iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53017

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 11 Issue V May 2023- Available at www.ijraset.com

FORM-based Document Understanding Sequential Model

Mr. Chetan Dalave[1], Mrs. Anushka Lodh[2], Mrs. Monika Bandal[3], Mrs. Priyanka Awalekar[4], Mrs. Vanita Babanne[5]
[1, 2, 3, 4]*Students* [5]*Asst Prof BE Computer Engineering, RMDSSOE Savitribai Phule Pune University, Pune, India*

Abstract: ***Image recognition and optical character recognition technologies have become an integral part of our daily lives due to the ever-increasing power of computing and the ubiquity of scanning devices. Sequence modeling has performed state-of-the-art on natural. Language and document comprehension tasks. However, this is the challenge of properly serializing tokens into form-like documents in practice due to the variety of layout patterns. We propose FormNet, a structure-aware layout model that reduces most serialization of forms. Form-based documents with the help of the OCR, we can search and recognize the text in documents and can easily convert them into human-readable text. It motivates us to make life simple to read the data which is unstructured as we convert it into a structured and simple format so that it becomes easy to read vast textual data.***

Keywords: Optical character recognition (OCR), Pre-processing, Classification Technique, Feature Extraction

I. INTRODUCTION

Form-Based Document understanding using a sequential model which will make unstructured data in a structured format like a key-value pair in csv format that will follow a left-to-right and top-to-bottom pattern. Printed documents can be changed quickly. Digital text files through optical character recognition and then edited by the user. This study shows how image-processing technologies can be used. Combination with optical character recognition to improve recognition accuracy improves the performance of extracting text from images. The results of the experiment show that the proposed system can accurately recognize text in images. Training and testing are done using pre-processed data sets i.e. PDF/DOC datasets used. This Image text recognition can be done using OCR (Optical Character Recognition). With the help of the OCR, we can search and recognize the text in documents and can easily convert them into human-readable text.

As Form based document is a growing research topic it has some limitations one of which is that through OCR it cannot scan and convert Handwritten text samples in a structured form. So, in this model, we are going to cover this limitation which will become a new feature of the Form-Based Document Understanding Sequential Model. Basically, this new model with the help of OCR will make human life easy because of its practical application automating the process of extracting and organizing valuable text data sources such as marketing documents, advertisements, and invoices.

To Model, the information in the structural form present in [1] Nomen Islam, Zeeshan Islam, and Nazia Noor (2016) introduced the OCR System that is Optical Character Recognition to solve the problem of unstructured text recognition. OCR is used to solve complex problems but becomes difficult because of different languages, fonts, and styles. There is a Review on OCR [2]Jay Dilipbhai Thanki, Priyanka Dineshbhai Davdaand Dr. Priya Swaminarayan(2021) in which all information on OCR i.e uses, application, advantages and how it will work is given. With the help of all this information, we are going to form a system of Form-Based Document Model.

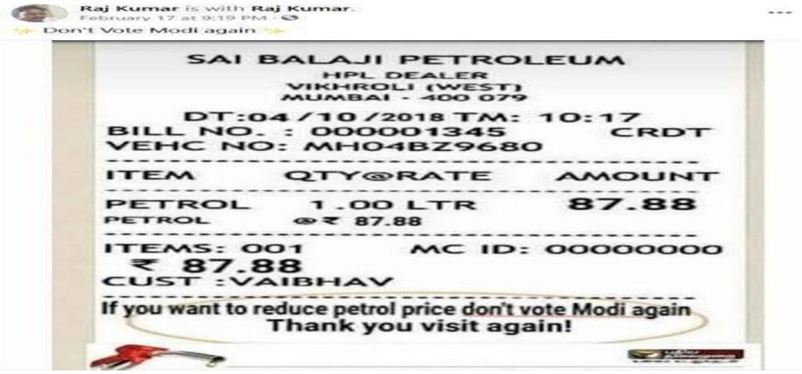


Figure 1: An example of a form document for extracting information

In the system, firstly the user has to register by putting all the details required and then login into the account after that a user can put the pdf of text image which can be any type of text document for example Bill of Grocery, etc. Users can put images of both printed and handwritten documents after adding an image OCR algorithm will start to extract text data from the image and then it will perform pre-processing on that text data such as Segmentation, Feature extraction, Classification, and lastly extract text-by-text recognition finally after all the process and module completion of the OCR algorithm the output will be generated by the system in the form of Key-value pair in csv in a sequential manner in a properly structured format .

