Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue S-3 Year 2023 Page 1083:1092

OCR Using Python and Its Application

Dr. Sumita Mukherjee[1], Hritik Tyagi[2], Purushautam Tyagi[3], Nikita Singh[4], Shraddha Bhardwaj[5]

[1]*Assistant Professor Amity University Dr Sumita Mukherjee, Email: Smukherjee2@amity.edu
[2]MCA Student, Amity University, Email: hritiktyagi729@gmail.com
[3]MBA Student, Amity University, Email: purushautamtyagi@gmail.com
[4]Student of MBA, Amity University, Email: nikitasin123@gmail.com
[5]Student of MBA, Amity University, Email: shraddha2407bhardwaj@gmail.com

***Corresponding Author:** Dr. Sumita Mukherjee
Email: Smukherjee2@amity.edu

Article History

Received: 18 June 2023
Revised: 01 September 2023
Accepted: 17 October 2023

Abstract

Optical Character Recognition (OCR) of papers has tremendous practical value given the prevalence of handwritten documents in human exchanges. A discipline known as optical character recognition makes it possible to convert many kinds of texts or photos into editable, searchable, and analysable data. In the past ten years, academics have developed systems that automatically evaluate printed and handwritten documents to convert them to electronic format. In the modern era, as demand for computer systems arose, the demand to convert paper text and computer vision also erose. To interact the computer with capability to read text from images, videos and images have been arose rapidly and many software companies came in role to fulfil this need. One of the active and difficult study areas in the world of pattern recognition and image processing has been handwriting recognition. Among its many uses are bank checks, reading assistance for the blind, and the conversion of any handwritten document into structural text. The main aim of this paper is to create a searchable pdf from the image and bring the application to easy use and deployable on premises and cloud.

Keyword: System, Software, Development

1. Introduction

OCR means Optical Character Recognition. A technology which is used to convert scanned images, PDFs, and other types of documents into machine-readable text. OCR software works by analysing the patterns of light and dark pixels in an image and trying to recognize the characters represented by those patterns.

OCR is commonly used in document management systems, where it is used to convert paper documents into digital form so that they can be easily searched, indexed, and archived. OCR is also used in the development of assistive technologies for people with visual impairments, as it can be used to convert printed text into speech or braille.

Overall, OCR technology is a useful tool for organizations and individuals looking to digitize and analyse information from paper documents. As OCR technology continues to evolve, it is likely that we will see even more advanced capabilities in the future, making it easier and more efficient to work with large amounts of information in digital form.

OCR technology has many advantages that make it useful tool for many applications, including:

1. **Timesaving:** OCR technology can automate the process of manually transcribing text, which can save a lot of time and effort, particularly for organizations with large volumes of documents.

2. **Improved accuracy:** While OCR technology is not perfect, it can significantly reduce errors compared to manual data entry. It can improve the accuracy and reliability of data and reduce the risk of costly mistakes.
3. **Better data analysis:** By digitizing text, OCR technology makes it possible to analyse and process data more efficiently. This can provide insights and information that may not have been possible to obtain from paper documents.

Overall, OCR technology has numerous applications across many different industries and can provide significant benefits in terms of efficiency, accuracy, and accessibility. As OCR technology continues to evolve, it is likely that we will see even more applications in the future.

OCR technology can be implemented using various methods, including:

1. **Software-based OCR:** This method involves using OCR software to scan and convert paper documents into digital formats. The software typically uses optical character recognition algorithms to detect and convert text characters.
2. **Cloud-based OCR:** This method involves using OCR technology that is hosted on cloud servers. Users can upload documents to the cloud service, where they are automatically processed and converted into digital formats.
3. **Mobile OCR:** This method involves using OCR technology that is integrated into mobile devices, such as smartphones and tablets. This allows users to scan and convert documents on-the-go, without the need for additional equipment.

OCR technology is not without its limitations and challenges. Some of these include Accuracy issues, language barriers, layout issues, cost, security, and privacy.

