iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45670

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Extract and Organize Information in Images with AI using IBM Services

Anu C S[1], Reshmabanu M Badiger[2], Sowmya N M[3], T Kavya[4], Soundarya Ramesh Raykar[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, BIET Davanagere*

Abstract: ***OCR is a short form of Optical character recognition or optical character reader. By the full form, we can understand it is something that can read content present in the image. Every image in the world contains any kind of object in it and some of them have characters that can be read by humans easily, programming a machine to read them can be called OCR. In machine learning, data mining is one of the major sections that cover the extraction of the data from the different platforms. OCR (Optical Character Recognition) is part of the data mining process that mainly deals with typed, handwritten, or printed documents. These documents hold the data mainly in the form of images. Extracting such data requires some optimised models which can detect and recognize the texts. Getting information from complex structured documents becomes difficult and hence they require some effective methodologies for information extraction. In this article, we will discuss OCR with IBM Watson Natural Language Understanding API, a deep learning-based tool for localizing and detecting the text in documents and images.***
Keywords: OCR, API, IBM and Natural Language.

I. INTRODUCTION

In the running world, there is growing demand for the software systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. These days there is a huge demand in "storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process". One simple way to store information in these paper documents in to computer system is to first scan the documents and then store them as IMAGES. But to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word. The reason for this difficulty is the font characteristics of the characters in paper documents are different to font of the characters in computer system. As a result, computer is unable to recognize the characters while reading them. This concept of storing the contents of paper documents in computer storage place and then reading and searching the content is called DOCUMENT PROCESSING. Sometimes in this document processing we need to process the information that is related to languages other than the English in the world. For this document processing we need a software system called CHARACTER RECOGNITION SYSTEM. This process is also called DOCUMENT IMAGE ANALYSIS (DIA). Thus our need is to develop character recognition software system to perform Document Image Analysis which transforms documents in paper format to electronic format. For this process there are various techniques in the world. Among all those techniques we have chosen Optical Character Recognition as main fundamental technique to recognize characters. The conversion of paper documents in to electronic format is an on-going task in many of the organizations particularly in Research and Development (R&D) area, in large business enterprises, in government institutions, so on. From our problem statement we can introduce the necessity of Optical Character Recognition in mobile electronic devices such as cell phones, digital cameras to acquire images and recognize them as a part 2 of face recognition and validation.

To effectively use Optical Character Recognition for character recognition in-order to perform Document Image Analysis (DIA), we are using use the IBM Watson™ Natural Language Understanding API to extract entities from documents using Jupyter Notebooks, and use a configuration file to build configurable and layered classification grammar.

II. LITERATURE SURVEY

In [1], The camera captured images containing text are having curved text lines because of distortions by page curl and the view angle of the camera. So it is necessary while scanning the document, the text should be straight and words are inline properly. But text lines segmentation of curled text is a difficult method for de-wrapping techniques. This paper presents the method based on image processing algorithms for segmentation and extraction of characters from curled text lines from document images. The algorithm performs the curved text segmentation using x-line and base line.

The words in the document image are identified and bounding boxes are plotted around the words. The properties of connected components are used for segmentation of words.

In [2], A new information extraction system by statistical shallow parsing in unconstrained hand-written documents is introduced. An entire text line is considered as an indivisible entity and is modeled with Hidden Markov Models. In this way, text line shallow parsing allows fast extraction of the relevant information in any document while rejecting at the same time irrelevant information. In this, two kinds of information have to be learnt 1. Character models (HMM- Hidden Markova models by Baum-Welch algorithm) 2. Transition between them with the help of probabilities.

In [3], An automatic document entry system is described that identifies the type of document and extracts textual information, such as titles or authors, from semi-formatted document images. The system registers documents, offers easy retrieval of documents used in a daily workflow, analyses the layout structure of documents by using document specific models, and assumes that each type of document is known in advance. In this paper we focus on a method for identifying the type of document.

