Rapid Scanning and OCR Technology

[1]Aditya Kumar Agarwal, [2]Abhishek Mishra, [3]Abhijeet Kumar, [4]Anaghashree CA, [5]Ms. Mangala HS
[1,2,3,4] *Student, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India*
[5]*Assistant Professor, Department of Computer Science and Engineering, 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 handwritten 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.

1. 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 virtually any 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 documents into machine- readable text documents. Once a scanned paper document goes through 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 paper documents 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;

Two findings were reached as a result of the tests.Uneven illumination did not significantly hinder the OCR process. The effects of noise were substantially worse, yet even with 10% of the noise and a high resolution (600 dpi), precision could still be achieved. The median filter seemed completely out of place in our situation, which further reduced accuracy. If the image resolution is lower than 300 dpi, the FineReader's accuracy suffers significantly. However, it's fascinating to see that upsampling a low-resolution image leads to a noticeable boost in accuracy.

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.

2. LITERATURE SURVEY

In [1] P. A. Khaustov; V. G. Spitsyn., (2021) has proposed an Algorithm for optical handwritten characters recognition based on structural components extraction. The paper describes a handwriting recognition algorithm based on extracting structural components from handwritten characters. The algorithm begins by preprocessing the input image to enhance the contrast and remove noise. Then, it segments the image into its individual characters using techniques such as thresholding and connected component analysis.Once the characters have been segmented, the algorithm extracts their structural components, such as strokes and junctions, which are the intersections between strokes.

These structural components are used to represent the characters in a more abstract form, which makes them easier to compare to a reference set of characters.The algorithm then uses a classifier to identify the characters based on their structural components. This can be done using a variety of techniques, such as support vector machines or neural networks. The classifier is trained on a set of reference characters, and it uses this training data to learn how to recognize the characters based on their structural components.

Finally, the algorithm outputs the recognized characters and their corresponding class labels. The authors of the paper report that their algorithm achieved good results on a dataset of handwritten characters, with an accuracy of over 90%.

In [2] Aarnav Pant; Babita Sonare and Abhishek Mule, (2020) proposed Handwritten Character Recognition using Neural Network for Encryption System his paper discusses the use of a neural network for handwritten character recognition in an encryption system. The authors propose a system in which handwritten characters are used as the keys for encrypting and decrypting messages. The neural network is trained on a dataset of handwritten characters and is able to recognize and classify them with high accuracy. The authors also discuss the benefits of using handwritten characters in an encryption system, including the high level of security and the ease of use for the user. Overall, the proposed system shows promise for use in secure communication systems.

In [3] Jose D. Bermudez Castro and Smith W. Arauco Canchumuni (2021) presented Improvement Optical Character Recognition for Structured Documents using Generative Adversarial Networks. his paper explores the use of generative adversarial networks (GANs) to improve optical character recognition (OCR) for structured documents. Structured documents, such as forms and tables, often have complex layouts that make OCR difficult. The authors propose using a GAN to generate synthetic training data for the OCR system, which can then be used to improve its performance on structured documents. The results show that the use of the GAN significantly improves the accuracy of OCR on structured documents. The authors also discuss the potential for using GANs in other applications related to OCR and document processing.

In [4] Anushri Arora and Aniruddh Chandratre (2017) has suggested Optical Character Recognition For Handwritten Forms With Dynamic Layout. This paper discusses the development of an optical character recognition (OCR) system for handwritten forms with dynamic layout. The authors propose a method for extracting and recognizing text from handwritten forms with variable layouts, such as those found in insurance claims or surveys. The system uses a combination of machine learning algorithms and human-in-the-loop verification to accurately extract and classify text from the forms. The authors also discuss the challenges and limitations of this approach, including the need for high-quality input images and the difficulty of recognizing handwriting from different writers. Overall, the proposed system shows promise for improving the

efficiency of data entry and processing in industries that rely on handwritten forms.

In [5] Vikas J. Dongre and Vijay H. Mankar., (2018) stated Devanagari offline handwritten numeral and character recognition using multiple features and neural network classifier. This paper presents a method for recognizing Devanagari numerals and characters from handwritten samples using a neural network classifier and multiple features. Devanagari is a script used to write several languages spoken in the Indian subcontinent, including Hindi and Nepali. The authors extract several features from the handwritten samples, including geometrical, statistical, and structural features. These features are then used to train a neural network classifier, which is able to achieve high recognition rates for both numerals and characters. The authors also compare their method to other existing methods and show that it performs competitively. This work is significant as it enables the recognition of Devanagari script, which is important for many 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 colored 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] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin (2017) proposed A Survey of OCR Applications The electronic conversion of handwritten, typewritten, or printed text into machine translated pictures is known as optical character recognition, or OCR. It is frequently used to extract text from electronic documents, search it, and publish the material online. The study provides an overview of OCR's uses in several domains and then details experiments with three key applications, including Captcha, institutional repositories, and optical music character recognition. We utilise evolutionary algorithms with an improved histogram equalization-based picture segmentation approach for optical character recognition. The article will serve as a useful literature review for scholars looking for

In [8] Hubert Michalak and Krzysztof Okarma (2019) propose Applications for optical character recognition (OCR) typically need for the utilisation of consistently lighted pictures, which flatbed scanners can produce. However, because to the quick advancement of mobile technology, document photos taken with the built-in cameras of current mobile phones and tablets are becoming more and more common. High resolution images that can be sufficient, for example, for insurance purposes, are also accepted by many businesses and administrative offices in addition to scanned copies of the papers. Sadly, such photos may have uneven lighting, which can cause issues with text identification using OCR programmes, particularly if QR, Aztec, or other common 2D codes are absent. In order to properly recognise text from camera photos, image preprocessing is necessary, including binarization. This cannot be done using standard global thresholding since local intensity variations are present. On the other hand, pixel-based adaptive approaches need a lot of time and don't always produce appropriate outcomes. A region-based method to image binarization, which is an extension of the well-known Niblack thresholding technique, is suggested in this study to address this gap and strike a compromise between recognition accuracy and fast processing speed.

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 proposed the Extraction for documents Digitization using OCR frameworkMany firms and organisations are currently taking into consideration the digital transformation. In order for analytics to be proposed, the data has been saved and organised into a useful way. To be used in the future, certain material that has been stored in the form of hard copies, scanned documents, photographs, and PDFs has to be converted into digital form. The goal of this study is to provide a method for extracting text from a physical document and converting it to digital form using optical character recognition (OCR), which aims to extract all text from photocopies and convert them into database structures. The experimental investigations revealed that the suggested method entirely searchable and editable the digitised documents with an average accuracy performance about 75.38% for extracting characteristics.

3. METHODOLOGY

OCR techniques often involve methods based on vision that extract textual regions and forecast the bounding box coordinates for those sections. The language processing techniques use RNNs, LSTMs, and Transformers to decode the feature-based information into textual data after receiving the bounding box data and picture features. The region proposal stage and the language processing stage are the two steps of deep learning-based OCR systems. Region Proposal: The initial step in OCR entails identifying text-rich regions in the image. Convolutional models that recognise text fragments and enclose them in bounding boxes are used to achieve this. Similar to the Region Proposal Network in object detection algorithms like Fast-RCNN, this network's task entails marking and extracting potential regions of interest. Along with information derived from the image, these areas serve as attention maps and are given to language processing algorithms. Language processing: RNNs and Transformers, two NLP-based networks, attempt to extract information from these regions and create comprehensible sentences using characteristics fed from the CNN layers. Recent studies have successfully investigated entirely CNN-based algorithms that recognise characters directly without going through this stage. These algorithms are particularly helpful to detect text that has limited temporal information to relay, such as signboards or vehicle registration plates.

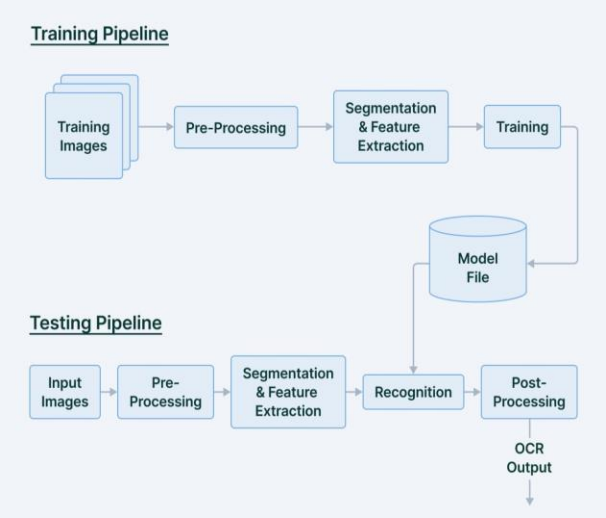

Fig 1: Flow Diagram for Training pipeline

In order to discriminate between empty and non-empty regions, optical character recognition divides the picture of a written character into pieces. The checksum of the resulting matrix is then identified (at least initially by a person) as relating to the character in the image, depending on the typeface or script used for the letter.

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

Searching the image content for clusters of pixels that resemble single characters and assigning each one to a different class. On the basis of generalised OCR templates or earlier models, the machine learning framework will then try to generate characteristics for the recurrent pixel clusters that it detects.

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

Humans review the outcomes following processing, with corrections being transmitted back into subsequent training sessions. The quality of the data may now need to be examined. While first training runs will do de-skewing, high contrast processing, and other useful ways to create a decent algorithm with minimum pre-processing, more laborious refinement of the data may be necessary. Data cleaning is time-consuming and expensive. both. This workflow is referred to as a stream processing pipeline, which includes the generation of the stream data, the processing of the data, and the delivery of the data to a final location. Stream processing has become a must-have for modern applications. Enterprises have turned to technologies that respond to data at the time at which it is created for a variety of use cases and applications, examples of which we'll cover below. Stream processing is most often applied to data that is generated as a series of events, such as data from IoT sensors, payment processing systems, and server

4. CONCLUSION

Although optical character recognition is a topic of ongoing study, the majority of efforts focus on recognition itself rather than preprocessing. We demonstrated in our paper that adequateOCR image preprocessing is crucial, especially for photos taken using a digital camera. Non-professional digital cameras are quickly taking over as the primary source of picture data, including for text scanning. Future iterations of OCR software are likely to include the kinds of techniques presented in this study. Behaviour from them, also this technique and procedure are relevant to any real-time data analysis. The information can be utilized for a different reason, for marketing. Furthermore, scientists, any website admin, blogger or individual with a site can find out about how to enhance their webpage.

Based on findings across a wide range of historical newspapers, it may be concluded that the core premise of the Otsu approach best approximates the reality behind historical printed texts. It has been demonstrated that the suggested multiresolutional variant consistently enhances performance.

The usefulness of more complicated binarization and picture preparation methods was not as evident as may logically be anticipated.

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

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

[12] 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).

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