Data Collection and Storage:

We stored all the intermediate images in a folder. At the end we store all the collected and processed data in a CSV file. Extracted data will appear in this format and can be saved in a text file or csv file.

```
ACCOUNT NUMBER ,       MICR CODE            ,         CHEQUE NAME   |
911020061653442 ,c902111c 11021301205 066580c 00 ,C:\Users\satya1\Desktop\cropte
911020061653443 ,c902111c 11021301205 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653444 ,c902111c 11021301208 0665800c 00 ,C:\Users\satya1\Desktop\cropt
911020061653445 ,c902111c 11021301205 06650808d 00 ,C:\Users\satya1\Desktop\crop
911020061653446 ,c902111c 11021301208 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653447 ,c902111c 11021301208 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653448 ,c902111c 11021301205 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653449 ,c902111c 11021301205 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653451 ,c902111c 11021301205 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653450 ,c902111c 11021301200 06658080 00 ,C:\Users\satya1\Desktop\cropt
911020061653372 ,c9002111c 11021301205 0665808d 00 ,C:\Users\satya1\Desktop\crop
```

2. Literature review:

The method of categorizing optical patterns in relation to alphanumeric or other characters is known as optical character recognition. Segmentation, feature extraction, and classification are also included [1]. This literature review explores Optical Character Recognition (OCR) technology, focusing on its implementation using Python and its diverse applications. The primary objective is to create searchable PDFs from images [2]. The process of converting printed text into editable text was done using optical character recognition (OCR) technology. OCR is an extremely useful and well-liked method that is used in many different applications. Techniques for text preparation and segmentation can affect OCR accuracy. [3] Various OCR methods, especially those employing Python, are discussed, including the use of OCR software such as PaddleOCR. Many works have been already done by many software companies like google, Microsoft, and amazon and some other companies which are paid and available on their cloud. [4]. Virtually generated 3D worlds have recently grown in prominence to reduce the requirement for manually tagged photographs. Unfortunately, producing realistic 3D content is difficult and time-consuming on its own [6]. Future research includes optimizing the system for mobile phone implementation with limited CPU and memory resources, and geo-tagging of the image using GPS coordinates and online database for various mobile applications [11]. Edge offers various textures and concise ideas about every object's shape. Over the past few decades, many

algorithms have been published and use. In this paper, a novel edge detection method is put forth to implement an ideal edge detection method, one that can handle varying light luminosity on colour images, operate under various lighting conditions, and offer the highest accuracy, maximum effectiveness, maximum signal-to-noise ratio (SNR), and minimum mean squared error (MSE) [12]. There is also some open-source software for OCR providing by some companies like tesseract-OCR by Google (Google have done excellent work on tesseract OCR but the open-source part of this software does not provide efficient results.) and some software are lanyocr, mmlabs, paddleocr, etc. The review concludes with insights into creating Docker images for Python applications and deploying them using Kubernetes. There are many ways to do so which are (i) OCR Software, (ii) Adobe Acrobat Pro, (iii) Google Drive, (iv) Microsoft Word, (v) Online OCR Tools. In this context we are using first one method which is **OCR Software.** In this method, we choose a OCR model to do extract text from the image of our choice using software and hardware resources. There is many OCR software in the market developed upon different programming languages. Some of which are open source, and some are paid ones. In this context, some of the OCR software examples given which are using python programming language and one which we are using to do OCR is paddleocr (paddleocr uses language python). In the end, we have discussed how to create docker image of the python application we have created and how to publish it on dockerhub and how to run it on on-premises using Kubernetes. It is most acceptable to group statistics, database technology, information discovery, pattern recognition, machine learning, business, natural disasters, and other fields under the umbrella of data mining [15]. Data mining will effectively introduce the computing strategies and techniques to retrieve the applicable and convenient information from combined large databases known as big data [16]. To do data mining we need to collect at least certain amount of data and some data maybe collect through images and videos. To process such type of data we need to convert data in images to computer readable text which can be done using OCR technology.

3. Materials and Methods

Methodology Using Python

Basically, OCR can be done using only python library or the software build upon these libraries.

1) OCR using python inbuilt libraries:

In python, OCR is done using **pytesseract** library. Steps to do ocr using python inbuilt library are:

i) first import cv function

ii) then open an image using imread function from cv2 library.

Iii) and then read text using function **image_to_string** from pytesseract library.