In [4], Extraction of texts from scanned copies of documents and text images is an important task in the recent scenario. Optical Character Recognition(OCR) is used to analyses text in images. The proposed algorithm deals with taking a scanned copy of a document as an input and extract texts from the image into a text format using Otsu's algorithm for segmentation and Hough transform method for skew detection. The system was confined to recognize English alphabets (A-Z, a-z) and numerals (0-9). OCR technique has been implemented to recognize characters. Validation tests were done on screenshots of typed texts and images of scanned documents from Internet sources. Experimental results indicate that the proposed algorithm is able to recognize alphabets written in Verdana font style with size 14 and a showed good results with rotated images. The proposed algorithm is successfully able to recognize characters from text images with an average accuracy of 93%. It also showed good results against rotation and scaling and was able to reduce noise from images to a good extent. The average rotation accuracy to correctly rectify skew from images was calculated to be 90%.

In [9], A method which takes the advantage of the gradient gray level to divide the 2D histogram region, and applies the traditional Otsu's thresholding method twice on two projection histograms to separate the regions of interest from the background. The experimental results show that method is robust against noise, and it requires less computational time. The objective of thresholding is to extract objects or regions of interest in an image from the back ground based on its gray level distribution. Many thresholding methods are exhaustively described and evaluated based on different error measures. One well-known thresholding method is Otsu's which selects the optimal threshold by maximizing the 'between-class variance'. However, using only a 1D 27 histogram of an image cannot reflect spatial information between image pixels, it is difficult to obtain satisfactory results when images contain noise.

Lui et al, thus extended the use of a 1D-histogram into a two dimensional (2D) histogram, which utilizes not only the gray level distribution of pixels, but also the average gray level distribution of their neighborhood to select the optimal threshold vector. This method gives better thresholding results than 1D Otsu's method, but it requires longer execution time.

In [10], getting an efficient method of removing noise from the images, before processing them for further analysis is a great challenge for the researchers. Noise can degrade the image at the time of capturing or transmission of the image. Before applying image processing tools to an image, noise removal from the images is done at 28 highest priority. The kind of the noise removal algorithms to remove the noise depends on the type of noise present in the image. Best results are obtained if the testing image model follows the assumptions and fails otherwise. In this paper, light is thrown on some important type of noise and a comparative analysis of noise removal techniques is done. This paper presents the results of applying different noise types to an image model and investigates the results of applying various noise reduction techniques.

III. SYSTEM DESIGN

System design thought as the application of theory of the systems for the development of the project. System design defines the architecture, data flow, use case, class, sequence and activity diagrams of the project development.

A. OCR

OCR is a short form of Optical character recognition or optical character reader. By the full form, we can understand it is something that can read content present in the image. Every image in the world contains any kind of object in it and some of them have characters that can be read by humans easily, programming a machine to read them can be called OCR. Programmatically we can say that it is the process of converting images of typed, handwritten, or printed text into machine-encoded text.

We mainly find the usage of this type of program in extracting data from printed paper data, the example of printed paper can be passports, invoices, statements, business cards, etc. OCR can also be considered the base programming for various projects like text mining, text-to-speech, cognitive computing, etc. while OCR is a program that is related to the field of computer vision in machine learning.

Talking about history, it all started with training models with images of all kinds of characters that need to be detected from documents. Nowadays we can find that there are models which are capable of detecting and recognizing characters with a high degree of accuracy for any kind of font and from any kind of document. Also, various models are capable of generating documents similar to the original and we can edit the generated document.

B. *Types of OCR*

There are various types of OCR. Some of the basic types of OCR are as follows:

1) *Optical character recognition(OCR):* Targets only a text at a time from the documents.
2) *Optical word recognition(OWR):* Targets one word at a time from the documents.
3) *Intelligent character recognition(ICR):* It is an update of standard OCR that can be capable of targeting one character at a time that is in the form of handwriting or cursive.
4) *Intelligent word recognition(IWR):* It is an update of standard OWR that can be capable of targeting one word at a time that is in the form of handwriting or cursive.

We can also categorize OCRs by their type of work environment in the following way:

a) *Offline OCR:* This kind of OCR works in offline mode where we normally use offline documents for the recognition of characters or words.
b) *Online OCR:* This kind of OCR is generally enabled in cloud storage and they are capable of recognising characters from the documents that are present in the cloud.
c) *Dynamic OCR:* This kind of OCR is capable of recognising characters from dynamic documents. Here the dynamic documents are those that keep characters and words in motion.