II. LITERATURE SURVEY

Sr. No.	Paper Title	Author	Year	Advantages	Disadvantages
1	FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction	Chen-Yu Lee† , ChunLiang Li† , Timothy Dozat‡ , Vincent Perot‡ , Guolong Su‡ , Nan Hua‡ , Joshua Ainslie‡ , Renshen Wang‡ , Yasuhisa Fujii‡ , Tomas Pfister†	2022	Unstructured data get converted into structural form with help of OCR algorithm	High costs. Limit communicatio n and collaboration
2	A Novel Method based on Character Segmentation for Slant Chinese Screen-render Text Detection and Recognition	Tianlun Zheng Xiaofeng Wang Xin Yuan and Shiqin Wang	2020	• Data validation methods can be used in FormBased Interfaces.	Document transportation . Editing problems
3	Summary of Scene Text Detection and Recognition	Yao Qin Zhi Zhang	2020	• No excessive training is required.	Security issues. Prone to damage
4	Research on Text Detection and Recognition Based on OCR Recognition Technology	Yuming He	2020	Larger processing power or memory is not needed.	• High costs. Limit communicatio n and collaboration.
5	Novel Approach for Image Text Recognition and Translation	Srinandan Komanduri Y. Mohana Roopa	2019	Having developed the Tesseract OCR Egnine for 10 years helped us a lots in finishing the product.	Lack of storage space. Paper documents can take up a lot of space, and the amount of paper will increase every day.

6	Review on optical character recognition	Muna Ahmed Awel Ali Imam Abidi.	2019	This technology enables computers to identify and process images of text that have been scanned or photographed.	This is because OCR technology is not 100% accurate, and it can sometimes make mistakes when converting images to text.
7	Text extraction using OCR: A Systematic Review.	Rishabh Mittal Anchal Garg	2020	Its ability to automate the process of data entry and text recognition, saving time and reducing the risk of errors.	OCR in RPA enables organizations to automate a greater volume of their operational business processes, especially those that still rely heavily on scanned paperwork such as those completed by customers form.
8	Text Detection Forgot About Document OCR: A systematic review	Krzysztof Olejniczak Milan Šulc	2023	An OCR is the ability to scan the characters accurately	An OCR is limited number of characters offered by it.
9	A Review on OCR Technology	Jay Dilipbhai Thanki Priyank Dineshbhai Davda Dr. Priya Swaminarayan	2021	: The latest software can re-create tables also as original layout.	The quality of the final image depends on the quality of the first image.
10	A Survey on Optical Character Recognition System.	Noman Islam Zeeshan Islam Nazia Noor.	2016	This process is much faster as compared to the manual typing the information into the system	OCR text works efficiently with the printed text only and not with handwritten text. Handwriting should be learned from PC.

III. ARCHITECTURE

1) *Input:* User should give the input as pdf or document .
2) *Preprocessing:* Data preprocessing is the process of converting raw data into a useful, understandable form. Real-world or raw data typically has inconsistent formatting, human errors, and may be incomplete. Data preprocessing solves such problems and makes data sets more complete and efficient for data analysis[1].
3) *Feature Extraction:* Feature extraction refers to the process of converting raw data into numerical features that can be processed while preserving the information in the original data set[2]. This gives better results than applying machine learning directly to the raw data[8].
4) *Segmentation:* Data segmentation is the process of taking the data you have and segmenting it and grouping similar data based on selected parameters so you can use it more effectively in marketing and operations[2]. Example of data segmentation could be: Gender. Consumers vs. Prospects.
5) *Classification (OCR algorithm working):* Optical character recognition (OCR) is the conversion of typed, handwritten, or printed text images into machine-encoded text[1]. With the help of OCR, a large number of paper-based documents, in multiple languages and formats, can be digitized into machine-readable text that not only facilitates storage but also makes previously inaccessible data accessible at the click of a button[2]. Also available Just think of the amount of archive boxes full of paper that lie in city or government basements. Such images and documents can be scanned as documents, document images, or scene images (such as text on signs and billboards).

6) *Extract Text from document or pdf:* With the help of Optical Character Recognition (OCR), you can extract any text from a PDF document into a plain text file[1]. And it's easy: just upload your PDF and let us do the rest. After providing your file, PDF2Go will use OCR to extract the text from your PDF and save it as a TXT file.

