iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60794

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 12 Issue IV Apr 2024- Available at www.ijraset.com

OCR: Optical Character Recognition

Kishan kushavaha[1], Anil kumar[2], Anshuman chakrawati[3]
Department of Computer Science and Engineering Galgotias College of Engineering and Technology (AKTU) Uttar Pradesh, India

Abstract:In many different fields, there is a great need to save the information contained in printed or written documents or images to the computer for later use. An easy way to save the information in these printed documents to your computer is to scan the document and save it as an image file. But it would be very difficult to read or query the text or other information in this image file to reuse this data. Therefore, there needs to be a technology that automatically collects and stores information (especially text) from image files. Optical character recognition is a field of research that seeks to create computer systems that can extract and process text from images. The purpose of OCR is to alter or convert any text or document containing text (such as handwritten, printed, or scanned images) into a digital format that can be converted to be processed for greater depth and processing. OCR thus enables machines to recognize text in these documents. There are some important issues that need to be recognized and addressed in order for automation to be implemented successfully. Characterization of the alphabet and the quality of characters in the document are just some of the recent challenges. Due to these problems, the computer may not recognize characters correctly. In this article, we examine OCR in four different ways. First, we provide detailed information about the difficulties that may arise during OCR. Secondly, we examine the general stages of OCR systems such as preprocessing, segmentation, normalization, feature extraction, classification and post-
processing. We will then focus on the development of OCR, its main applications and uses, and finally discuss the brief history of OCR. This session therefore provides an excellent review of the state-of-the-art in this field..

I. INTRODUCTION

We always insist on designing and building systems that can recognize patterns. Automatic optical recognition, facial recognition, fingerprint recognition, speech recognition, DNA recognition, etc. thanks to good results. Optical character recognition is a field of research that seeks to create computer systems that can extract and process text from images. Nowadays, there is a great need to save the information in printed or written documents on computer storage disks so that they can be used on computers later. An easy way to store the information in this file on the computer is to scan it and save it as an image file. But it would be very difficult to read or query the text or other information in this image file to reuse this data. Therefore, there needs to be a technology that automatically collects and stores information (especially text) from image files. Of course, this is not a very important task. In order for automation to be implemented successfully, some important issues need to be raised and resolved. Characterization of the alphabet and the quality of characters in the document are just some of the recent challenges. Due to these problems, the computer may not recognize characters correctly. Therefore, a strategy to recognize the behavior of the Data Analysis Image (DIA) is needed to overcome these problems and create electronic documents from the converted data. new in text format [2]

Similarly, Optical Character Recognition (OCR) is the process or transformation of any text or document containing text (such as written, printed or scanned images) into a digital format that can be manipulated for deep processing. Optical character recognition technology allows machines to recognize letters in such documents. In a real world example, it is like the human body's heart and eyes being together. Eyes can detect, display and extract text from images, but it is actually the human brain that performs the process of detecting or extracting the text that the eye reads [1].

Photocopiers that perform human tasks like reading are an age-old dream. But in the last few years, machine learning has gone from a dream to a reality. Text is everywhere in our daily lives, both in the form of documents (newspapers, books, emails, newspapers, etc.) and in natural situations where people can read it at any time (signs, screens, time). People Unfortunately, blind and visually impaired people are excluded from this information because their blindness does not allow them to access this information, which restricts their path, movement in space is not limited.

OCR-based speech synthesis will be useful to the extent that visually impaired people can interact with the environment in the same way as normal people [3].

©IJRASET: All Rights are Reserved

II. LITERATURE SURVEY

Behavior recognition is not a new problem, but its roots can be traced to the systems that create computer technology. The oldest O CR system is not a computer, but a mechanical device that can recognize characters, but its speed is very slow and its accuracy is lo w. In 1951, M. Sheppard created the reading robot GISMO, which can be considered the oldest work of OCR today. GISMO can rea d individual musical notes and print the text on the page.

only knows 23 characters. The machine also has the ability to create text pages. In 1954, J. Rainbow built a machine that could read capitalized English in one minute. Early OCR systems were criticized for errors and slow recognition. For this reason, not much res earch was done on this subject in the 1960s and 1970s. The structure is found only in government institutions and large businesses s uch as banks, newspapers, and airlines. OCR fonts

should be standardized for easy OCR recognition tasks. Therefore, ANSI and EMCA created OCRA and OCRB in the 1970s, which provide good assurance [4]. There has been a lot of research on OCR in the last three years. This has led to the emergence of multili ngual, text, and full-

font OCR [4]. Despite much research, the ability of machines to read text is still far from humans. Therefore, current OCR research i s dedicated to improving OCR accuracy and speed for a variety documents printed/written in a borderless environment. There is cur rently no open source or commercial software available for complex languages like Urdu or Sindhi etc.