C. *Applications of OCR*

Before going deep into OCR implementation we are required to know the places where we may be required to use OCR.

Some examples of such use cases are as follows:

1) Data entry.
2) Number plate recognition.
3) Traffic sign and vehicle recognition.
4) Information extractor from important documents such as passports, Aadhar cards, etc.
5) Scanned data editor.
6) Electronic books from printed books.
7) Assistive technology for blind and visually impaired users.

D. *System Architecture*

Following Modules are implemented and shown in the architecture diagram in fig. 1

1) Sign up for IBM Watson Studio
2) Classification of image Documents
3) Text Extraction Using Optical Character Recognition
4) Entity Extraction and Document Classification
5) Analyse the Results

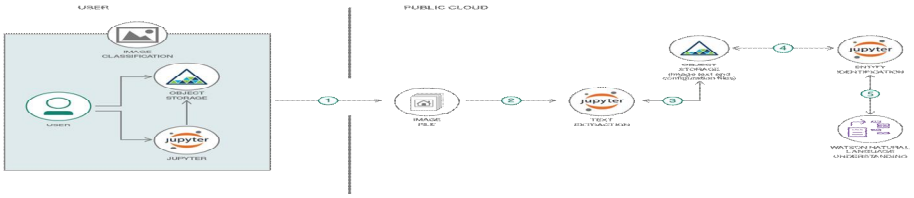


Fig. 1 Architecture Diagram

If we have not already signed up for Watson Studio then you can sign up here. By signing up for IBM Watson Studio, two services will be created - Spark and Object Store in your IBM Cloud account.

Code pattern demonstrates how images, specifically document images like id cards, application forms, cheque leaf, can be classified using Convolutional Neural Network (CNN). Even though there are code patterns for image classification, none of them showcase how to use CNN to classify images using Keras libraries. Many organisations process application forms, such as loan applications, from its customers. Along with the application forms, customers provide supporting documents needed for processing applications. Some of these supporting documents could be identity proof document, address proof document. Generally application forms, along with supporting documents, are scanned and captured into the organisation's systems for further processing of applications. When the system is fed with a set of all these scanned documents, it needs to identify the form document so that it can process it further. This code pattern shows how to classify images and identify application form document among them. This code pattern covers the following aspects: Dataset preparation for training and testing and running notebook for image classification.

Previous section identified application form document among the list of all image documents. This section extracts text from the application form document that was identified in the above section. The text is then saved as text document on to Object Storage, that was created in Part1 of this code pattern. We will use tesseract OCR for text extraction. We need to install tesseract engine on our local machine. And so we will run the next notebook on local. The output of this section will be text content of image document, which will be saved to Object Storage as form-doc-x.txt, where x is the nth document. e,g, if it's first form document the file is stored as form-doc-1.txt. This file will be used later by another notebook to extract information from text extracted.

Extracted text is stored in Object Storage. The following processes occurs:

- Add the extracted text file to Watson Studio Create IBM Cloud services
- Create the following IBM Cloud service and give a unique name for the service: Watson Natural Language Understanding
- Create notebook
- Upload text data and configuration data to Object Storage
- Create Watson Natural Language Understanding (NLU) service
- Add Watson NLU credentials to notebook
- Add Object Storage credentials to notebook

Analyse the result.

Featured technologies

- Jupyter Notebooks: An open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.
- Artificial Intelligence: Intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans.
- Machine Learning: Uses statistical techniques to give computer systems the ability to "learn" with data
- Natural Language Processing: the ability of a computer program to understand human language as it is spoken. NLP is a component of Artificial Intelligence
- Python: An interpreted high-level programming language for general-purpose programming

IV. RESULTS

Fig. 2 Home Page

This page will give you information about OCR and upload button to upload the document for text extraction

