DEEP LEARNING – DRIVEN OPTICAL CHARACTER RECOGNITION

K. Vijayakumar[1], Abhay Kumar[2], Ankit Raj Sharma[3]
[1]Department of CSE-AIML, AMC Engineering College, Bangalore
Email:- kmk.vijay@gmail.com
[2]Department of CSE-AIML, AMC Engineering College, Bangalore
Email:- abhayacc201@gmail.com
[3]Department of CSE-AIML, AMC Engineering College, Bangalore
Email:- sharmaankit9849@gmail.com

Abstract
The technological translation of typewritten, handwritten, or printed text into different machine-coded text—whether either from a scanned document or from a photo of a document—is known as optical character recognition. Right now, entering data into computers is still mostly done via keyboarding, which is also perhaps the most labor- and time-intensive process. OCR has been extensively studied over the last three decades and has successfully replicated human reading. It involves mechanically or electronically converting scanned images of printed, typewritten, or handwritten text. OCR converts printed words into digital format so that they can be electronically searched through and used in automated processes. Text that is machine-encoded and suitable for text mining, machine translation, and text-to-speech applications is created from photos. OCR offers a quick, easy, and affordable solution for reading documents with handwritten text or preset font sizes and styles. Information gated from various printed paper records, including business cards, passports, paper invoices, bank statements, digital receipts, postal delivery of printed bank statements, and other appropriate paperwork, is common.

Key Words: Optical Character Recognition, Tesseract, OpenCV, Python

1. INTRODUCTION

One of the most popular uses of computer vision is text recognition, which is employed by several tech giants like Apple, Google, and others. Apple just revealed that iOS 15 would have the "Live Text" functionality. The Google Search and Photos app for iOS devices. one must point the camera toward an image or text that is shown on a paper or board sign. Whether it is an email address or phone number, the Live Text function can identify the text that is Show in the picture. These functions rely on OCR, a service or technology. OCR has been the only way to convert printouts into computer-processable data for decades, and it remains the technique of choice for converting paper bills into data extracted and connected to financial systems.

Conventional OCR systems are frequently susceptible to noise, contrast, and picture resolution, which can skew the text and reduce recognition precision. Although image preprocessing has advanced, methods are unable to reliably handle the variety of scenarios in which text is collected. The usefulness of OCR in real-world scenarios when picture circumstances are not optimal is restricted by its inability to handle photos of differing quality in a robust manner.

2. LITERATURE REVIEW

Since the early days of technology—long before computers—optical character recognition has faced difficulties. The initial devices were the slowest processing machines with lower quality character recognition capabilities. The first device was patented in 1929 by Gustav Tauschek in Germany, and Handel in the United States in 1933. Tauschek's device recognized characters using photodetectors and templates. In 1949, the Radio Corporation of America (RCA) began developing a character recognition system for the visually impaired. The equipment was very expensive and inefficient at first.

1951 is regarded as a turning point in the history of OCRs since someone from the M. Sheppard group invented GISMO at that time, laying the groundwork for modern OCR. In addition to having symbols for every character from A to Z, GISMO can detect music. In addition to having symbols for every character from A to Z, GISMO can detect music. J. Rainbow successfully develop an OCR in 1952 that could only process one English capital letter per minute. The necessity for OCR fonts to be standardized was highlighted by these early OCR initiatives, and as a result, OCR-A and OCR-B, as well as ANSI and ECMA, were interbacked, improving OCR. OCRs were originally commercialized by the Kurzweil Computer Corporation in 1978, mainly for converting hard copy materials into digital libraries.

As Kurzweil Computer joined Xerox as a subsidiary in 1967, they developed the pixel-sharpening technology. The other approaches, however, were statistical judgments and parametric procedures utilizing Gaussian distributions. However, this needed a lot of processing power. For OCR enhancement, a variety of nonlinear multimodal function optimization techniques, fuzzy logic, and neural networks are employed. For instance, crowding causes a bimodal histogram to provide multimodal optimization issues, which in turn cause handwritten English and other OCR processes to the Document Image Analysis (DIA).

Though recent advancements have been achieved, OCR is still unable to comprehend text at the human level. Picture extraction and develop a threshold-setting system that optimizes algorithm computation using genetic algorithms. New capabilities in the field of OCR have been made possible by recent developments in study, and the body of knowledge even extends beyond computers and robots to human-like levels. The available OCR types on the cloud-based OCR and the issues they aim to resolve are the main topics of this section.

