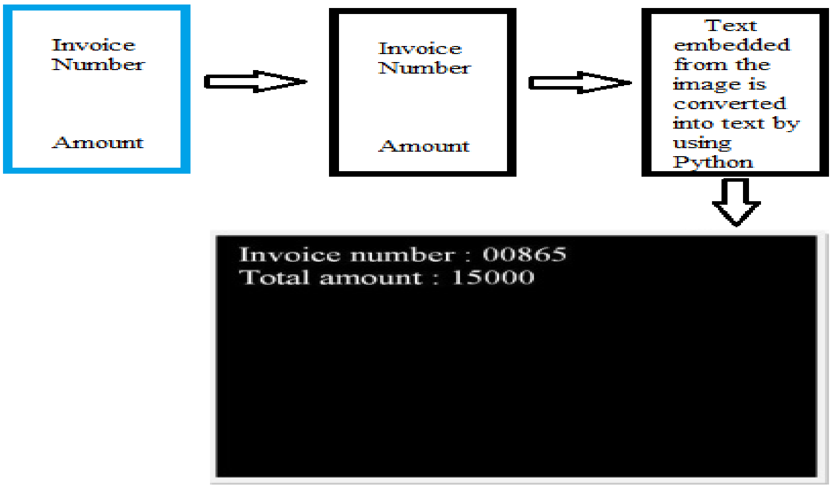
Block Diagram

Supported Operating System, Windows 7(32 or 64 bit), Windows 8(32 or 64 bit), Windows 10 (32 or 64 bit)Supported Development Environment, Python, anaconda, Tesseract, OpenCV, matplotlib, Optical Character Recognition (OCR) is a tool used to get the text information or printed text or handwritten text embedded in the scanned copy of the image. Basically this Optical Character Recognition (OCR) is of two steps. The two steps involved in the Optical Character Recognition (OCR) tool are detecting text from the scanned copy of the image or the receipt and extract the information embedded in the text or to recognize the text from the scanned copy of the image or the receipt. Optical Character Recognition (OCR) tool consist of both hardware and software equipment. This hardware and software equipment enables Optical Character Recognition (OCR) tool to get the information embedded in the scanned copy of image or receipt and to read the information or to recognize the text embedded in the scanned copy of image or receipt. Optical Character Recognition (OCR) tool needs two steps to perform its task. The two steps involved in Optical Character Recognition (OCR) are detecting the text and recognizing the text. Firstly, Optical Character Recognition (OCR) tool requires the scanned copy of the receipt or image. And then Optical Character Recognition (OCR) tool copies all the scanned images or receipts and converts the image or receipt into black and white color. This scanned image or receipt is analyzed to recognize the text, characters embedded in the scanned copy of image or receipt as shown in fig.2.4.1.2. There are two methods or algorithms to recognize the characters and text embedded in the scanned copy of the image or receipt. They are pattern recognition, featured detection. In Optical Character Recognition (OCR) tool there are several examples of programs. This programs in Optical Character Recognition (OCR) tool are used to get the various fonts or formats of the text. These fonts and formats are used to compare the text that is embedded scanned copy of the image or receipt. This is called pattern recognition. There are also several examples which help in understanding the rules of writing specific letters, numbers and characters. These rules are used to compare with the various kinds of letters, numbers and characters that are embedded from the scanned copy of image or receipt. This is called feature detection. Optical Character Recognition (OCR) tool has several advantages. It saves serious amount of time, decreases the errors caused by human due to any reason and minimizes the effort. It can save huge amount of data that is images or receipts and it can compress the file. One can edit the image or receipt that is scanned into the Optical Character Recognition (OCR) tool and can search the image or document or file that is scanned in Optical Character Recognition (OCR) tool at any point of time[1].

Python tesseract software is an Optical Character Recognition (OCR) tool. This python tesseract software is used to process the invoices of the vendor to recognize invoice number and total amount. There are Optical Character Recognition (OCR) tools available but python tesseract software is more efficient. It can recognize text embedded in the scanned copy of the image or receipt more accurately. Tessaract software can be accessed with many programming languages. In this project python programming language is used. This python tesseract software recognizes the text embedded in the scanned copy of the image or receipt as shown in fig.2.4.2. This python tesseract software has the capability to recognize the text embedded in very large document or file als[2].

OpenCV means open source computer vision. Open source computer vision is a library used for processing a image [3]. This open source library is based on machine learning, it is a machine learning software library. Open source computer vision is built to provide the infrastructure to the applications that are based on computer vision. In open source computer vision one can write code according to their need. The code in the open source computer vision can be edited or modified according to our requirement. This open source computer vision has various number of algorithms or examples to perform various tasks. There are around two thousand five hundred algorithms in the open source computer vision. These algorithms can used for various tasks such as detecting the faces, recognizing faces, identifying the objects, differentiating human actions in videos, to track the movements of the camera, to track the movements of the objects, to extract the objects that are in 3D models, to stitch the images together to provide a image comprising of both the images in one image as entire scene with highest resolution, to find the images in the database that are similar, to remove the excess light, color, flash from the image, to identify the eye movements in a video etc. there are several thousands of people who are using this open source computer vision. This open source computer vision is hugely used in many companies, government, research people etc. some of the companies that are using this open source computer vision are Microsoft, Google, Honda, IBM, Yahoo, Sony, Intel etc. not only these well established companies, there are various number of startups that are using this open source computer vision library. The startup companies that are using this open source computer vision library are VideoSurf, Applied Minds etc[4]

Matplotlib is a library that is used for visualization in python.[5] It is a multiplatform visualization library. This matplotlib visualization library is used for plotting in python. It is an amazing library for plotting 2D arrays in python. Matplotlib visualization library is built on NumPy arrays. Matplotlib visualization library was introduced by John Hunter in the year 2002. Matplotlib visualization library has various applications, it is one of the greatest visualization library. This Matplotlib visualization library has various benefits in plotting. It can allow visual access to large lumps of data in very easy way. This Matplotlib visualization library has several plots like scatter, histogram, line, bar etc.; Any plotting works that are in python and NumPy are done by this powerful plotting library. Matplotlib visualization library is a tool that is used in various applications for plotting the images. It is very necessary to visualize the data, so it can be achieved by this Matplotlib visualization library.[6] It is almost MATLAB like interface. The difference between Matplotlib visualization library and MATLAB is it uses python for visualizing the data. Below fig 1. Illustrates the block diagram


Invoice Number
Amount
Invoice Number
Amount
Text embedded from the image is converted into text by using Python

```
Invoice number : 00865
Total amount : 15000
```

FIGURE 1. *Block Diagram representing the text embedded from the image*

FLOWCHART AND EXPLANATION

There are various steps involved in automating the invoices to recognize the invoice number and total amount.

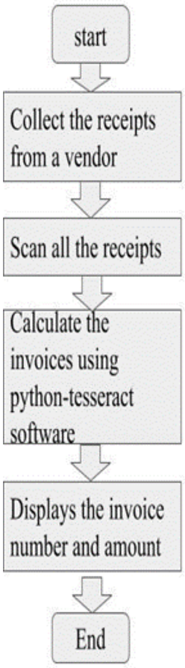
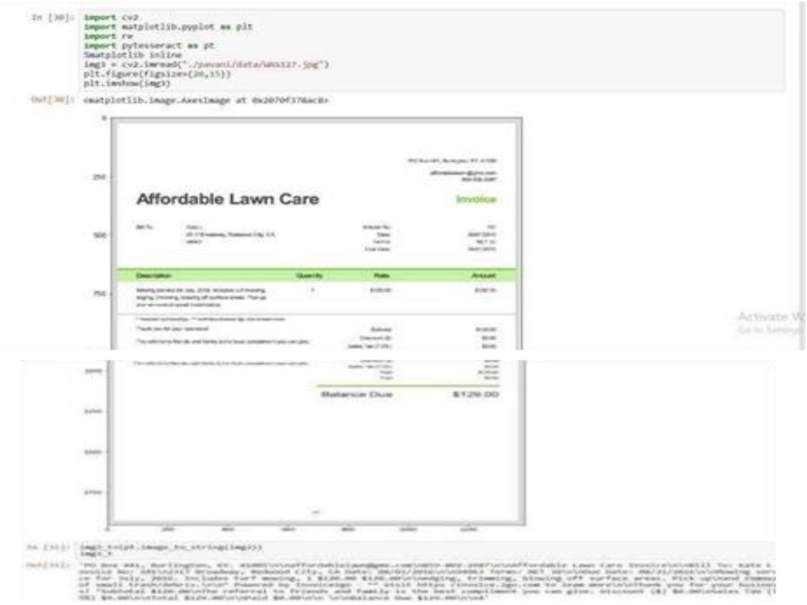

start
Collect the receipts from a vendor
Scan all the receipts
Calculate the invoices using python-tesseract software
Displays the invoice number and amount
End

FIGURE 2. *Flow chart of the code written for the process*

Those steps are collecting the receipts, scanning the receipts without any errors, preprocessing the image, passing it through python tesseract software to get the invoice number and total amount of the receipts given by the vendors as shown in fig 2. Before running the code we have to install all the required software. First we have to create a folder in which it stores the receipts of a vendor [7],[8],[9]. Then we must open the anaconda software, launch the Jupiter notebook. Then a window opens select desktop then check whether the created folder is present, then click on new which is on open top right, thus we have created python notebook 3, in that we have to write the code. First import all the required libraries like regular expressions, cv2, etc.., Write the code to import all the libraries, then write code to import image and then we have to convert image to text so write required by importing tesseract, from the converted text we have to find invoice number and total amount of a vendor by writing the required code, after each code we have run the program. Then output is displayed.

RESULTS

Fig.3 represents the output of the software shows the imported calculated invoices and total amount all receipt values and total amount of all receipts which are given by vendors to the database



```
In [30]: import cv2
         import matplotlib.pyplot as plt
         import re
         import pytesseract as pt
         %matplotlib inline
         img3 = cv2.imread("./pavani/data/WA5327.jpg")
         plt.figure(figsize=(20,15))
         plt.imshow(img3)
Out[30]: <matplotlib.image.AxesImage at 0x2070f378ac8>
```

```
In [31]: img3_t=(pt.image_to_string(img3))
         img3_t
Out[31]: [illegible]
```

Figure 3. *Result shown by all receipts in software*

DISCUSSION and CONCLUSIONS

Process the invoices to recognize invoice number and total amount in less time and cost by automation. By utilizing python-tesseract software in automation of invoice processing the accuracy of the captured data is improved and the errors get minimized creating efficiencies. In earlier days all the accounting tasks were handled manually and in resulting, there were lots of errors. With automation of invoice processing, error rates are minimized, preventing all types of error- resulting issues, such as delayed approvals type spent on searching for and correcting mistakes. Improves the accuracy of data and eliminates the risk of missing invoices. Reduces the paper and printing costs.

REFERENCES

1. Online:https://searchcontentmanagement.techtarget.com/definition/OCR-optical-character-recognition
2. Online: https://www.anaconda.com//
3. Online: https://sourceforge.net/projects/opencvlibrary/
4. Online: https://docs.opencv.org/master/d9/df8/tutorial_root. Html.
5. Online: https://matplotlib.org/3.2.1/tutorials/index.html
6. Online: https://www.edureka.co/blog/python-matplotlib-tutorial/
7. Prasanna Lakshmi, K., Reddy, C.R.K.A survey on different trends in Data Streams ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455.
8. Kumar, P., Singhal, A., Mehta, S., Mittal, A.Real-time moving object detection algorithm on high-resolution videos using GPUs (2016) Journal of Real-Time Image Processing, 11 (1), pp. 93-109.
9. Padmavathi, K., Sri Ramakrishna, K.Classification of ECG signal during Atrial Fibrillation using Autoregressive modeling(2015) Procedia Computer Science, 46, pp. 53-59.