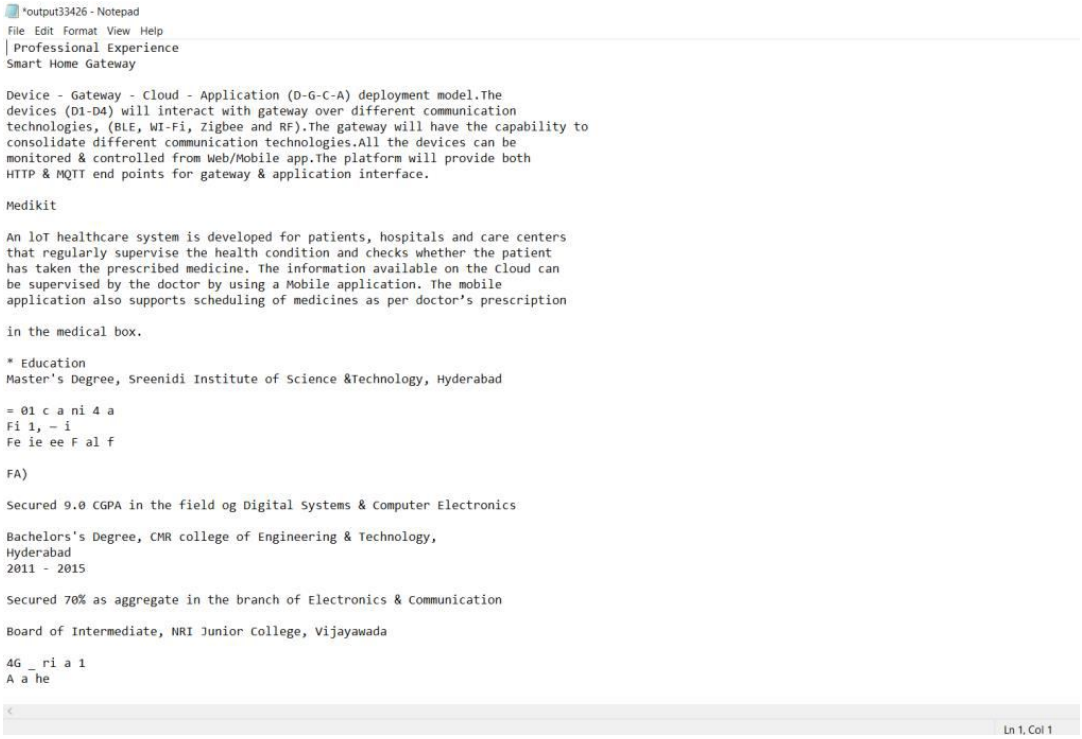


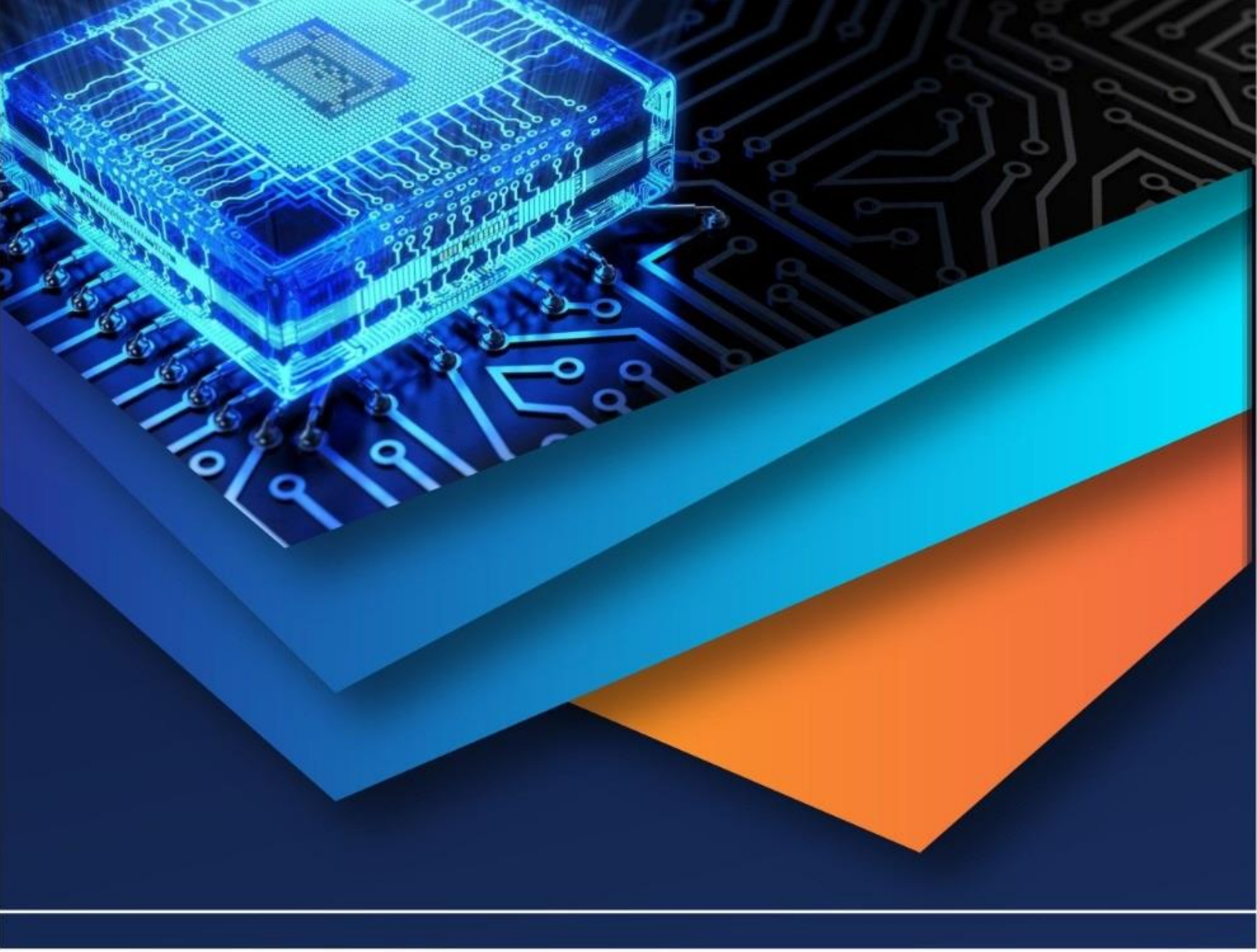
Fig. 3 Displays the text document where the output is stored

