IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105118 |

Restoration of Partially Visible Text in A Scanned Document

Priyesh jain[1], Pratheeksha Shetty [2], Abhishek P shetty[3], Mr. Rajesh N Kamath [4]

Student, Dept. of Information Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India[1,2,3]

Assistant Professor, Dept. of Information Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India[5]

ABSTRACT: Abstract— Machine learning has emerged as a new area of artificial intelligence study in recent years. Complex systems like those that use machine learning, such as those that use speech recognition, natural language processing, online search, recommendation systems, intelligent robotics, artificial intelligence, etc., have been used successfully. Since machines were created by people, their intellect will never surpass that of people. A challenging issue in the realm of character identification has been the recognition of Hindi characters. It is challenging to utilize our conventional approach to automatically recognize it in languages like English where number characters are small. There is still much potential for development even if several local and international software providers have introduced a rate of characters automatic identification system with good recognition. OCR and various machine learning approaches now make it possible to forecast and restore partially visible letters on documents and photos. The majority of papers in many recent domestic works of literature focus on study into the automated recognition of a limited set of characters. It is challenging to apply to a large character set recognition object; hence, software is mostly developed for unnoticed text recognition

KEYWORDS: TESSERACT, OCR, AND BINARIZATION

1. INTRODUCTION

Every image that is scanned has some degree of partial visibility. The causes could include an increase in ink, an excessive amount of darkness, ink that is smeared across the paper, or letters that are only half printed. The goal is to use image processing to recover the actual letter based on the sentence or the letter's significance. Additionally, letters may be lost or deleted in earlier printed papers. This will diminish the value of keeping the document.

In this paper, on the scanned page, printed document, and image, we have compiled several methods for predicting and recovering the partially visible letters.

II. ANALYSIS OF EXISTING APPROACHES

There are numerous OCR-based programmers that can extract text from images, but they are insufficiently accurate to perform the same purpose for handwritten writing. Additionally, typing every phrase would be a laborious and time-consuming operation. It is simple to manage them all if we consider the advantages of turning handwritten documents into online computerized ones. Additionally, updated information can be added, something that cannot be done on a straightforward text document that has been simply scanned.destination and the calculated lifetime of each path. Lifetime of the path is used as a fitness function. Fitness function will select the highest chromosomes which is having highest lifetime. Cross over and mutation operators are used to enhance the selection. In [5] authors improved AODV protocol by implementing a balanced energy consumption idea into route discovery process. RREQ message will be forwarded when the nodes have sufficient amount of energy to transmit the message otherwise message will be dropped. This condition will be checked with threshold value which is dynamically changing. It allows a node with over used battery to refuse to route the traffic in order to prolong the network life. In [6] Authors had modified the route table of AODV adding power factor field. Only active nodes can take part in rout selection and remaining nodes can be idle. The lifetime of a node is calculated and transmitted along with Hello packets. In [7] authors considered the individual battery power of the node and number of hops, as the large number of hops will help in reducing the range of the transmission power. Route discovery has been done in the same way as being done in on-demand routing

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

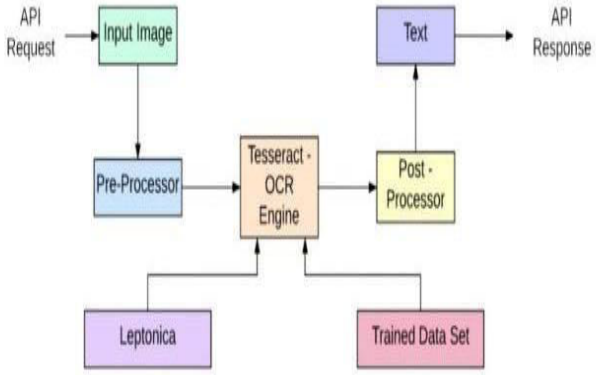
|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105118 |

algorithms. After packet has been reached to the destination, destination will wait for time δt and collects all the packets. After time δt it calls the optimization function to select the path and send RREP. Optimization function uses the individual node's battery energy; if node is having low energy level then optimization function will not use that node.

