iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: XI Month of publication: November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47414

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 XI Nov 2022- Available at www.ijraset.com

Optical Character Recognition Using Tesseract

Dr. Suraya Mubeen[1], Jally Brahmani[2], Datha Pavan Kalyan[3], Ayesha Jagirdar[4], A. Praveen Kumar[5]
[1, 2, 3, 4, 5]Dept.of ECE, CMR Technical Campus

Abstract: Optical Character Recognition (OCR) is a process or technology in which text within a digital image is recognized. With rapid pace of technology, people want quicker, handy and reliable tools, which can fulfil their daily needs. With this moto we had gone forward and analyzed the existing tools and made up this Android App, which provides seamless experience (No ads and easy-to-use), and great accuracy.

The main objective of this project is to allow automatic extraction of the information that a user wants from the paper document and using it wherever it is needed. In this project, OCR uses Tesseract as an engine to display the text to the user and uses a Deep learning model to classify the letters and display them to the user. It adds a new neural network (LSTM) based OCR engine which is focused on line recognition but also still supports the legacy Tesseract OCR engine which works by recognizing character patterns.

Keywords: OCR, Tesseract, LSTM, Legacy, Android

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".

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 of face recognition and validation.

II. LITERATURE SURVEY

Several experiments have been carried out over the years by different groups of researchers.

Here are some of the following groups:

1) Yue Jiet Chong, Kein Huat Chua, Mohammad Babrdel, Lee Cheun Hau, Li Wang has proposed a deep learning model based on Single Shot Detector (SSD) Mobile-net V2 and an optical character recognition (Tesseract OCR) engine are developed for the low-cost digitization of analogue meter readings. The model is developed in Python and the evaluation is carried for various types of meters, illumination conditions, and backgrounds. The results show that the deep learning model and OCR accuracies are 95% and 93%, respectively.

2) Vernon Estrada Bugayong, Jocelyn Flores Villaverde, Noel B. Linsanga has proposed an optical character recognition system capable of interpreting captured images of hard disk drive and solid-state drive labels with high accuracy. The images captured using a vision camera went through different stages of image pre-processing via OpenCV-Python and recognition through Google Tesseract.

3) Pranamya P Bhat , Pratheeksha P , Pankhuri Tayal has come up with the idea of Digitization of mechanical meter reading using OCR. A camera fixed in front of the mechanical meter takes the image of the meter every month and sends it over to the database. Using the image, the current reading is extracted (by OCR methods). Current reading is subtracted from the previous months reading and stored back in the database. For the customer to view the generated bill, they will be required to log in into the web application.

III. PROPOSED SYSTEM

Our proposed system is OCR using tesseract which is a character recognition app (OCR) that supports recognition of the characters of multiple languages (120+ languages). It extract text from images and copy data to clipboard. Process multiple languages in single image. Process images directly from the gallery on your device via the share menu. Recognizes Maths equations with great accuracy.

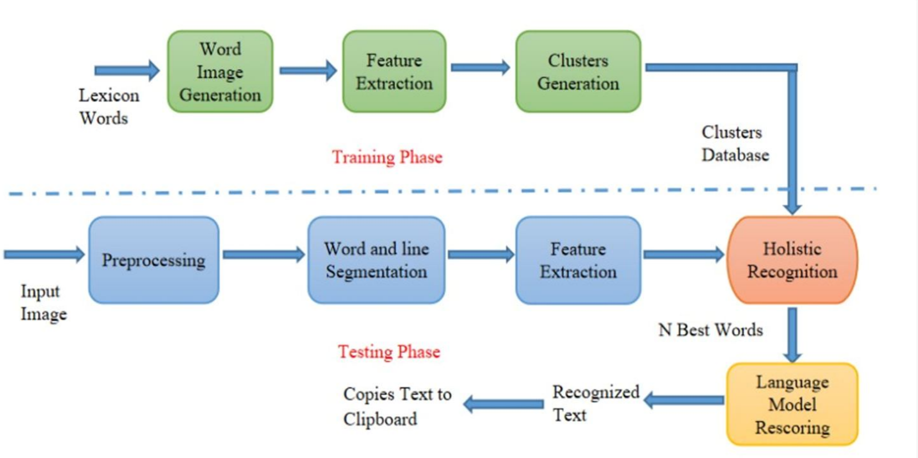
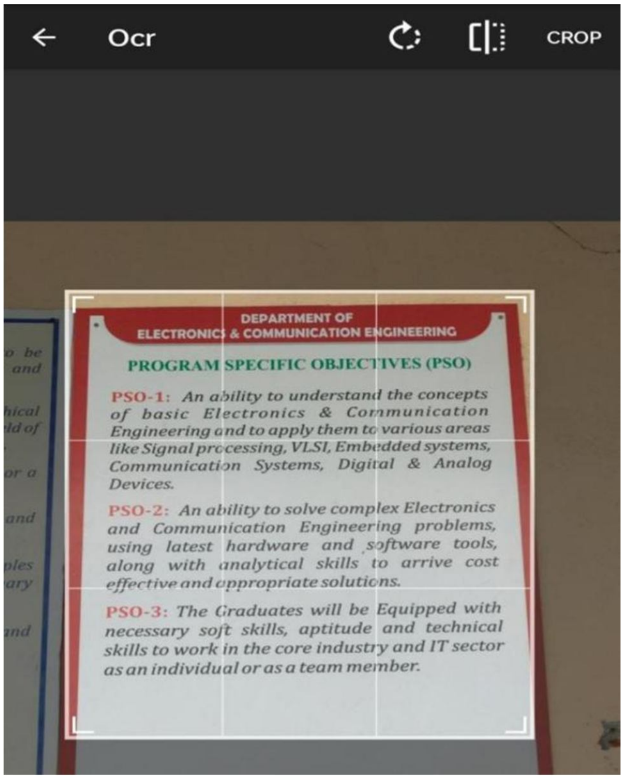


Figure 3.1: Architecture of the Model.

IV. RESULT

It's unrealistic to expect any OCR system, even state-of-the-art OCR engines, to be 100% accurate. Obviously, the accuracy of the conversion is important, and the proposed OCR software provides 85 to 98 percent accuracy. In most cases, this level of accuracy is acceptable.

The results of the proposed model is shown below -



DEPARTMENT OF
ELECTRONICS & COMMUNICATION ENGINEERING

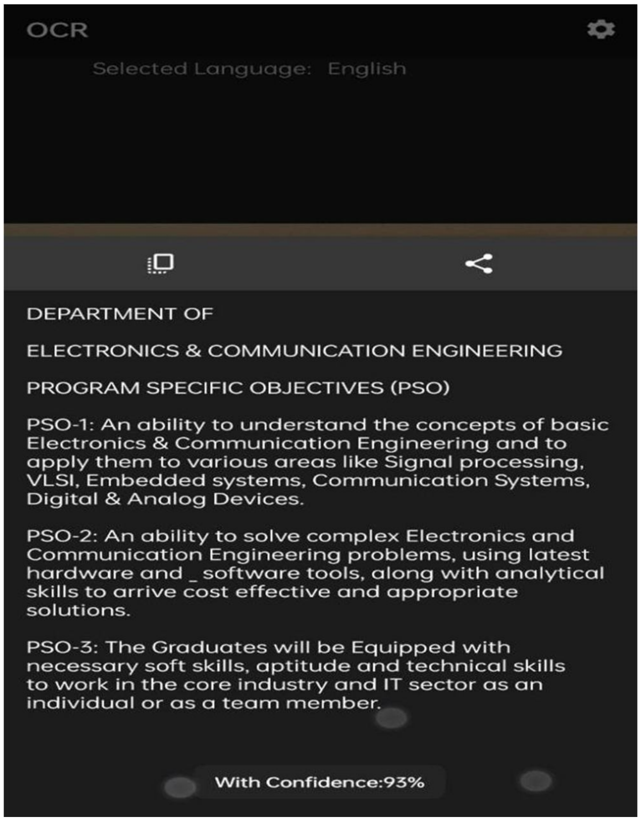
PROGRAM SPECIFIC OBJECTIVES (PSO)

PSO-1: An ability to understand the concepts of basic Electronics & Communication Engineering and to apply them to various areas like Signal processing, VLSI, Embedded systems, Communication Systems, Digital & Analog Devices.

PSO-2: An ability to solve complex Electronics and Communication Engineering problems, using latest hardware and software tools, along with analytical skills to arrive cost effective and appropriate solutions.

PSO-3: The Graduates will be Equipped with necessary soft skills, aptitude and technical skills to work in the core industry and IT sector as an individual or as a team member.

Fig 4.1(a): Input Image



DEPARTMENT OF

ELECTRONICS & COMMUNICATION ENGINEERING

PROGRAM SPECIFIC OBJECTIVES (PSO)

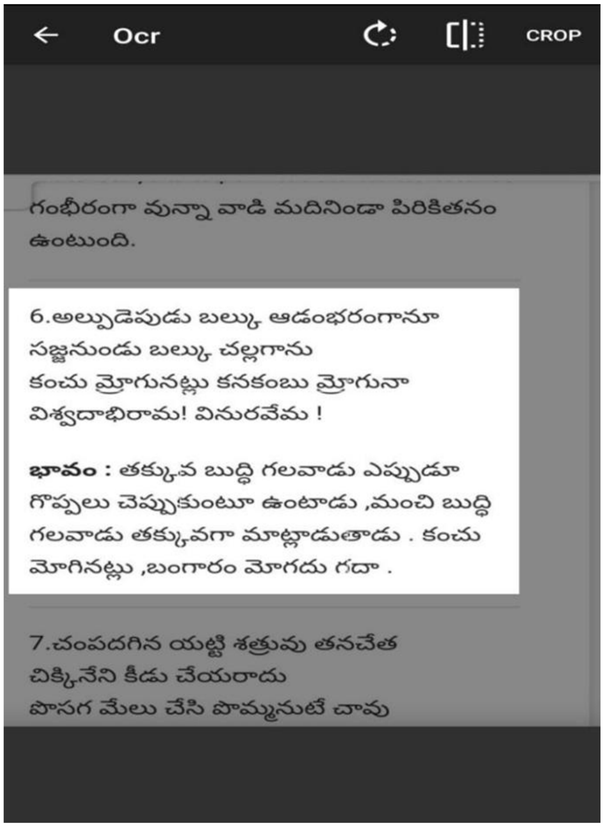
PSO-1: An ability to understand the concepts of basic Electronics & Communication Engineering and to apply them to various areas like Signal processing, VLSI, Embedded systems, Communication Systems, Digital & Analog Devices.

PSO-2: An ability to solve complex Electronics and Communication Engineering problems, using latest hardware and _ software tools, along with analytical skills to arrive cost effective and appropriate solutions.

PSO-3: The Graduates will be Equipped with necessary soft skills, aptitude and technical skills to work in the core industry and IT sector as an individual or as a team member.

With Confidence:93%

Fig 4.1(b): Text extraction from the input image



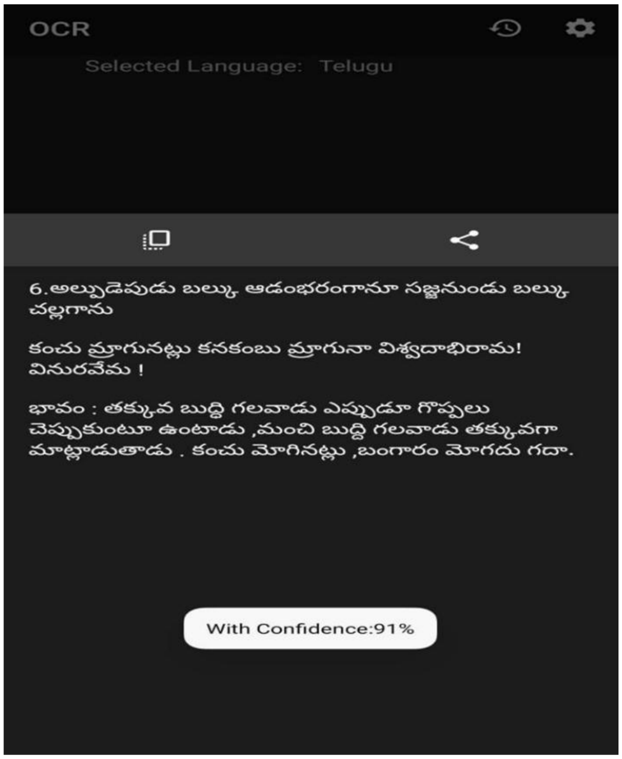
గంభీరంగా వున్నా వాడి మదినిండా పిరికితనం ఉంటుంది.

6.అల్పుడెపుడు బల్కు ఆడంబరంగానూ
సజ్జనుండు బల్కు చల్లగాను
కంచు మ్రోగునట్లు కనకంబు మ్రోగునా
విశ్వదాభిరామ! వినురవేమ !

భావం : తక్కువ బుద్ధి గలవాడు ఎప్పుడూ గొప్పలు చెప్పుకుంటూ ఉంటాడు ,మంచి బుద్ధి గలవాడు తక్కువగా మాట్లాడుతాడు . కంచు మోగినట్లు ,బంగారం మోగదు గదా .

7.చంపదగిన యట్టి శత్రువు తనచేత
చిక్కినేని కీడు చేయరాదు
పొసగ మేలు చేసి పొమ్మనుటే చావు

Fig 4.2(a): Input Image



6.అల్పుడెపుడు బల్కు ఆడంభరంగానూ సజ్జనుండు బల్కు చల్లగాను

కంచు మ్రోగునట్లు కనకంబు మ్రోగునా విశ్వదాభిరామ! వినురవేమ !

భావం : తక్కువ బుద్ధి గలవాడు ఎప్పుడూ గొప్పలు చెప్పుకుంటూ ఉంటాడు ,మంచి బుద్ధి గలవాడు తక్కువగా మాట్లాడుతాడు . కంచు మోగినట్లు ,బంగారం మోగదు గదా.

With Confidence:91%

Fig 4.1(b): Text extraction from telugu image

V. CONCLUSION

To sum up, developed model is able to detect & extract text from images. However, accuracy is maximum for tesseract engine moreover output is saved in text file and copies to clipboard. Tesseract performs well when document images follow the next guidelines: clean segmentation of the foreground text from background, horizontally aligned and scaled appropriately high-quality image without blurriness and noise. The model's key drawback is the blur or light images. In order to improve its accuracy, our future approach will be based on gathering more enhanced techniques from various sources and improving the proposed image processing algorithm.

REFERENCES

[1] "Yue Jiet Chong, Kein Huat Chua, Mohammad Babrdel, Lee Cheun Hau, Li Wang", "Deep Learning and Optical Character Recognition for Digitization of Meter Reading" , ISBN : 978-1-6654-8703-0, 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), May 2022.

[2] "Vernon Estrada Bugayong, Jocelyn Flores Villaverde, Noel B. Linsangan", "Google Tesseract: Optical Character Recognition (OCR) on HDD / SSD Labels Using Machine Vision", ISBN:978-1-6654-8380-3, 2022 14th International Conference on Computer and Automation Engineering (ICCAE), March 2022.

[3] "Pranamya P Bhat , Pratheeksha P , Pankhuri Tayal", "Digitization of Mechanical Meter Bill Generation using OCR", International Journal of Engineering Research & Technology (IJERT), VOL NO : 11, ISSUE NO : 01, ISSN : 2278-0181, January 2022.

[4] "Nikita Kotwal , Gauri Unnithan , Ashlesh Sheth , Nehal Kadaganchi", "Optical Character Recognition using Tesseract Engine", International Journal of Engineering Research & Technology (IJERT), VOL NO : 10, ISSUE NO : 09, ISSN : 2278-0181, September 2021.

[5] "Firhan Maulana Rusli, Kevin AkbaAdhiguna, Hendy Irawan", "Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing", 2021 9th International Conference on Information [x/[and Communication Technology (ICoICT), ISBN:978-1-6654-0447-1, August 2021.

[6] "Tao Ma, Min Yue, Chao Yuan, Haibo Yuan", "File Text Recognition and Management System Based on Tesseract-OCR", 2021 3rd International Conference on Applied Machine Learning (ICAML), ISBN:978-1-6654-2125-6, July 2021.

[7] "Saurabh Dome, Asha P Sathe", "Optical Charater Recognition using Tesseract and Classification", 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), ISBN:978-1-7281-8519-4, March 2021.

[8] "AS Revathi, Nishi A Modi", "Comparative Analysis of Text Extraction from Color Images using Tesseract and OpenCV", 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), ISBN : 978-93-80544-43-4, March 2021.

[9] "Isuri Anuradha, Chamila Liyanage, Harsha Wijayawardhana, Ruvan Weerasinghe", "Deep Learning Based Sinhala Optical Character Recognition (OCR)", 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), ISSN : 2472-7598, November 2020.

[10] "Saumitra Godbole, Dhananjay Joijode, Kshitij Kadam, Sameer Karoshi", "Detection of Medicine Information with Optical Character Recognition using Android", 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), ISBN : 978-1-7281-8794-5, October 2020.