Let's briefly introduce the history of OCR. Gustav Tauschek received a patent for OCR in Germany in 1929, followed by Handel, who received a US patent for OCR in the United States in 1933. In 1935 Tauschek also received a US patent. path. Tauschek machines are mechanical devices that use patterns and photographs. RCA engineers developed the first computer-based OCR to aid the blind at the U.S. Veterans Administration in 1949, but their device did not convert printed text into machine language; Transfer to the machine and speak the text.It turned out to be too expensive and was not accepted after testing.[2] [3].

In 1978 Kurzweil Computer Products began selling a commercial version of the Optical Identification Computer program. LexisNexis was one of the first customers and purchased a program to upload legal documents and information to the online database. Around 1965, Reader's Digest and RCA teamed up to create an OCR reader system designed to digitize bar codes on Reader's Digest coupons returned from advertising. Two years later, Kurzweil sold his company to Xerox, which wanted to add business to converting paper into computers. Kurzweil Computer The product became a subsidiary of Xerox; called Scansoft,now Nuance Communications.

OCR is a technology that allows you to convert different documents, such as scanned documents, PDF

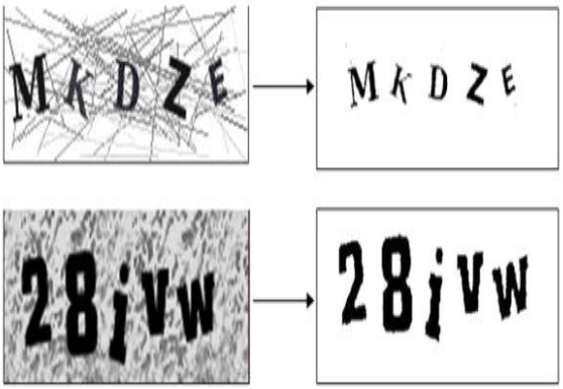
files, or images taken with a digital camera, into editable and searchable documents. Images taken with a digital camera

are different from scanned documents or photos. They often have artifacts such as edge distortion and dim lighting, making it difficu lt for most OCR applications

to recognize text. We chose Tesseract because of its broad acceptance, scalability and flexibility, community of active developers, an d working “outside the box.” To implement the behavior, our application must go through three important steps. The first of these is segmentation, where a binary input image is given to define each

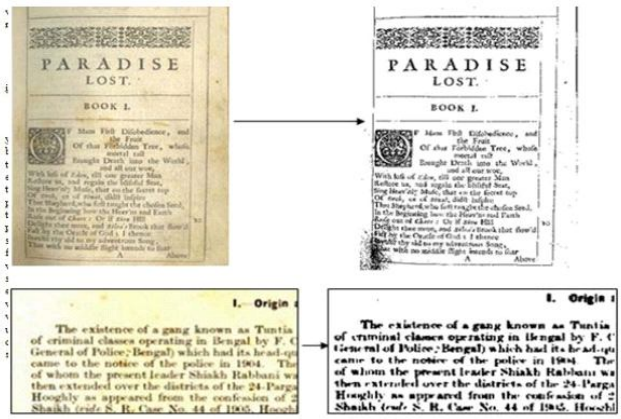
glyph (usually a simple unit representing one or more consecutive

characters). The second step is feature extraction, that is, calculating the number vector from each glyph to be used as input for the ANN [5].



OCR-figure 1

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



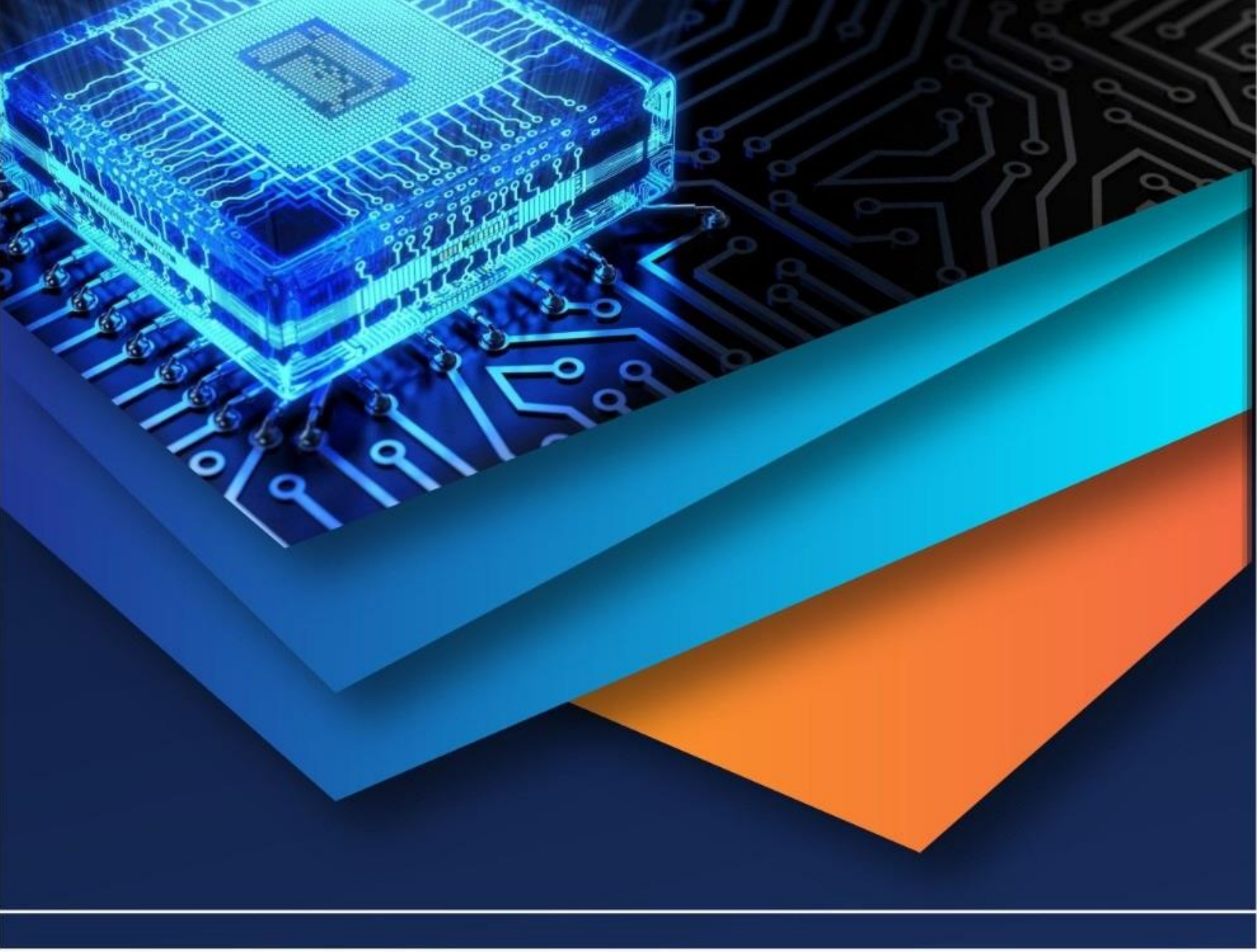
ORC-figure 2

III. CONCLUSION

Optical character recognition involves recognizing optically complete characters. Reliable translation of text from real-world images is a challenging issue due to environmental changes and has been made easier by even the best open source OCR engi nes.

REFERENCES

[1] Patel C, Patel A, Patel D. Optical character recognition by open source OCR tool tesseract: A case study. International Journal of Computer Applications. 2012 Jan 1;55(10).
[2] Ye Q, Doermann D. Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence. 2015 Jul 1;37(7):1480-500.
[3] S.M.A. Haque, et al.Automatic detection and translation of bengali text on road sign for visually impaired.
[4] Mahmoud, S.A., & Al-Badr, B., 1995, Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), 49-77.Filatov, Dmytro, and Ghulam Nabi Ahmad Hassan Yar. "Brain tumor diagnosis and classification via pre-trained convolutional neural networks." arXiv preprint arXiv:2208.00768 (2022).
[5] R.W. Smith, The Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)