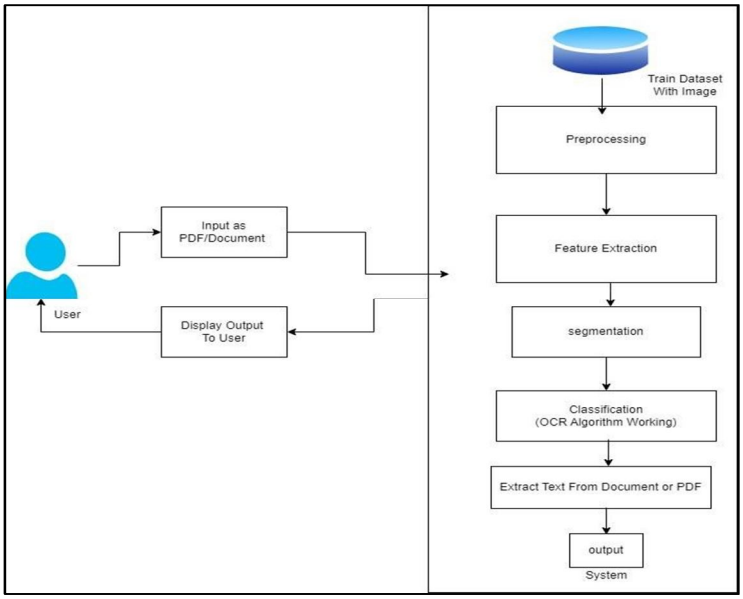


Fig2: System Architecture.

IV. METHODOLOGY

OCR is a technology that analyze the text of a page and turns the letters into code that may be used to process information.OCR is a technique for detecting for printed or handwritten text characters inside digital images of paper files, such as scanning paper files ,such as scanning paper records(optical character recognition).

Working of OCR

A. Image acquisition

A scanner reads documents and converts them into binary data. The OCR software analyze the Scanner images and classifies the light areas as background and the dark areas as text.

B. Preprocessing

OCR software first cleans the image and removes errors to make it readable. Here's some of it to read. These are some of its cleaning techniques:

1) Tilt or tilt the scanned document slightly to correct alignment problems during scanning..
2) Despeckling or removing the edges of text images.
3) Clearing boxes and lines in an image.
4) Script recognition for multi-language OCR technology

C. Text recognition

The two main types of OCR algorithms or software processes that OCR software uses for recognition are called pattern matching and feature extraction.

D. Segmentation

Character Segmentation is the most crucial step for any OCR(Optical Character Recognition)System .The selection of the segmentation algorithm being used is the key factor in deciding the accuracy of OCR system .If there is a good segmentation of characters,the recognition accuracy will also be high.

E. Features Extraction

Features extraction breaks down or decompose the glyphs into features such as lines ,closed loop ,line direction ,and line intersection . It then uses these features to find the best match or the nearest neighbor among its various stored glyphs into features such as lines,closed loops ,line direction ,and line intersection. It then uses these features to find the best match or nearest neighbor among its various stored glyphs.

F. Traning a Neural Network

An optical character recognition (OCR) system, which uses a multilayer perceptron neural network classifier, has the advantage of being fast (highly parallel), easily trainable, and able to create arbitrary partitions of the input feature space is capable of.