2.1 System Architecture

Proposed Model, Architecture, and Algorithm Image Acquisition, Pre-processing, Text Recognition, Pattern Matching, Feature Extraction, and Post-processing



2.2 OCR

Pre-processing is one of the OCR techniques.

The image or document is initially pre-processed by OCR to enhance data recognition. The following are some approaches for this pre-processing:

Grey scale is used to eliminate the positive and negative flecks in DE speckle. De-skew - If the document is not completely horizontal while scanning, we must make it that way in order to improve recognition.

Grey scale is utilised for binarization in this instance. Text separation is accomplished by binarization. For this process, numerous commercial algorithms are employed. Character recognition quality and thoughtful selections are necessary because the type of input image—such as a scanned document, scene text image, historical deteriorated document, etc.—determines the quality of the binarization method used to produce the binary result. Box removal and line cleaning are done effectively here. It deals with the data's organisational elements, like as columns, rows, and so forth.

Line and word detection: This feature looks for certain word or character patterns in scanned text.

Script recognition - In this step, the language type is detected and identified so that the appropriate OCR may be used precisely. Segmentation For the sake of recognition, a single character that has been split into many pieces by artefacts needs to be connected.

2.3 post-processing

Lexical analysis boosts OCR recognition's precision. It is challenging for the software to distinguish words in a document that are not included in the vocabulary. Software like Tesseract maintains the lexicon for quick recognition, increasing the software's accuracy. The OCR software's output can be in text format or in the document's format. The output and software are key factors. It might resemble a well-formatted PDF file or a plain text document. Additionally, it includes the photos with the correct alignments and effects. K closest neighbor machine learning methods are employed for nouns and some popular words that are grammatically incorrect. Grammar analysis also aids in identifying unknown terms so that those can be assumed, increasing accuracy.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105118 |

2.4 Pre-processing

Documents were submitted using the OCR engines' default settings without any further pre-processing because the experiment was designed to gauge performance right out of the box. Footnote7 Although this is a rare application of Tesseract, it treats the engines equally and highlights how heavily picture pre-processing is reliant on Tesseract. A total of 42,504 requests for document processing using all three OCR engines were made using the English corpus. Since Tesseract does not support Arabic, the Arabic corpus was only submitted to Tesseract and Document AI for a total of 8800 processing requests.

iii.literature survey

Taking into account the numerous project characteristics and the scope of the project, a literature survey displays the many analyses and research done in the topic of interest as well as the results that have already been published. It contains research done by numerous analysts, along with their methods and findings. It is the most crucial section of the report since it specifies the path the research will take.

3.1 BACKGROUND RESEARCH

The majority of information today is printed on paper. Documents that need to be digitized and utilized for archiving, indexing, or information retrieval are becoming more and more prevalent today, such as scanned copies of office documents seen in publications, advertising, and web sites. Due to the many characteristics of text, robust and effective text extraction from these documents is a difficult challenge to solve. The fundamental Text extraction method was created to transform the Data from paper-based documents that can be

processed by computers to create editable, reusable documents. Software for text recognition is an automatica tool that extracts text from image files so you don't have to manually enter it in again. When working with scanned documents and image files that contain text that the computer might not be able to recognize, this is very crucial. This issue is fixed by making the documents text-searchable using text recognition software, also known as OCR.

[1] Text Extraction and Detection from Images using Machine Learning Techniques By Shivani Surana, Komal Pathak, Mehul Gagnani, Vidhan Shrivastava, Mahesh T R, Sindhu Madhuri A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022

[2] A New Defect Detection Method for Improving Text Detection and Recognition Performances in Natural Scene Images By H.Mokayed, P. Shivakumara, M. Liwicki and U. Pal, 2020 Swedish Workshop on Data Science (SweDS), 2020

[3] An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents by Gabriel Pereira e Silva and Rafael DueireLins, 2011 International Conference on Document Analysis and Recognition, 2011

[4] Machine learning with text recognition by Subodh L. Wasankar , Harshad Mahajan, Deovrat Deshmukh, Hemant Munot , 2010 IEEE International Conference on Computational Intelligence and Computing Research, 2010

[5] Support Vector Machine (SVM) for English Handwritten Character Recognition by Dewi Nasien, HabibollahHaron, Siti SophiayatiYuhaniz , 2010 Second International Conference on Computer Engineering and Applications, 2010,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105118 |

IV. RESULTS AND DISCUSSION

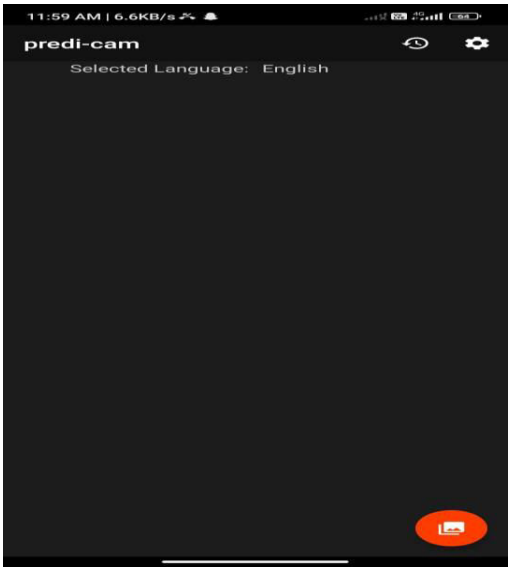


Figure 4.1 Home Screen

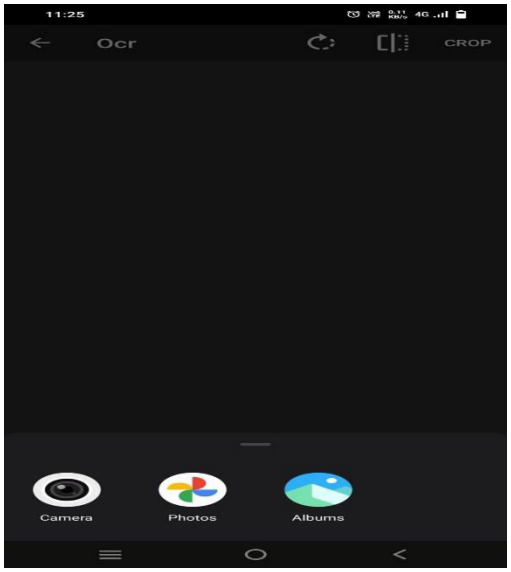


Figure 4.2 Image Selection Method

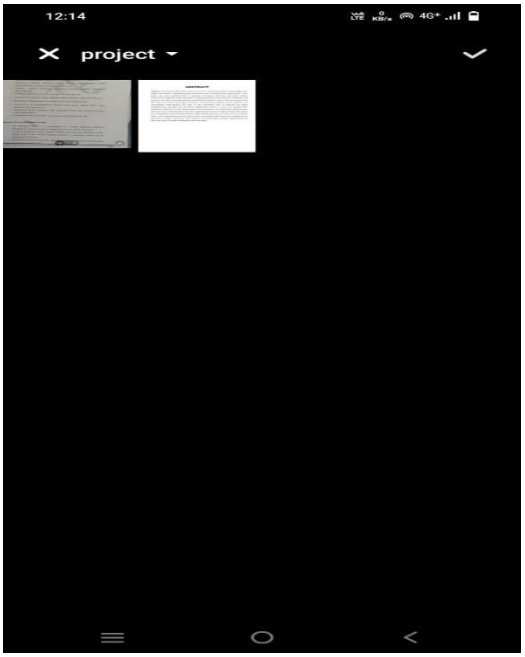


Figure 4.4 Select Image from Gallery

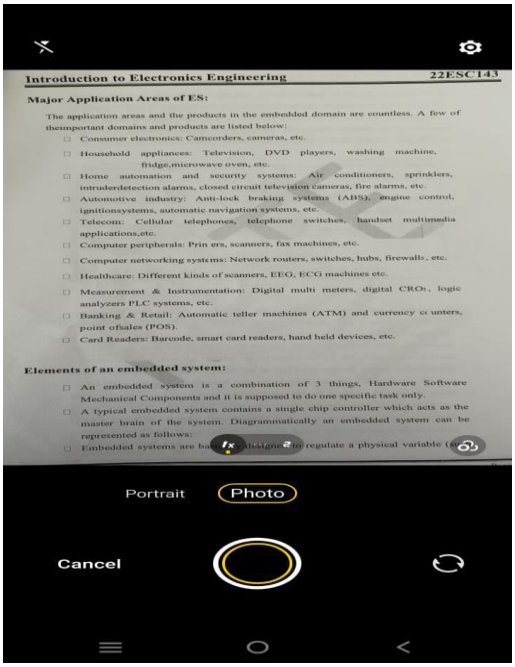


Figure 4.3 Input Using Camera

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105118 |

Figure 4.5 cropper

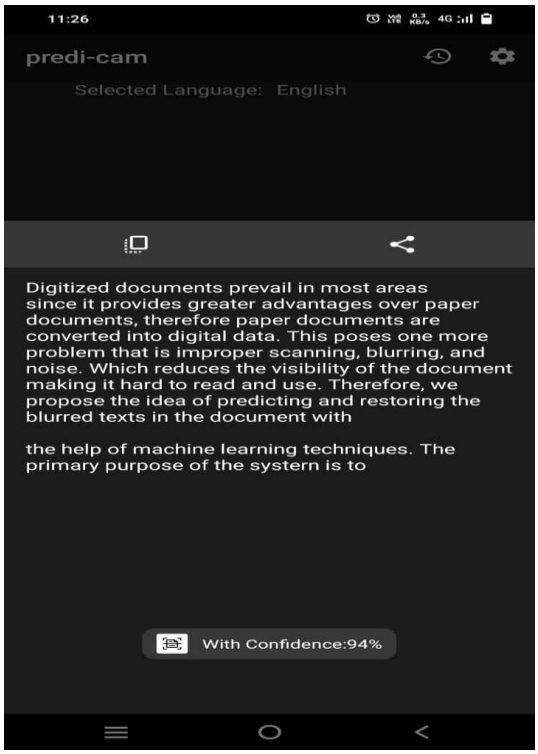


Figure 4.6 Output

V. CONCLUSION

This discipline has advanced due to a demand for human-computer interaction. The researchers in this area are working very hard to create character recognition systems that can actually be used to benefit society. The goal of the approaches like OCR and Tesseract presented in this paper is to increase recognition rates. According to the survey, deployed identification systems successfully identify high-quality characters, but they fall short when faced with overlapped, similarly shaped, or ambiguous characters. The goal of this session is to clarify how damaged documents and images might have their partially viewable data collected and assessed. This study discovered that software that uses algorithms, such as OCR, tesseract, etc., is typically used effectively for the intended purpose.

REFERENCES

[1] S. Surana, K. Pathak, M. Gagnani, V. Shrivastava, M. T. R and S. Madhuri G, "Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022

[2] H. Mokayed, P. Shivakumara, M. Liwicki and U. Pal, "A New Defect Detection Method for Improving Text Detection and Recognition Performances in Natural Scene Images," 2020 Swedish Workshop on Data Science (SweDS), 2020.

[3] Z. Dai and J. Lücke, "Autonomous Document Cleaning—A Generative Approach to Reconstruct Strongly Corrupted Scanned Texts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp. 1950-1962, 1 Oct. 2014, doi: 10.1109/TPAMI.2014.2313126.

[4] S. L. Wasankar, H. Mahajan, D. Deshmukh and H. Munot, "Machine learning with text recognition," 2010 IEEE International Conference on Computational Intelligence and Computing Research, 2010.

[5] D. Nasien, H. Haron and S. S. Yuhaniz, "Support Vector Machine (SVM) for English Handwritten Character Recognition," 2010 Second International Conference on Computer Engineering and Applications, 2010.

[6] G. P. e. Silva and R. D. ins, "An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents," 2011 International Conference on Document Analysis and Recognition, 2011, pp. 553-557, doi: 10.1109/ICDAR.2011.117.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details