Algorithm:

```
import pytesseract
custom_config = r'-l eng+por --psm 6'
txt = pytesseract.image_to_string(img, config=custom_config)
print(txt)
```

2) Ocr using software build upon python libraries:

There are many libraries which are open source and can be used to do ocr. Some examples of such types of software are Open MM Labs, lanyocr, easyocr, paddleocr, ocropus, etc.

i) OCR using Open MM Lab:

Open MM Lab is an open-source platform that aims to promote research and development in the field of multimedia machine learning. It is an initiative launched by Multimedia Laboratory (MMLab) of The Chinese University of Hong Kong (CUHK) and is currently maintained by a team of developers from MMLab and other organizations.

Installation Guide:

```
conda create -n open-mmlab python=3.8 pytorch=1.10 cudatoolkit=11.3 torchvision -c pytorch -y
conda activate open-mmlab
pip3 install openmim
git clone https://github.com/open-mmlab/mmocr.git
cd mmocr
mim install -e .
```

Simple Program to do OCR ::
from mmocr.apis import MMOCRInferencer
ocr = MMOCRInferencer(det='DBNet', rec='CRNN')
ocr('demo/demo_text_ocr.jpg', show=True, print_result=True)

ii) OCR using lanyocr :::
LanOCR is an open-source Optical Character Recognition (OCR) system developed by the Natural Language Processing Group at Lanzhou University in China. OCR is a technology that enables machines to read and interpret text from scanned images, PDFs, or other documents. OCR is used extensively in document digitization, data extraction, and automated processing.

Installing Lanyocr ::
git clone https://github.com/JC1DA/lanyocr
cd lanyocr
pip install lanyocr

Code to do OCR ::
python detect.py --merge_rotated_boxes true --merge_vertical true --image_path images/example1.jpg
(example1.jpg is the image for ocr.)

iii) OCR using easyocr :::
EasyOCR is an open-source Optical Character Recognition (OCR) tool developed by Jaided AI, a startup based in Thailand. It is a Python-based library that enables machines to read and interpret text from images, PDFs, and other documents. EasyOCR has gained significant popularity due to its simplicity, ease of use, and multilingual support.

Installing easyocr ::
pip install git+https://github.com/JaidedAI/EasyOCR.git

Code to do ocr :::
import easyocr
reader = easyocr.Reader(['ch_sim','en']) # this needs to run only once to load the model into memory
result = reader.readtext('image_name')

iv) OCR using paddleocr :::
git clone https://github.com/PaddlePaddle/PaddleOCR
cd PaddleOCR
pip install paddleocr
pip install paddlepaddle

v) OCR using Ocropus
Installation::
Install outside python enviroment(anaconda):
git clone https://github.com/ocropus/ocropy
cd ocropy
pip install

Install inside python enviroment:
conda create -n ocropus_env python=2.7
conda activate ocropus_env
conda install --file requirements.txt.
wget -nd https://github.com/zuphilip/ocropy-models/raw/master/en-default.pyrnn.gz
mv en-default.pyrnn.gz models/
python setup.py install.
Docker: A platform called Docker enables programmers to automatically deploy, scale, and manage applications inside of small, portable containers. Software can be packaged together with its dependencies and runtime using containers so that it can function consistently in development, testing, and production environments, among others. Docker has established itself as a crucial tool for DevOps and contemporary software development.
Installing Docker on ubuntu:
1. Update the package index and install required packages: sudo apt update.
2. Add the official Docker GPG key to ensure the authenticity of the Docker repository: sudo apt install apt-transport-https ca-certificates curl software-properties-common
4. Add the Docker repository to APT sources: curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o /usr/share/keyrings/docker-archive-keyring.gpg
5. Update the package index again: sudo apt update
6. Install Docker: sudo apt install docker-ce
7. Start and enable the Docker service: sudo systemctl start docker sudo systemctl enable docker
8. Verify that Docker is running: sudo systemctl status docker
9. Optionally, add your user to the "docker" group to use Docker without sudo: sudo usermod -aG docker $USER

4. Result and Discussion

Result obtained by applying OCR Methodology

1. Searchable pdf:

In the era of digitalization, the importance of searchable PDFs has increased dramatically. A searchable PDF is a type of document that allows users to search for specific words, phrases, or characters within the text of the document. In this paper, we present a comprehensive study on the techniques and applications of searchable PDFs. We discuss the advantages and disadvantages of searchable PDFs over other types of documents. We also present various techniques for creating searchable PDFs, including Optical Character Recognition (OCR) and automated indexing. Finally, we explore the applications of searchable PDFs in various fields, including education, healthcare, legal, and business.

The PDF (Portable Document Format) is one of the most widely used document formats in the world. It is a file format that preserves the layout, fonts, and graphics of a document, regardless of the software or hardware used to view it. A searchable PDF is a type of PDF that includes an OCR (Optical Character Recognition) layer, which allows the text within the PDF to be searched, selected, and copied. The OCR technology recognizes text within the image of the document and then converts it into searchable and editable text.

Advantages of Searchable PDFs: Searchable PDFs have numerous advantages over other types of documents. For instance, they are much easier to search and navigate, which makes them more convenient for users. Additionally, they allow users to copy and paste text, which can save time and effort. Furthermore, searchable PDFs are more accessible to individuals with visual impairments, as they can use screen readers to read the text within the PDF.

Techniques for Creating Searchable PDFs: There are several techniques for creating searchable PDFs. One of the most common techniques is OCR, which involves using software to recognize the text within an image of a document and then converting it into searchable text. OCR technology has advanced significantly in recent years, and now it can recognize various fonts, languages, and even handwriting. Another technique for creating searchable PDFs is automated indexing, which involves automatically extracting and indexing the text within a document. This technique can be useful for large-scale document processing, such as digitizing archives and libraries.

Applications of Searchable PDFs: Searchable PDFs have numerous applications in various fields, including education, healthcare, legal, and business. In education, searchable PDFs can be used to create digital textbooks that allow students to search for specific concepts and keywords. In healthcare, searchable PDFs can be used to create patient records that can be searched and shared with other healthcare providers. In legal, searchable PDFs can be used to create electronic legal documents that can be searched and shared with other attorneys. In business, searchable PDFs can be used to create electronic contracts, invoices, and receipts that can be searched and shared with other business partners.

In conclusion, searchable PDFs are becoming increasingly important in the digital age, and their advantages over other types of documents are clear. By using OCR and automated indexing techniques, searchable PDFs can be created quickly and efficiently. Furthermore, their applications in various fields make them essential for modern document processing and management.

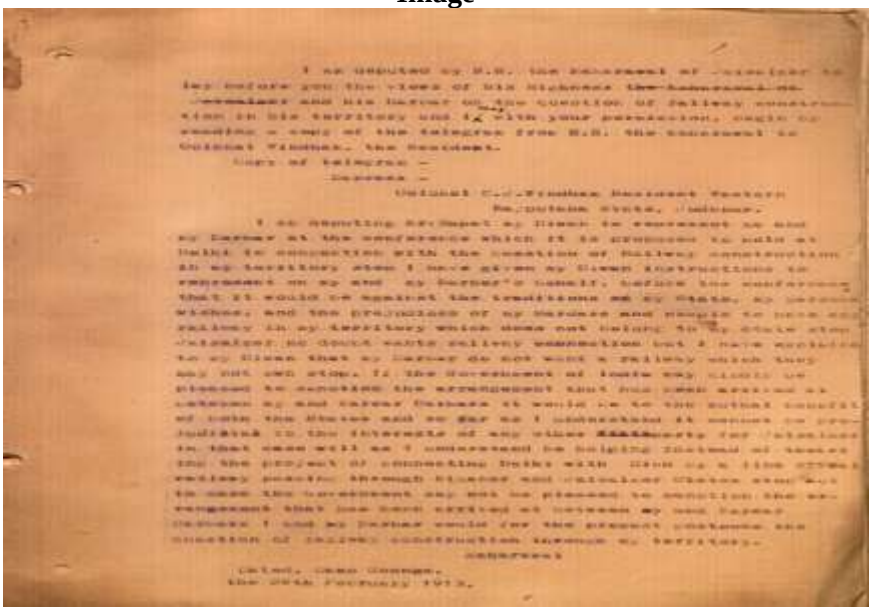
To create a searchable pdf from an image, we must follow some instructions which are:

i) To convert given image into pdf.
ii) now we need to extract text from the image, i.e., doing OCR on image (you can choose OCR software of your type but in this paddleocr is being chosen. *Important: writing OCR text on blank pdf is based on the output of the OCR software, in the case of paddleocr it returns the boxes of the lines of the text and text itself in a tuple and that information is used to plot character on the blank pdf.).
iii) and then we need another pdf which are of same width and size of image having written text as per written on the image (means that words on the image and another pdf should be on same coordinate).
iv) after this we must merge the image pdf and the text pdf (one thing to care about here is that image pdf should overlay over the text pdf.).

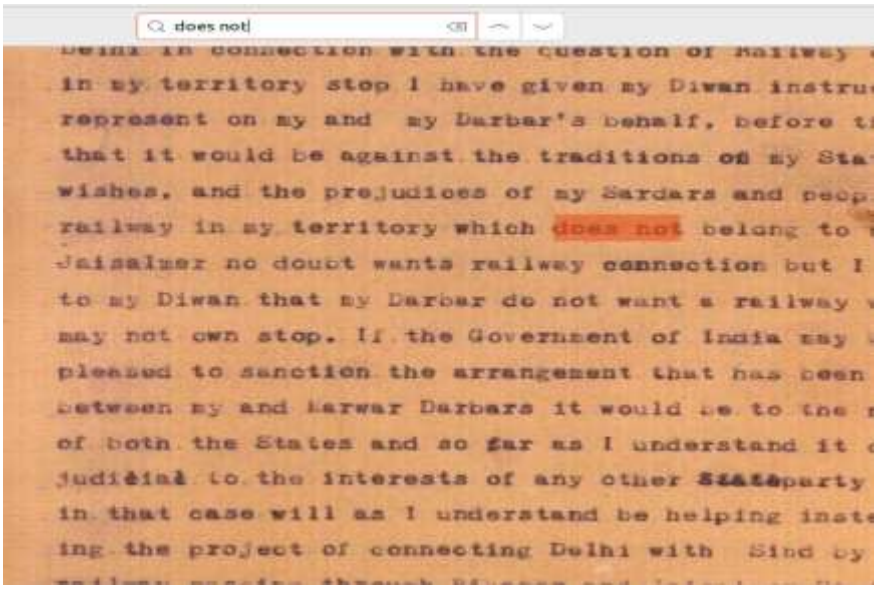
We have different part of code doing different work using the given image that are:
i) Convert to pdf
ii) OCR
iii) Draw on pdf
iv) Merge pages

OUTPUT

Image



Output:



Cheque Truncation

As per the Reserve Bank of India regulation, the check used by various banks in India must be as per the below template. This helps us quickly identify an area of interest in e-processed check.

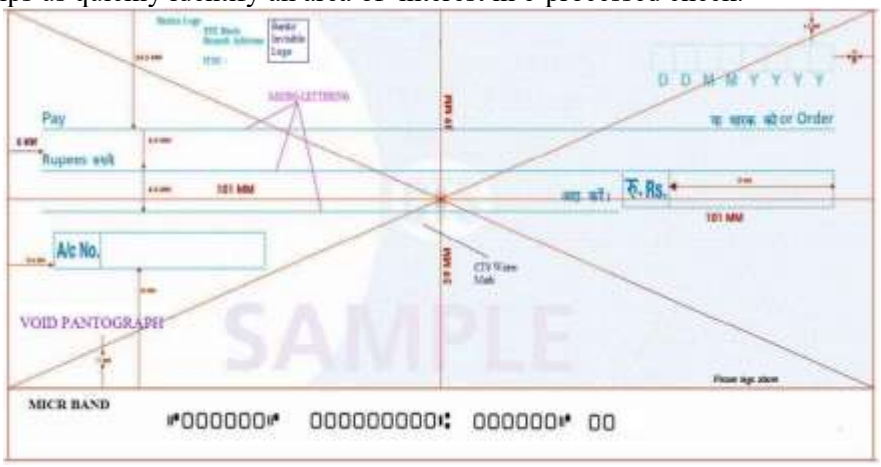


Image Registration and Digit Recognition

Image Registration: The process of image registration is an automatic or manual operation that attempts to discover matching points between two photographs and spatially align them to minimize the desired error, i.e., uniform distance measurement between two images. To perform image registration, our approach is to find the image homograph and the region of interest of the template image. We perform a perspective distortion operation using homograph and then take the difference between the image and the template image.