V. CONCLUSION

In this work, we have discussed the optical character recognition(OCR) that is used to extract information from printed documents and we have discussed the types and use cases of OCR. Along with this, we have seen how we can implement OCR using the IBM services. We extracted text using optical character recognition; also we use the IBM Watson Natural Language Understanding API to extract entities from documents using Jupyter Notebooks, and use a configuration file to build configurable and layered classification grammar. Thus, we developed an efficient and effective system for extracting information from the images as well as pdf documents. Future investigations on other aspects need to be pursued for developing solid image to text detection and recognition applications and related multimedia retrieval and annotation applications.

REFERENCES

[1] Shejwal, M. A., & Bharkad, S. D. (2017). Segmentation and extraction of text from curved text lines using image processing approach. 2017 International Conference on Information, Communication, Instrumentation and Control.
[2] Thomas, S., Chatelain, C., Heutte, L., & Paquet, T. (2010). An Information Extraction Model for Unconstrained Handwritten Documents. 2010 20th International Conference on Pattern Recognition.
[3] Kochi, T., & Saitoh, T. (1999). User-defined template for identifying document type and extracting information from documents. Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99.
[4] Agrawal, N., & Kaur, A. (2018). An Algorithmic Approach for Text Recognition from Printed/Typed Text Images. 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).
[5] S. P. Chowdhury, S. Mandal, A. K. Das. Automated Segmentation of Math-Zones from Document Images
[6] Zhen-Long BAI and Qiang HUO. An Approach to Extracting the Target Text Line from a Document Image Captured by a Pen Scanner.
[7] K.N. Natei, J. Viradiya, S. Sasikumar. Extracting Text from Image Document and Displaying Its Related Information.
[8] Chandan Singha,Nitin Bhatiab,Amandeep Kaurc. Hough transform based fast skew detection and accurate skew correction methods. 65
[9] Puthipong Sthitpattanapongsa and Thitiwan Srinark. A Two-stage Otsu's Thresholding Based Method on a 2D Histogram.
[10] U. Bhattacharya, S. K. Parui and S. Mondal, Devanagari and Bangla Text Extraction from Natural Scene Image.
[11] https://developer.ibm.com/patterns/image-recognition-and-information-extraction-from-image-documents-using-keras-and-watson-nlu/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY