IJARSCT

Impact Factor: 7.301

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Volume 3, Issue 14, May 2023

Research Paper on Text Extraction using OCR

Prof. Anuradha Thorat[1], Mayur Zagade[2], Shivani More[3], Manish Pasalkar[4], Anand Narute[5]

Assistant Professor, Department of Information Technology[1]

Student, Department of Information Technology[2,3,4,5]

Zeal College of Engineering and Research, Pune, India

Abstract: *Text extraction from images is a challenging task with numerous applications in fields such as document digitization, information retrieval, and image understanding. Extracting text accurately and efficiently from images is crucial for enabling automated processes and facilitating the analysis of large volumes of visual data. This abstract provides an overview of text extraction from images, focusing on recent advancements and techniques employed in this area. Traditional methods for text extraction from images involved preprocessing steps, such as binarization and noise removal, followed by techniques like connected component analysis and optical character recognition (OCR). These approaches often faced challenges in handling complex backgrounds, varying fonts, and distorted or degraded text.*

Keywords*:* Digitization, Binarization, Extracting.

I. INTRODUCTION

Text extraction from images using Optical Character Recognition (OCR) is a prominent field in computer vision and document processing. OCR technology enables the conversion of text contained in images or scanned documents into editable and searchable digital formats. This process plays a vital role in various applications, including document digitization, information retrieval, automated data entry, and text analysisOCR systems aim to accurately recognize and extract text from images, overcoming challenges such as varying fonts, styles, sizes, and image quality. The process involves several stages, including image preprocessing, textdetection, character segmentation, and text recognition.Image preprocessing techniques are employed toenhancethe quality of input images and improve the accuracy of subsequent OCR tasks. Common preprocessing steps include noise reduction, image binarization, skew correction, and geometric normalization. These techniques help mitigate noise, correct image distortions, and ensure that text regions are properly aligned.Text detection is a crucial step in OCR, where algorithms analyze the image to identify regions likely to contain text. Traditional text detection methods relied on handcrafted features, such as edge detection, morphological operations, and connected component analysis. However, with the advent of deep learning, convolutional neural networks (CNNs) have proven to be effective in automatically detecting text regions within images.

II LITERATURE REVIEW

[1]**Name**: A Novel Method based on Character Segmentation for Slant Chinese Screen-render Text Detection and Recognition

Author: Tianlun Zheng 1,2, Xiaofeng Wang1,2,*, Xin Yuan 1,2, and Shiqin Wang

Abstract: Screen rendering text has broad application prospects in the fields of medical records, dictionary screen capture, and screen-assisted reading. However, Chinese screen rendering text always has the challenges of small font size and low resolution. Obtaining a screen-rendered text image in a natural scene will have a certain tilt angle. These all pose great challenges for screen text recognition. This paper proposes a method based on character segmentation

[2] **Name:**Research on Text Detection and Recognition Based on OCR Recognition Technology

Author: Yuming He

Abstract: Image recognition and optical character recognition technologies have become an integral part of our everyday life due in part to the ever-increasing power of computing and the ubiquity of scanning devices. Printed documents can be quickly converted into digital text files through optical character recognition and then be edited by the user. Consequently, minimal time is required to digitize documents; this is particularly helpful when archiving volumes of printed materials

[3]**Name:**Summary of Scene Text Detection and Recognition
Author: Yao Qin1, 2, 3, Zhi Zhang1
Abstract: In recent years, scene text recognition has received much attention, and has a wealth of application scenarios, such as: photo translation, image retrieval, scene understanding and so on. However, the text in the scene is also faced with many problems, such as: light changes, deformation text, text string recognition under background noise interference, text skew and degree of curvature, and a large number of artistic fonts. Solving the above problems will always be a challenging thing

III. METHODOLOGY

Image Preprocessing

Convert the input image to grayscale to simplify subsequent processing steps.Apply noise reduction techniques, such as Gaussian blurring or median filtering, to eliminate unwanted noise or artifacts.Perform image binarization to separate the text regions from the background. This can be achieved through thresholding or adaptive thresholding algorithms.

Text Detection

Utilize a text detection algorithm, such as a convolutional neural network (CNN), to identify potential text regions within the preprocessed imageTrain the text detection model using annotated datasets that include images with labeled text regions.Employ techniques like sliding windows, region proposal networks (RPNs), or connected component analysis to generate bounding boxes around the detected text regions.

Character Segmentation:

If the text regions contain connected or overlapping characters, apply character segmentation techniques to separate individual characters or groups of characters.Employ methods such as contour analysis, connected component labeling, or clustering algorithms to identify and separate the characters within the text regions.Validate the segmentation results through geometric constraints, stroke width analysis, or other heuristics to ensure accurate separation of characters.

Text Recognition:

Employ an OCR model, such as a recurrent neural network (RNN), convolutional-recurrent neural network (CRNN), or transformer-based model, to recognize the text within the segmented characters or character groups.Train the OCR model using annotated datasets containing images paired with corresponding ground truth text.Preprocess the segmented characters, such as resizing, normalization, or padding, to match the input requirements of the OCR modelUtilize the trained OCR model to predict the textual content of each character or character group.Employ post-processing techniques, such as language models, dictionary lookups, or confidence score thresholds, to refine and improve the accuracy of the recognized text.

Text Extraction and Output:

Combine the recognized characters or character groups to form complete words or sentences.Preserve the spatial layout and structure of the text by considering the bounding box information obtained during text detection.Output the extracted text in a digital format, such as plain text or a searchable document format, to facilitate further processing or analysis

IV. SYSTEM ARCHITECTURE

[1] Requirement Analysis - Requirement Analysis is the most important and necessary stage in SDLC. The senior members of the team perform it with inputs from all the stakeholders and domain experts or SMEs in the industry. Planning for the quality assurance requirements and identifications of the risks associated with the projects is also done at this stage. Business analyst and Project organizer set up a meeting with the client to gather all the data like what the customer wants to build, who will be the end user, what is the objective of the product. Before creating a product, a core understanding or knowledge of the product is very necessary.

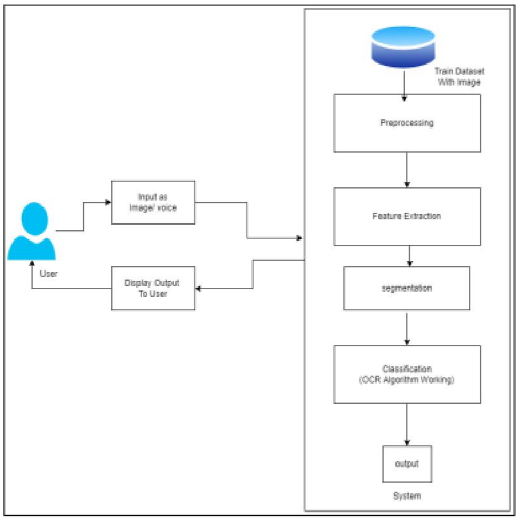


Fig: - System Architecture

[2] System Design - The next phase is about to bring down all the knowledge of requirements, analysis, and design of the software project. This phase is the product of the last two, like inputs from the customer and requirement gathering. 3. Implementation - In this phase of SDLC, the actual development begins, and the programming is built. The implementation of design begins concerning writing code. Developers have to follow the coding guidelines described by their management and programming tools like compilers, interpreters, debuggers, etc. are used to develop and implement the code.

[4] Testing - After the code is generated, it is tested against the requirements to make sure that the products are solving the needs addressed and gathered during the requirements stage. During this stage, unit testing, integration testing, system testing, acceptance testing are done.

[5] Deployment - Once the software is certified, and no bugs or errors are staSted, then it is deployed. Then based on the assessment, the software may be released as it is or with suggested enhancement in the object segment. After the software is deployed, then its maintenance begins.

[6] Maintenance - Once when the client starts using the developed systems, then the real issues come up and requirements to be solved from time to time. This procedure where the care is taken for the developed product is known as maintenance

IV. IMPLEMENTATION PLAN

Model Training : The training and testing is done using pre-processed dataset. Image dataset used . The Text Recognition of Image can be done Using OCR.

Model Testing and Evaluation:

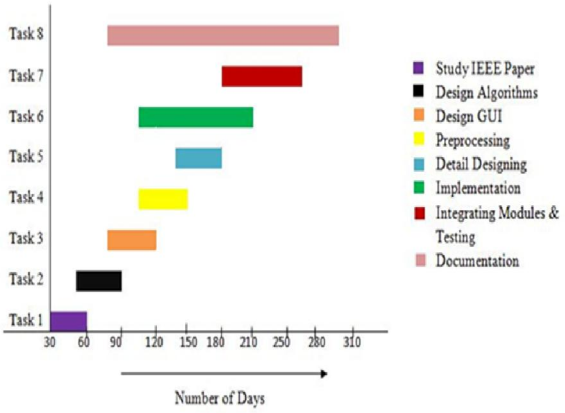
Evaluate the trained model's performance on a separate test dataset to measure its accuracy, precision, recall, and other relevant metrics.

- Front End: Python , Tkinter
- Backend: MySqLite
- Framework: Spyder
- Algorithms: OCR Algorithm

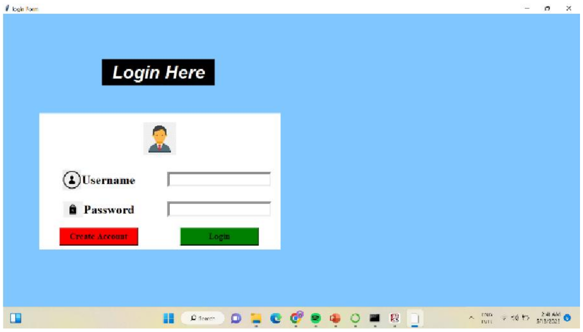
DOI: 10.48175/568

ISSN
2581-9429
IJARSCT

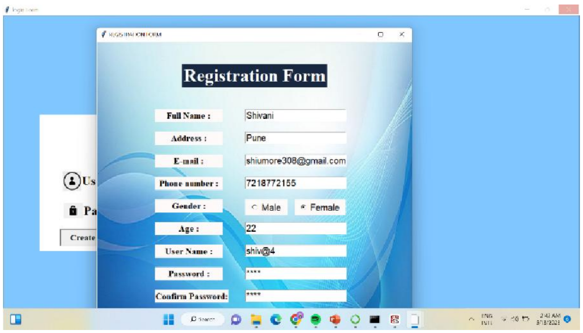
GRAPH :



V. RESULT



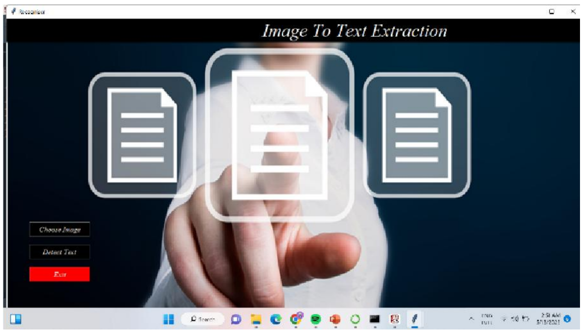
5.1: Login Page



Registration Form

Full Name :	Shivani
Address :	Pune
E-mail :	shiumore308@gmail.com
Phone number :	7218772155
Gender :	○ Male ● Female
Age :	22
User Name :	shiv@4
Password :	****
Confirm Password:	****

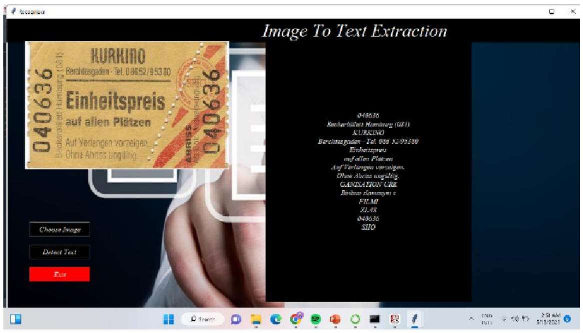
5.2: Registration Page



5.3 User Dashboard

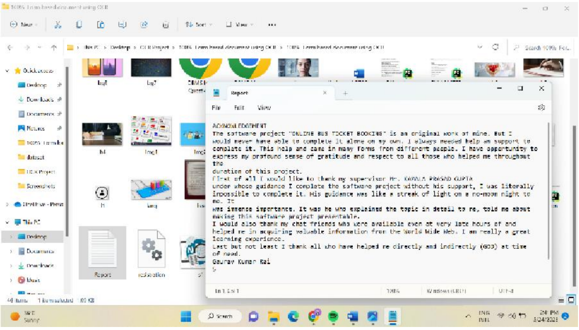
DOI: 10.48175/568

ISSN
2581-9429
IJARSCT
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY



```
040636
Beckerbillett Hamburg (081)
KURKINO
Berchtesgaden · Tel. 086 52/95380
Einheitspreis
auf allen Plätzen
Auf Verlangen vorzeigen.
Ohne Abriss ungültig.
GANISATION URE
Binhan dansmpm z
FILMP
ZLAS
040636
SIIO
```

5.4: Text Extraction Successful



```
ACKNOWLEDGEMENT
The software project "ONLINE BUS TICKET BOOKING" is an original work of mine. But I
would never have able to complete it alone on my own. I always needed help an support to
complete it. This help and came in many forms from different people. I have opportunity to
express my profound sense of gratitude and respect to all those who helped me throughout
the
duration of this project.
First of all I would like to thank my supervisor Mr. KAMALA PRASAD GUPTA
under whose guidance I complete the software project without his support, I was literally
impossible to complete it. His guidance was like a streak of light on a no-moon night to
me. It
was immense importance. It was he who explained the topic in detail to me, told me about
making this software project presentable.
I would also thank my chat friends who were available even at very late hours of and
helped me in acquiring valuable information from the World Wide Web. I am really a great
learning experience.
Last but not least I thank all who have helped me directly and indirectly (GOD) at time
of need.
Gaurav Kumar Rai
5
```

5.5: Output Generated As Text File

VI. CONCLUSION

Need several kinds of images as sources of information for elucidation and analysis. When an image is transformed from one form to another such as digitizing, scanning, and communicating, storing, etc. degradation occurs. Therefore, the output image has to undertake a process called image enhancement, which contains of a group of methods that seek to develop the visual presence of an image. Image enhancement is fundamentally enlightening the interpretability or awareness of information in images for human listeners and providing better input for other automatic image processing systems. New features can be added to improve the accuracy of recognition. These algorithms can be tried on large database of handwritten text. There is a need to develop the standard database for recognition of text. The proposed work can be extended to work on degraded text or broken characters. Recognition of digits in the text, half characters and compound characters can be done to improve the word recognition rate. This extracted text can be further converted to audio so make physically challenged i.e. blind people easily understand which text has been converted from the image.

REFERENCES

[1] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5676–5685.

[2] S. Unar, A. H. Jalbani, M. M. Jawaid, M. Shaikh, and A. A. Chandio, "Artificial urdu text detection and localization from individual video frames," Mehran Univ. Res. J. Eng. Technol., vol. 37, no. 2, pp. 429–438, 2018.

[3] A. Mirza, M. Fayyaz, Z. Seher, and I. Siddiqi, "Urdu caption text detection using textural features," in Proc. 2nd Medit. Conf. Pattern Recognit. Artif. Intell., 2018, pp. 70–75.

[4] C. Yao. MSRA Text Detection 500 Database (MSRA-TD500). Accessed: Aug. 2018 [Online].

[5] A. A. Chandio and M. Pickering, "Convolutional feature fusion for multilanguage text detection in natural scene images," in Proc. 2nd Int. Conf. Comput., Math. Eng. Technol. (iCoMET), Jan. 2019, pp. 1–6.

[6] A. A. Chandio and M. Pickering, "Convolutional feature fusion for multilanguage text detection in natural scene images," in Proc. 2nd Int. Conf. Comput., Math. Eng. Technol. (iCoMET), Jan. 2019, pp.1–6.