Date Digit Recognition:

The date consists of handwritten digits. The date position can be determined using the dimension shown on the CTS check.

Access:

1. We will perform image registration and template matching to remove frames and printed text from the image.
2. We use the MNIST dataset, which consists of 60,000 training images and 10,000 test images of handwritten digits with labels.
3. Using the dataset, we used the K-Nearest Neighbor (KNN) classification method to predict the digits.

4. After performing the template matching, we found the number of digits using the measure label that indicate the connected area of the image.
5. Then we resized each digit image to 28*28 and then used our KNN trained model to predict digits from the image.

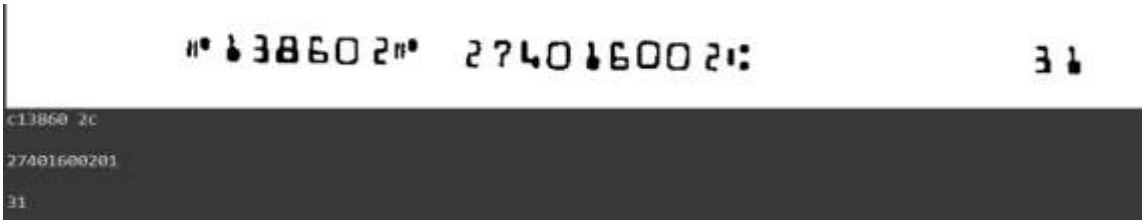
Amount Digit Recognition:

The amount in digits also consists of handwritten digits. The location of the amount box can be determined using the dimension shown on the CTS check.
Access:

1. To predict the amount, we will use the same method as before.
2. We will perform Image Registration and Template matching, and then use the previously built model using the MNIST dataset and the KNN classifier.
3. After performing the template matching, we used the measure label command to find the number of digits that indicate the linked area of the image.
4. Then we resized each digit image to 28*28 and then used our KNN trained model to predict digits from the image.

MICR Code recognition:

The MICR is a 9-digit code printed on the bottom of the check and helps identify the check.
The font used for MICR is unique and cannot be recognized by regular OCR.
Access:
1. We use pytesseract which is a wrapper for Google Tesseract-OCR Engine
2. We used a trained model for MICR characters. Link to the model
3. Using the trained data for OCR, we recognized MICR characters.



4. Account Number:

1059163844

4. Signature:

Here, we used the concept that the size of the handwritten part of the signature will be greater than the printed part of the cheque.
After that we know that the signature would be found in the bottom right quarter of the cheque, hence we have extracted the signature from there.

Handwritten text:

Handwriting OCR (optical character recognition) is the extraction of handwritten text from paper, scans, and other low-quality digital documents. Before manual OCR, there was traditional OCR. Traditional OCR relies on fonts and symbols that have been sufficiently "researched" to detect virtually all permutations of machine-printed text. But therein lies the limitation of classic OCR: while it is excellent at extracting text from paper, it is unable to read handwriting. Simply put, there is too much variety.
OCR for handwritten text requires significantly more advanced technology than standard OCR. This kind of optical character recognition uses a well-trained machine learning model and powerful computer vision engines to interpret text like a human. Machine learning is a subfield of artificial intelligence that allows computers to automatically learn and iterate from experience without explicit instructions, instead relying on patterns and

inference. Another category of artificial intelligence that can automate tasks that the human visual system can perform is computer vision. A combination of highly trained machine learning models and computer vision engines enables handwriting OCR to mimic the way humans read handwriting.
Paddle OCR is an easy-to-use and open-source OCR repository that provides ultra-lightweight OCR systems and more than 80 types of multilingual recognition models. We use Paddle OCR to read both handwritten and printed Hindi text.

1. Payee Name:

2. Amount:

3. Date:

10072050

5. Conclusion:

In conclusion, OCR technology has changed the way we deal with and communicate with textual information. Its efficiency, accuracy, availability, cost-effectiveness, and ability to integrate with other technologies make it a valuable tool across various industries, simplifying operations, improving accessibility, and opening new possibilities for data analysis and automation.
In this paper, by using OCR technology, we have successfully created a searchable pdf and cheque processing which is beneficiary for many types of companies (there are many more things we can do using OCR). And tell how to create docker image of python application and show a demo to run it. In searchable PDFs are becoming increasingly important in the digital age, and their advantages over other types of documents are clear. By using OCR and automated indexing techniques, searchable PDFs can be created quickly and efficiently. Furthermore, their applications in various fields make them essential for modern document processing and management.

References:

1. R. Parthiban, R. Ezhilarasi and D. Saravanan, "Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262379.
2. Olalekan Joseph ONI, Franklin Oladiipo ASAHIAH, Computational modelling of an optical character recognition system for Yorùbá printed text images, Scientific African, Volume 9,2020,e00415,ISSN2468-2276,https://doi.org/10.1016/j.sciaf.2020.e00415.(https://www.sciencedirect.com/science/article/pii/S2468227620301538)
3. M. R. M. Ribeiro, D. Jùlio, V. Abelha, A. Abelha and J. Machado, "A Comparative Study of Optical Character Recognition in Health Information System," *2019 International Conference in Engineering Applications (ICEA)*, Sao Miguel, Portugal, 2019, pp. 1-5, doi: 10.1109/CEAP.2019.8883448.
4. Raj, A. D., & Ramya, K. Handwritten text recognition using OCR with deep learning techniques. International Journal of Innovative Technology and Exploring Engineering, 11(6), 582-586. DOI:2022.
5. Sinha, A., Verma, V., & Das, S. K. (2022). Optical character recognition using Python: A comprehensive review. International Journal of Engineering Science and Computing, 12(3), 32440-32448.
6. Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, An-dreas Geiger and Carsten Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes", *International Journal of Computer Vision*, vol. 126, no. 9, pp. 961-972, 2018.

7. M. Rizvi, H. Raza, S. Tahzeeb and S. Jaffry, "Optical Character Recognition Based Intelligent Database Management System for Examination Process Control," *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Islamabad, Pakistan, 2019, pp. 500-507, doi: 10.1109/IBCAST.2019.8667127.
8. Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu. "A fast adaptive binarization method for complex scene images." In Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 1889-1892. IEEE, 2012.
9. Sumetphong, Chaivatna, and Supachai Tangwongsan. "An Optimal Approach towards Recognizing Broken Thai Characters in OCR Systems." Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on. IEEE, 2012.
10. Al Salman, Abdul Malik, et al. "A novel approach for Braille images segmentation." Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE, 2012.
11. Mutholib, Abdul, Teddy Surya Gunawan, and Mira Kartiwi. "Design and implementation of automatic number plate recognition on android platform." Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012.
12. Jana, P., & Sarkar, A. (2023). Secure Data Transmission in the Era of 6G: Challenges and Solutions. *Algorithm Asynchronous*, *1*(1), 8–15. Retrieved from https://hasmed.org/index.php/jourasy/article/view/46
13. Harguess, J., & Harvey, D. Optical character recognition using Python. Journal of Information Systems Education, 29(3), 175-182.DOI: 2018.
14. Sahay, R., & Bharti, P. Optical character recognition for printed Devanagari script using Python. International Journal of Recent Technology and Engineering, 8(2S3), 77-81.DOI:2019.
15. Das, A., & Dutta, S. An overview of OCR with Tesseract using Python. International Journal of Computer Sciences and Engineering, 8(7), 238-242.DOI:2020.
16. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13272-13281 © Research India Publications. http://www.ripublication.com Scopus Indexed. Integrating Application with Algorithms of Association Rule used in Descriptive Data Modelling, through which Data Mining can be Implemented for Future Prediction.
17. Retrieval Number: B2448078219/19©BEIESPDOI: 10.35940/ijrte.B2448.078219 Journal Website: www.ijrte.org International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019. Collaborating Data Mining Modeling with Big Data Analytics for Disaster Prediction