G. Postprocessing

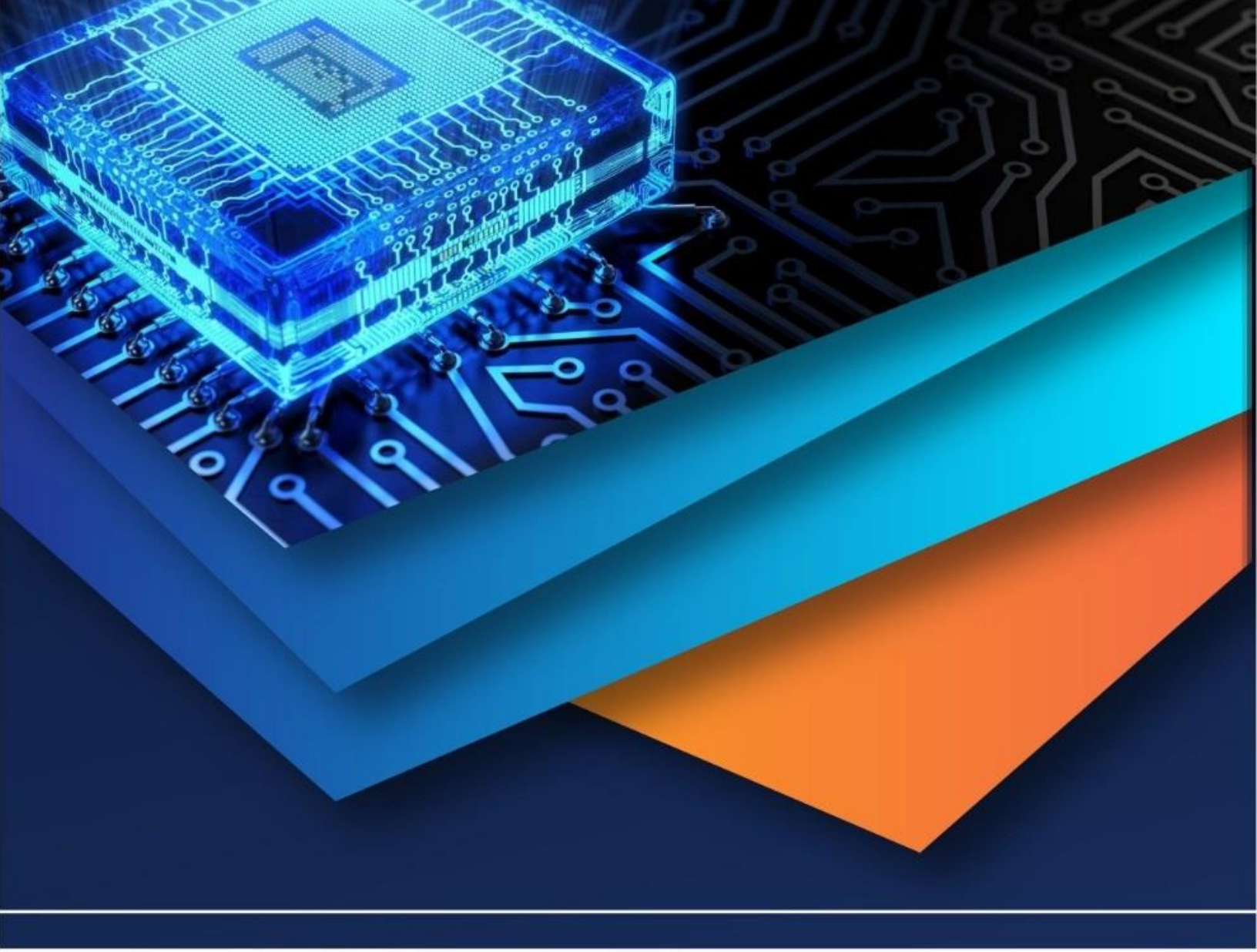
After analysis, the system converts the extracted text data into a computerized file. Some OCR systems can create annotated PDF files that contain earlier and later versions of the scanned document.

V. CONCLUSION

We present FormNet, a novel model architecture for form-based document understanding, this paper is to organize, classify, summarize, and analyze the methods of scene text detection, text recognition and end-to-end text detection and recognition, The most popular handwriting recognition technique is optical character recognition (OCR). It allows us to scan handwritten documents and printed document and then convert them into basic text and in csv using computer vision.

REFERENCES

[1] Milan Aggarwal, Hiresh Gupta, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. Form2seq: A framework for higher-order form structure extraction in EMNLP.
[2] Joshua Ainslie, Santiago Ontan˜on, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured data in transformers in EMNLP.
[3] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Ste- Fan Pletschacher. 2009.
[4] A realistic dataset for performance evaluation of document layout analysis. In ICDAR. Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021.
[5] Docformer: End-to-end transformer for document understanding. In ICCV.Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020.
[6] Unilmv2: Pseudomasked language models for unified language model pre- Training. In ICML.Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013.
[7] Rule-based information extraction is dead! Long live rule-based information Extraction systems! In EMNLP.Brian Davis, Bryan Morse, Scott Cohen, Brian Price, and Chris Tensmeyer. 2019.
[8] FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction. Chen-Yu Lee† , ChunLiang Li† , Timothy Dozat‡ , Vincent Perot‡ , Guolong Su‡ , Nan Hua‡ , Joshua Ainslie‡ , Renshen Wang‡ , Yasuhisa Fujii‡ , Tomas Pfister† ,2022
[9] Review on optical character recognition, Muna Ahmed Awel, Ali Imam Abidi, 2019
[10] Text extraction using OCR: A Systematic Review, Rishabh Mittal, Anchal Garg, 2020



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY