IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 3, March 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003056 |

Use cases on PDF OCR using RPA

Sakshi Dhavale, Harshada Kakade, Harshada Sarwade, Vishakha Kshirsagar

Ms. B. V. Jadhav

PCP Student, Department of Computer, Pimpri Chinchwad Polytechnic, Akurdi, Pune, India

Department of Computer, Pimpri Chinchwad Polytechnic, Akurdi, Pune, India

ABSTRACT: Optical character recognition (OCR) is a main part of robotic process automation (RPA) solution.OCR is a technology which is used to extract text from images, documents, scanned structured and semi-structured receipts and invoices. It converts typed, handwritten or printed text into machine-encoded text-this data can then be used in electronic business processes without someone manually capturing it. For example, if you scan your document with the help of printer it will create an image of it and suppose if you want to make a editable document so that you can make a changes in it,that will not possible but with the help of our project we can convert this non-editable document into an editable format. We can reduce large amount of manual work of labours by using this project.According to the performance of the submissions, we believe there is still large gap on the expected information extraction performance.

KEYWORDS: Extract text, explore technology, Accurate result, Store data.

I. INTRODUCTION

Optical character recognition (OCR) is a main part of robotic process automation (RPA) solution. OCR is a technology which is used to extract text from images, documents, scanned structured and semi-structured receipts and invoices.On the other hand, extracting key texts from receipts and invoices and save the texts to structured documents can serve many applications and services, such as efficient archiving, fast indexing and document analytics. OCR also faces big challenges. It uses in many practical applications such as name card recognition, Adhar card recognition, license recognition and hand written text recognition. OCR helps to reduce the manual work of labour, due to this it can be backend workflows where workers become free to take the more responsibilities on their own. Some OCR software can only converts it into text or some other software converts the characters to editable text but our software will convert an image into text with layout of the text as well as its size or format.

II. LITERATURE SURVEY

There are various techniques that can be used to extract text from images, documents, scanned structured and semi structured receipts. So, there is one technology which is used to generate usecases on PDF OCR that is Robotic Process Automation (RPA). Using UiPath we can digitize data from documents and then it will processed and analyzed. Using this, we can analyze the data so it will helps in variety of things like handwriting, checkboxes, signature, Adhar-card recognition.It will alsoprovidesvarious benefits like accurate and flexible document processing, reduced risk of human error, increased operational efficiency. This process of recognition of characters is carried out in 4 steps as follows: firstlypreprocess the input data, secondly it will detect characters and lines,third it will postprocess the recognized data and lastly it will give the output as images.

III. PROPOSAL

Optical Character Recognition(OCR) detects and extracts text from images. If you ever had to retype a document that you could not find on your computer, you will know how hectic this work can be. Our OCR module will regain the text from a scan and will allow you to modify the document. Whether you have your document as paper or a non-editable PDF, out OCR engine will allow you to convert it back to an editable format. You can select individual images as well as process all images inside your document. OCR is the best technology in the industry. It will converts computer fonts in medium quality images with high accuracy. We can also use our OCR on low quality scans or on handwriting and save a lot of time as compared to manually rebuilding every single line of our document.There is one technology which is used to generate usecases on PDF OCR that is Robotic Process Automation (RPA). Using UiPath we can digitize

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003056 |

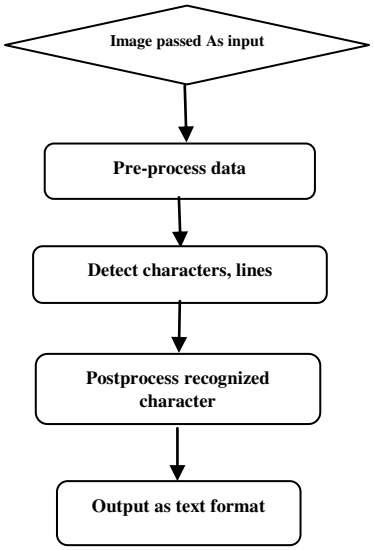
data from documents and then it will processed and analyzed. Using this, we can analyze the data so it will help in variety of operations.

IV. RESULTS AND DISCUSSION

The final result is carried out in following steps: firstly pre-process the input data, secondly it will detect characters and lines ,third it will postprocess the recognized data and lastly it will give the output as images. The final result of our project is to produce accurate text from images and scanned documents.This project will surely help you in real time application like Adhar-card recognition, license recognition and many more. So, we can say that this project provides features like detecting texts from various scanned documents.We have tested this project with the help of number of users and they all had given us good feedback. It is cost effective soit will helps in our daily requirements.

V. SIMULATION RESULT

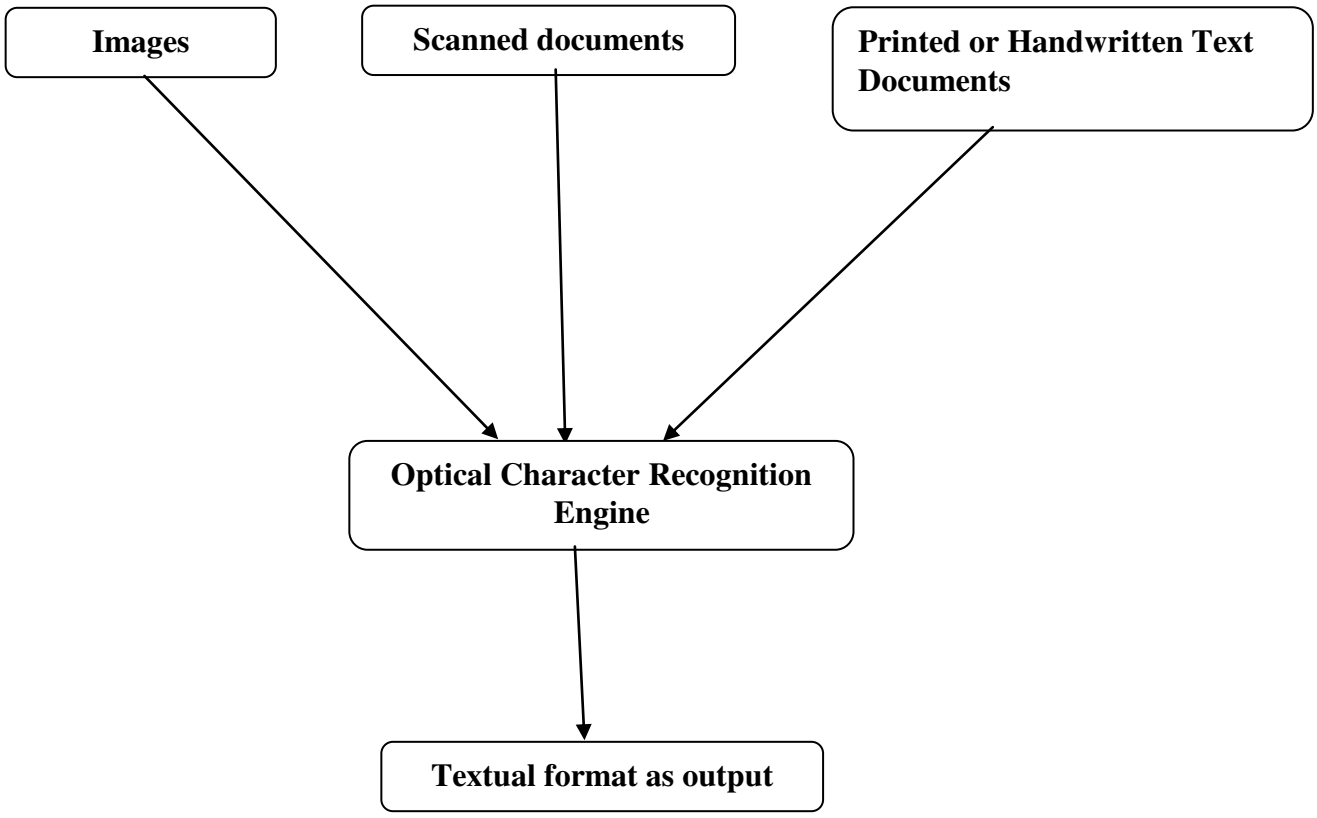
The simulation results will surely produced texts from the images and scanned documents as we passed as input. Our OCR engine will preprocess the data. Then it will detect all the lines, characters and words present in our input data. Then it will analyzed and recognized characters and will produced text from the input which is passed by user like images, scanned document and handwritten documents as a output. All the documents such as images, typed or handwritten or printed text as well as scanned documents will gone through OCR engine then it will produce output as text.



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003056 |



VI. CONCLUSION

This project will provide us a way to extract text. It will surely reduce the manual task of labours. It provides us way to edit a documentslike scanned or pdf, organizing documents and understanding text. The process of extracting the text from the images, scanned documents, handwritten text or printed text is still very much challenging process. Now, there is large amount of unstructured data we have seen, so there is big need of extracting this text for various applications. OCR will convert document into images, due to this reason there is large amount of demand for OCR. So, we have developed this software to reduced this problem which will surely help you.

REFERENCES

[1] "CT-1205CL-SMT Buzzer." Retrieved from http://www.digikey.com/product-detail/en/CT-1205CL-SMT/102-1267-1-ND/610975.
[2] "XM7 USB port Data sheet." Retrieved from http://www.digikey.com/product-detail/en/XM7A-0442A/OR1070-ND/2755612
[3] "TPS61032 (ACTIVE) 5-V Output, 1-A, 96% Efficient Boost Converter." Texas Instruments, Jan 2012.
[4] "LM 2679-5.0 (ACTIVE) 5-V Output, 5-A, 96% Efficient Buck Converter." Texas Instruments, Jan 2012.
[4]"IEEE Code of Ethics" Retrieved from http://www.ieee.org/about/corporate/governance/p7-8.html
https://www.ibm.com/cloud/blog/optical-character-recognition
https://www.hyland.com/en/resources/terminology/data-capture/what-is-optical-character-recognition-ocr
https://anyline.com/news/what-is-ocr

Books:
1)The Robotic Process Automation handbook by Tom Taulli
2)Robotic Process Automation by Richard Murdoch.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details