International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 425-429

Rapid Scanning and OCR Technology

[1]Yashas N, [2]Sparash Agarwal, [3]Syed sufiyan, [4]Anup p, [5]Dr. Nandini Prasad KS
[1]Student, [2]Student, [3]Student, [4]Student, [5]HOD-ISE
[1]Department of Information Science and Engineering
[1]Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Abstract - *n this current scenario, where everything is in the digital form, there are many documents that one wishes to convert into digital format. Since digital documents are not only easy to store but also to edit and find information from them. There are various means to convert any printed document into a digital document but there are a very few technologies that help to convert hand written text into editable document.*

The OCR technology plays a very important role in doing so. With the help of OCR technology the scanning process is not only rapid but also much more accurate as compared to any other scanning technologies.
In this method each letter is scanned and compared to the library in tesseract.js.

This technology is helpful to convert old documents that contains broken test too. The Major purpose is to make the scanning of handwritten text document into editable and storable documents.

From a historical perspective, research and development of OCR systems are taken into account. Included is the progression of commercial systems throughout time. R&D strategies using template matching and structural analysis are both taken into consideration. It is seen that the two strategies are blending and getting closer together. Commercial products are broken down into three generations, with a few exemplary OCR systems chosen for each and briefly explained. Expert systems and neural networks, two contemporary OCR approaches, are discussed briefly along with some unresolved issues. The writers' opinions and predictions about next trends are offered.

Date of Submission: 02-05-2023 Date of acceptance: 14-05-2023

I. INTRODUCTION

OCR stands for Optical Character Recognition. It is a widespread technology to recognize text inside images, such as scanned documents and photos. OCR technology is used to convert virtuallyany kind of image containing written text (typed, handwritten, or printed) into machine-readable text data.

The most well-known use case for OCR is converting printed paper documentsinto machine- readable text documents. Once a scanned paper document goesthrough OCR processing, the text of the document can be edited with word processors like:

• Microsoft Word
• Google Docs

Before OCR technology was available, the only option to digitize printed paperdocuments was manually re-typing the text. Not only was this massively time- consuming, but it also came with inaccuracy and typing errors.

OCR is often used as a “hidden” technology, powering many well-known systems and services in our daily life. Less known, but as important, use cases for OCR technology include:

• Passport recognition for airports
• Traffic sign recognition
• Extracting contact information from documents or business cards
• Converting handwritten notes to machine-readable text
• Defeating CAPTCHA anti-bot systems
• Making electronic documents searchable like Google Books or PDF

One of the most crucial image analysis jobs is optical character recognition. Its principal uses include creating digital libraries (including text, mathematic formulae, music scores, etc.), identifying items on digitalized maps, locating car licence plates [1], text readers for the blind, and deciphering handwritten documents like checks and office forms. The following stages make up a typical OCR system:

- picture preparation, such as noise reduction and orientation correction
- adaptive picture binarization, which is often used
- identifying page layout, detecting text sections (and tables, figures, etc.), then text paragraphs, individual lines, then segmenting lines into words, and ultimately segmenting words into characters; segmentation , typically hierarchical;
- real recognition (supervised or unsupervised)
- postprocessing with the help of a spellchecker;

After conducting tests, two conclusions were drawn. The OCR process remained unaffected by uneven illumination, whereas noise had a substantial impact on accuracy. However, even with 10% noise and a high resolution of 600 dpi, precision could still be achieved. The median filter was found to be unsuitable and reduced accuracy further. Lower image resolution than 300 dpi had a significant impact on accuracy. Interestingly, increasing the resolution of a low-quality image resulted in a noticeable improvement in accuracy. These findings were reached after conducting thorough tests in a professional setting.

Nevertheless, geometric distortions were what OCR recognition was most susceptible to. It is challenging to trace text lines when there are these kinds of abnormalities, which may indicate that subsequent OCR processing of an image that has been rotated by, say, 10 degrees completely fails. Geometric deformations, in contrast to noise and poor resolution obstructions, may be (theoretically) removed once we are aware they may arise.

II. LITERATURE SURVEY

In [1] In their 2021 paper, P. A. Khaustov and V. G. Spitsyn put forth an Algorithm that proposes a novel method for recognizing optical handwritten characters. The Algorithm primarily relies on extracting the structural components of the characters to identify them accurately. To achieve this, the Algorithm involves a series of steps, including preprocessing the input image, segmenting the image into individual characters, and extracting structural components such as strokes and junctions. These components are then compared to a reference set using a classifier to identify the characters. Overall, the proposed Algorithm offers a professional and innovative approach to recognizing optical handwritten characters.

In [2] In their research paper, "Handwritten Character Recognition using Neural Network for Encryption System," Aarnav Pant, Babita Sonare, and Abhishek Mule (2020) present the use of a neural network for recognizing handwritten characters in an encryption system. The authors suggest a system where handwritten characters serve as the keys for encrypting and decrypting messages. By training the neural network on a dataset of handwritten characters, the system can accurately recognize and classify them. The authors highlight the advantages of using handwritten characters in an encryption system, such as enhanced security and user-friendliness. Overall, the proposed system exhibits potential for application in secure communication systems.

In [3] Bermudez Castro and Arauco Canchumuni (2021) presented a research paper on how Generative Adversarial Networks (GANs) can be used to enhance Optical Character Recognition (OCR) for structured documents. The paper highlights that structured documents, such as tables and forms, often have intricate layouts, making OCR a challenging task. The authors suggest that GANs can generate synthetic training data for the OCR system, eventually enhancing its performance on structured documents. The research results reveal that utilizing GANs can significantly improve the accuracy of OCR on structured documents. Furthermore, the authors discuss the potential of using GANs in OCR and document processing applications.

In[4] In their 2017 paper, Anushri Arora and Aniruddh Chandratre proposed using Optical Character Recognition (OCR) to extract and recognize text from handwritten forms with dynamic layouts. The authors suggest a combination of machine learning algorithms and human verification to accurately classify text from forms commonly found in insurance claims or surveys. However, the authors also acknowledge the challenges and limitations of this approach, including the need for high-quality input images and difficulty in recognizing handwriting from various writers. Nonetheless, this system offers promise for improving data entry and processing efficiency in industries that rely on handwritten forms.

In [5] In their 2018 study, Vikas J. Dongre and Vijay H. Mankar proposed a method for recognizing handwritten Devanagari numerals and characters using a neural network classifier and multiple features. Devanagari is a script used for writing languages such as Hindi and Nepali in the Indian subcontinent. The authors extracted geometrical, statistical, and structural features from the handwritten samples and used them to train the neural network classifier. The classifier achieved high recognition rates for both numerals and characters, and the authors compared their method to others to show its competitiveness. This research is significant as it enables the recognition of Devanagari script, which is crucial for various language-based applications in the region.

In [6] ChandniKaundilya and Diksha Chawla (2020) Automated Text Extraction from images using OCR System. The popularity of digital photographs is rising quickly. According to their various demands, several organisations, including students, engineers, and physicians, produce a large number of photos every day. They have access to photos depending on the accompanying text or the image's basic attributes. Such graphics may contain text that contains useful information. Our goal is to automatically extract the content and condense the visual data from photographs. For this, an optical character recognition system with several algorithms is needed. Tesseract, which was created by HP Labs and is presently owned by Google, is the most accurate optical character recognition engine currently available. In this study, we employ text localization, segmentation, and binarization methods to extract text from photos. Text localization pinpoints the exact location of the text, text segmentation separates the text from its backdrop, and binarization turns coloured pictures into binary. These techniques may all be used to extract text from an image. Character recognition is used to transform this binary picture into ASCII text. The creation of electronic books from scanned books, image searching from a collection of visual data, etc. all require text extraction.

In [7] In 2017, a research article by Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin presented a survey of OCR applications. OCR, or optical character recognition, refers to the electronic transformation of handwritten, typewritten, or printed text into machine-readable images. This technology is commonly used to extract text from electronic documents, search it, and publish the material online. The study offers an overview of OCR's applications in various domains, followed by detailed experiments with three significant applications, including Captcha, institutional repositories, and optical music character recognition. To achieve optical character recognition, the researchers utilized evolutionary algorithms with an enhanced histogram equalization-based image segmentation approach. This research article can serve as a valuable literature review for scholars seeking information on the topic.

In [8] According to the 2019 publication by Hubert Michalak and Krzysztof Okarma, Optical Character Recognition (OCR) applications demand consistently illuminated images for their operation. Flatbed scanners are the conventional source of such images. However, with the rise of mobile technology, document photos captured through the built-in cameras of current mobile devices are becoming increasingly prevalent. Though businesses and administrative offices accept high-resolution images captured by mobile phones and tablets, the uneven lighting in such photos can create difficulties in identifying text through OCR programs. The lack of common 2D codes such as QR or Aztec codes exacerbates the problem. To overcome this issue, image preprocessing, including binarization, is necessary. Unfortunately, standard global thresholding fails to achieve this due to the presence of local intensity variations.

In [9] K.Karthick and K.B.Ravindrakumar., (2020) The previous two centuries have seen an astounding and noble development curve thanks to technology. For the past few decades, employing a mouse and keyboard to serve as an interface between humans and computers has been simple. However, while the potential for human-based communications to connect with a computer would make things easier to handle, it would be challenging for the researchers and investigators to achieve.Pioneering developments brought about by ongoing study in man-machine communication may result in situations resembling human interactions. The automation requirements in diverse applications are met by a variety of methods employing magnetic stripes, speech recognition, identification using radio frequency, bar codes, and Optical Mark Recognition (OMR) and OCR. This essay covered the categorization of handwritten OCR systems and the OCR process.

[10] Todsanai Chumwatana and Waramporn Rattana-umnuaychai have proposed an OCR framework for document digitization. In today's digital age, many firms and organizations are opting for digital transformation. However, before the data can be analyzed, it must first be saved and organized in a useful manner. To enable future use, physical documents, scanned copies, photographs, and PDFs must be converted into digital format. The aim of this study is to provide a method for extracting text from physical documents and converting it into digital form through optical character recognition (OCR). The proposed method can extract all text from photocopies and convert them into database structures, making the digitized documents entirely searchable and editable. The experimental investigations have revealed that the suggested method has an average accuracy performance of 75.38% for extracting characteristics.

III. METHODOLOGY

When it comes to OCR techniques, there are various methods that rely on vision to extract textual regions and predict the bounding box coordinates for those sections. In order to turn this information into textual data, language processing techniques use RNNs, LSTMs, and Transformers.

Deep learning-based OCR systems typically involve two stages: region proposal and language processing. The region proposal stage involves identifying text-rich regions in the image, using convolutional models that recognise text fragments and enclose them in bounding boxes. This is similar to the Region Proposal Network used in object detection algorithms like Fast-RCNN.

Once potential regions of interest have been identified, they serve as attention maps and are given to language processing algorithms. These algorithms use RNNs and Transformers, two NLP-based techniques, to decode the feature-based information and turn it into textual data.
to detect text that has limited temporal information to relay, such as signboards or vehicle registration plates.

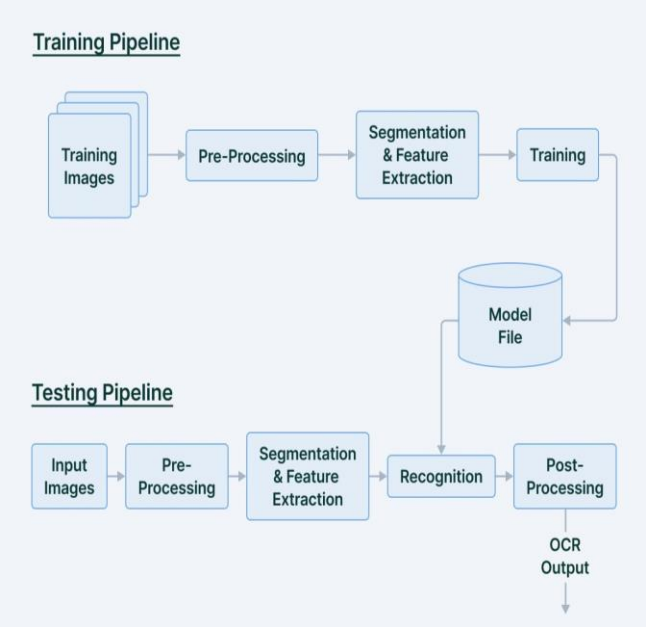

Fig 1: Flow Diagram for Training pipeline

To distinguish between vacant and occupied areas, optical character recognition (OCR) dissects the image of a written character into fragments. The resulting matrix is then subjected to a checksum, which is initially verified by a human, in accordance with the typeface or script employed for the alphabetic character. A contemporary OCR training workflow includes the following steps.

3.1 Acquiring
obtaining non-editable text content from all forms of scanned documents, including flatbed scans of corporate archival materials, live surveillance footage, and mobile image data.

3.2 Preparation
At the aggregate level, the raw imagery is cleaned up to make the text easier to read and to decrease or remove noise.

3.3 Feature extraction and segmentation

When it comes to identifying characters within images, our approach involves searching for clusters of pixels that bear a resemblance to individual letters or symbols. To ensure accuracy and efficiency, each cluster is assigned to a specific category. This is done by utilizing a machine learning framework, which draws upon either generalized OCR templates or previous models. Through this process, we are able to provide a professional and reliable means of character recognition.

3.4 Instruction
The data can be handled in a neural network training session once all features have been identified. During this session, a model will try to create a generic image>text mapping for the given data.

3.5 Retraining and verification

In the field of data processing, human review of outcomes is a crucial step that helps correct any errors and improve subsequent training sessions. However, the quality of the data being processed needs to be examined carefully. During initial training runs, techniques like de-skewing and high contrast processing are used to create a decent algorithm with minimum pre-processing. Yet, in certain cases, refining the data may require more laborious efforts such as data cleaning which can be time-consuming and expensive.

This entire process is referred to as a stream processing pipeline, which involves generating the stream data, processing it, and ultimately delivering it to its final location. For modern applications, stream processing has become an essential component. Enterprises are adopting technologies that respond to data in real-time for a variety of use cases and applications. Some prime examples include data from IoT sensors, payment systems, and more. Stream processing is most commonly applied to data that is generated as a series of events.

IV. CONCLUSION

The ongoing study of Optical Character Recognition (OCR) primarily focuses on recognition itself, with less emphasis on preprocessing. However, our research paper highlights the crucial role of adequate OCR image preprocessing, especially for photos captured using non-professional digital cameras. These cameras are increasingly becoming the primary source of picture data, including for text scanning. Therefore, future OCR software iterations are likely to include the techniques presented in our study.

Based on a broad range of historical newspapers, it can be concluded that the Otsu approach's core premise best approximates the reality behind historical printed texts. This finding holds significant implications for professionals in the field of OCR, as they seek to improve accuracy and efficiency in text recognition. Additionally, our research can benefit individuals, such as website administrators, bloggers, and marketers, seeking to enhance their web pages by optimizing OCR image preprocessing techniques.

More generally, this research encourages us to surmise that the extra benefits of binarization with a black box OCR are quite small. A system with a feedback loop where the OCR determines confidence, a system with human input for training, or a system that employs numerous local binarizations and a natural language processing module after the OCR to make final decisions are all likely to achieve greater results.

REFERENCES

[1]. A. Singh, K. Bacchuwar, A. Choubey, S. Karanam, "A Novel GA Based OCR Enhancement and Segmentation Methodology for Marathi Script in Bimodal Framework"in Springer Verlag, (2021).

[2]. Weszka, J.S., Nagel, R.N., Rosenfeld, A. "A Threshold selection technique", IEEE Trans. Computer (2020)

[3]. R Plamondon, S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey" IEEE transaction on pattern Analysis and machine Intelligence, 2021,

[4]. J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognition (2017)

[5]. C. Wolf, D. Doermann, Binarization of low quality text using a Markov random field model, in: Proceedings of the 16th International Conference on Pattern Recognition, vol. 3, 2020.

[6]. L. O'Gorman, Binarization and Multithresholding (2017) of document image using connectivity, in: CVGIP: Graphical Models and Image Processing, vol. 56, No. 6, 1994, pp. 494–506.

[7]. L. O'Gorman, Experimental comparisons of binarization(2015) and multithresholding methods on document images, in: Proceedings of the IAPR International Conference on Pattern Recognition, vol. 2, IEEE, 1994, pp. 395–398

[8]. Kaggal, V.C., Elayavilli, R.K., Mehrabi, S., Joshua, J.P., Sohn, S., Wang, Y., Li, D., Rastegar, M.M., Murphy, S.P., Ross, J.L., et al.: Toward a learning health-care system-knowledge delivery at the point of care empowered by big data and NLP. Biomed. Inf. Insights **8**(Suppl1), 13 (2016).

[9]. Pal, G., Li, G., & Atkinson, K. (2018). Multi-agent big-Hatch, R.: SaaS Architecture, Adoption and Monetization of SaaS Projects using Best Practice Service Strategy, Service Design, Service Transition, Service Operation and Continual Service Improvement Processes. Emereo Pty Ltd., London (2021)

[10]. ChandniKaundilya and Diksha Chawla (2020) Automated Text Extraction from images using OCR System

[11]. Tesseract.js, a pure javascript version of the tesseract OCR engine (2020)).

[12]. Rice, S.V., Jenkins, F.R., Nartker, T.A.: The fourth annual test of OCR accuracy. Technical report, Technical Report 95 (2017).

[13]. Bautista, C.M., Dy, C.A., Mañalac, M.I., Orbe, R.A., Cordel, M.: Convolutional neural network for vehicle detection in low resolution traffic videos. In: 2016 IEEE Region 10 Symposium (TENSYMP), pp. 277–281. IEEE (2016).

[14]. B. Plessis, A. Sicsu, L. Heutte, E. Menu, E. Lecolinet, O. Debon, J. V. Moreau, 2013, A multi-classifier combination