IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

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 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104240 |

OPTICAL CHARACTER RECOGNITION

Mohit Gokhale, Binish Moosa, Aayush Gupta, Saakshi Mishra, Mrs. Kadambari Patil

Department of Computer Engineering, Thakur Polytechnic, Mumbai, India

Department of Computer Engineering, Thakur Polytechnic, Mumbai, India

Department of Computer Engineering, Thakur Polytechnic, Mumbai, India

Department of Computer Engineering, Thakur Polytechnic, Mumbai, India

Guide, Department of Computer Engineering, Thakur Polytechnic, Mumbai, India

ABSTRACT: In many different fields, there is a high demand for storing information to a computer storage disk from the data available in printed or handwritten documents or images to later reutilize this information by means of computers. One simple way to store information to a computer system from these printed documents could be first to scan the documents and then store them as image files. But to re-utilize this information, it would be very difficult to read or query text or other information from these image files. Therefore, a technique to automatically retrieve and store information, in particular text, from image files is needed. Optical character recognition is an active research area that attempts to develop a computer system with the ability to extract and process text from images automatically. The objective of OCR is to achieve modification or conversion of any form of text or text-containing documents.

KEYWORDS: OCR, AI, Machine Learning, Image Analysis.

I. INTRODUCTION

Optical character recognition is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, or from subtitle text on an image. Widely used as a form of data entry from printed paper data records – whether passport documents, invoices, bank statements, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly and displayed on-line. This can then be used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. OCR systems are made up of a combination of hardware and software that is used to convert physical documents into machine-readable text. Hardware, such as an optical scanner or specialized circuit board, is used to copy or read text while software typically handles the advanced processing. Software may also use artificial intelligence (AI) to implement intelligent character recognition (ICR), to identify languages or handwritings.

OCR basically exemplifies text recognition. It is a system/software that can automatically recognize text from images, printed text or even handwritten text, but the performance of these systems are directly dependent on the quality of input given. If the input given is not proper then the results can be disapproving. This application is for the Android mobile operating system that combines Google's open-source OCR engine, Tesseract, text recognition OCR engine.

II. HISTORY OF OCR

Optical character recognition is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, or from subtitle text on an image. Widely used as a form of data entry from printed paper data records – whether passport documents, invoices, bank statements, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly and displayed on-line. This can then be used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. OCR systems are made up of a combination of hardware and software that is used to convert physical documents into machine-readable text. Hardware, such as an optical scanner or specialized circuit board, is used to copy or read text while software typically handles the advanced

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

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104240 |

processing. Software may also use artificial intelligence (AI) to implement intelligent character recognition (ICR), to identify languages or handwritings.

OCR basically exemplifies text recognition. It is a system/software that can automatically recognize text from images, printed text or even handwritten text, but the performance of these systems are directly dependent on the quality of input given. If the input given is not proper then the results can be disapproving. This application is for the Android mobile operating system that combines Google's open-source OCR engine, Tesseract, text recognition OCR engine.

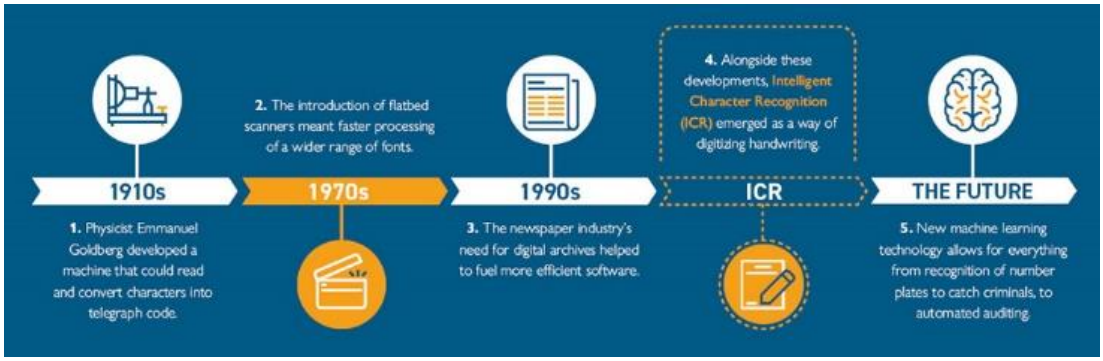

Fig 1. Brief history of OCR

III. HOW DOES OCR WORK

The OCR engine or OCR software works by using the following steps:

i. Image acquisition - A scanner reads documents and converts them to binary data. The OCR software analyzes the scanned image and classifies the light areas as background and the dark areas as text.

ii. Preprocessing - The OCR software first cleans the image and removes errors to prepare it for reading. These are some of its cleaning techniques:
- De-skewing or tilting the scanned document slightly to fix alignment issues during the scan.
- De-speckling or removing any digital image spots or smoothing the edges of text images.
- Cleaning up boxes and lines in the image.
- Script recognition for multi-language OCR technology

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

a) Pattern matching - Pattern matching works by isolating a character image, called a glyph, and comparing it with a similarly stored glyph. Pattern recognition works only if the stored glyph has a similar font and scale to the input glyph. This method works well with scanned images of documents that have been typed in a known font.

b) Feature extraction - Feature extraction breaks down or decomposes the glyphs into features such as lines, closed loops, line direction, and line intersections. It then uses these features to find the best match or the nearest neighbor among its various stored glyphs. iv. Postprocessing After analysis, the system converts the extracted text data into a computerized file. Some OCR systems can create annotated PDF files that include both the before and after versions of the scanned document.

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

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104240 |

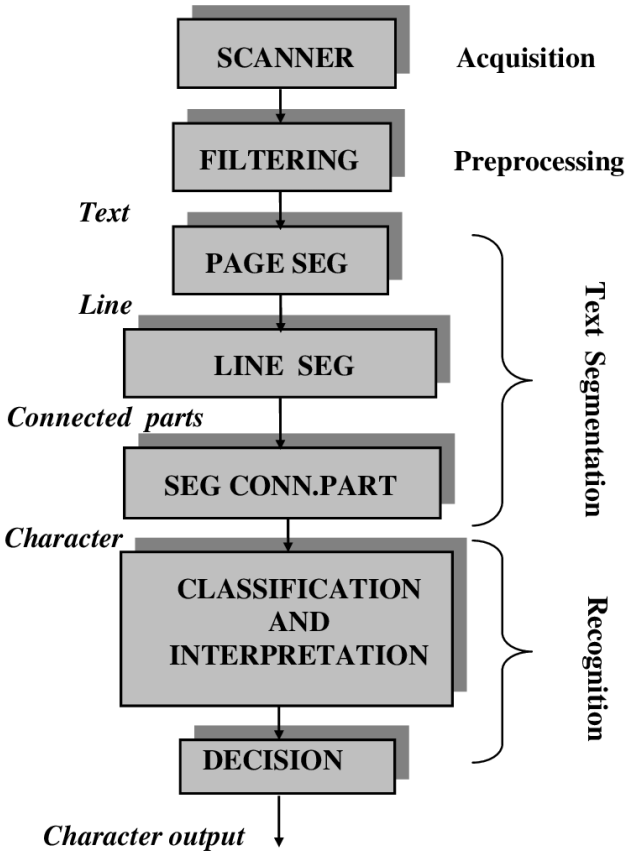

Fig 2. Working of OCR

IV. FINAL PURPOSE OF USING OCR

Many technology companies, with a greater vision, are implementing another kind of Artificial Intelligence, which is Deep Learning. In Deep Learning, a neural network simulates the activity of the human brain to ensure algorithms do not have to rely on history to determine the accuracy but can do it themselves. Deep Learning not only helps OCR recognize text but also identify meaningful information. With modern OCR, banks can quickly and accurately extract important information in credit loan contracts. Therefore, many technology companies, with a greater vision, are implementing another kind of Artificial Intelligence, which is Deep Learning. In Deep Learning, a neural network simulates the activity of the human brain to ensure algorithms do not have to rely on history to determine the accuracy but can do it themselves. Deep Learning not only helps OCR recognize text but also identify meaningful information. With modern OCR, banks can quickly and accurately extract important information in credit loan contracts.

V. FUTURE SCOPE

Recently, the new generation of engineers revived OCR by the integration with Machine Learning built on Artificial Intelligence (AI). This new technology is not limited by the comparison between characters based on the rules of traditional OCR software. With Machine Learning, algorithms are trained with a large amount of data. The new OCR program will accumulate knowledge and learn to recognize any character. Many high-quality OCR solutions were born. However, those solutions have not certainly solved the specific problem of each business.

VI. CONCLUSION

OCR is important because it allows us to digitalize our documents and make them searchable. It also helps in the creation of PDFs, so that we can share them easily. For example, if you have scanned a document from your old hardcopy and want to convert it into a digital document, you will need OCR software. It will help you scan the image and get the text written in that specific image by recognizing the character's with the help of technologies such as

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

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104240 |

artificial intelligence and machine learning The implications of OCR are much more than just digitizing documents. It not only helps in the creation of PDFs but also helps in digitizing books, magazines and newspapers which can be used for research purposes or simply as a source of information and the data can be archived later as well..

REFERENCES

[1] Juha-Pekka Soininen , Kari Kolehmainen, Hannu Tanner,"Irrigation water saving estimation using soil moisture forecast simulation*'',* 978-1-6654-0533-1/21/$31.00 ©2021 IEEE

[2] Surender Singh, Surender Singh," Sustainable and Smart Agriculture: A Holistic Approach", 978-1-6654-3789-9/22/$31.00 ©2022 IEEE

[3] Rishabh Sachan, Sanmukh Kaur, Anil Kumar Shukla,"Smart Irrigation and Security System for Agricultural Crops and Trees", 978-1-6654-1703-7/21/$31.00 ©2021 IEEE.

[4] G. Santhakumar, R. Vadivelu, K. Harshini," Smart Irrigation System for Agriculture using Wireless Sensor Network", 978-1-6654-0521-8/20/$31.00 ©2021 IEEE

[5] R.Karthikamani , Harikumar Rajaguru "IoT based Smart Irrigation System using Raspberry Pi", 978-1-6654-1806-5/21/$31.00 ©2021 IEEE

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