You can provide handwriting and printed OCR input. Conventional OCR systems relied mostly on document layout and common character styles. Nevertheless, OCR technology faces difficult hurdles due to the differences in writing styles and character kinds of many languages. Online and offline OCR systems can be identified based on the type of task and data involved. To acquire less complexity, an online OCR program processes the written text data in parallel with the user's writing speed. Simultaneously, offline systems work with static data, which forces the computers to perform more intricate pattern recognition. Users have come to love the online systems because of their accuracy, ease of creation, and tablet appendix. OCR types are shown in the insertion image.

1.1 PROBLEM STATEMENT

Many printed books and newspapers on a various topics. The difficulty here is getting software to identify characters in the computer system when information is scanned through paper materials. Every time we scan a document using a scanner, the document is saved in the computer system as an image (jpeg, gif, etc.). The user is unable to read or alter these images. However, it is quite challenging to read each section separately and do a word-by-word and line-by-line search of these documents' contents in order to reuse this information. "Storing the information available in these paper documents into a computer storage disk and later editing or reusing this information by searching" is a highly sought-after practice these days. The persistent shortcomings of OCR technology have significant effects on accessibility and data management. Errors in important data, inefficiencies in information retrieval, and a rise in the amount of manual labor required for data correction might result from inaccurate text recognition. These difficulties can lead to operational hiccups and decreased productivity in industries like legal and financial services, which mostly depend on precise document processing.

There are many types of document layouts available, ranging from simple single-column designs to intricate multi-column layouts that incorporate a variety of elements, such as annotations, tables, and photographs. In many cases, OCR systems today cannot handle complex document formats or layouts. As a result of irregular formatting, embedded graphics, or non-linear text layouts, this limitation can result in inaccurate or partial text extraction for documents.

Real-time processing capabilities are required in many practical applications, such as live text recognition and mobile document scanning. Current OCR systems frequently encounter difficulties striking a balance between accuracy and processing speed, which can cause delays or performance sacrifices. Addressing the essential difficulty of managing real-time data efficiently while retaining high accuracy is imperative.

1.2 COMPONENT OF OCR SYSTEM

An OCR system typically comprises of a few essential parts, as shown in a typical configuration. Using an optical scanner, the analog document is first digitalized to start the process.

Fig 1: Optical Character Recognition System

1. Image Acquisition

- **Scanners/Cameras:** Devices used to capture the images of the documents.
- **Input Formats:** Handling various input formats such as JPEG, PNG, TIFF, PDF, etc.

2. Pre-processing

- **Noise Reduction:** Techniques to remove noise from images to enhance text clarity.
- **Binarization:** converting images to black-and-white to simplify text extraction.

3. User Interface

- **Display Interface:** Providing a visual interface for users to view and verify the OCR output.
- **Editing Tools:** Allowing users to manually correct OCR errors and adjust the output.

4. Integration and APIs

- **APIs and SDKs:** Providing interfaces for integrating OCR capabilities into other software applications.
- **Batch Processing:** Supporting the processes of multiple documents/images simultaneously.

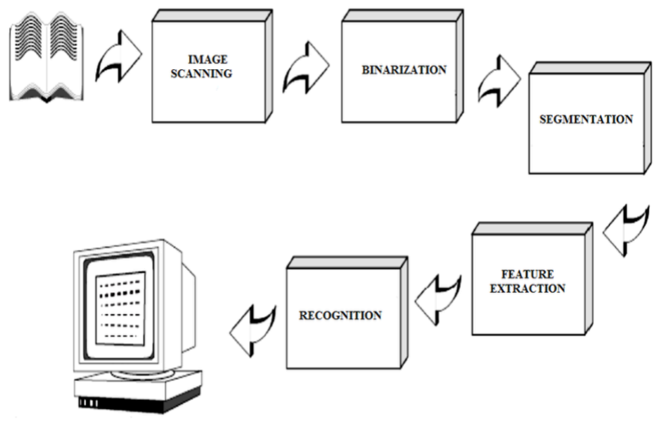

Fig 2: Component of OCR System

2. PROPOSED WORK

The technology known as optical character recognition, or OCR, transforms various document formats. An OCR project's suggested workflow is as follows:

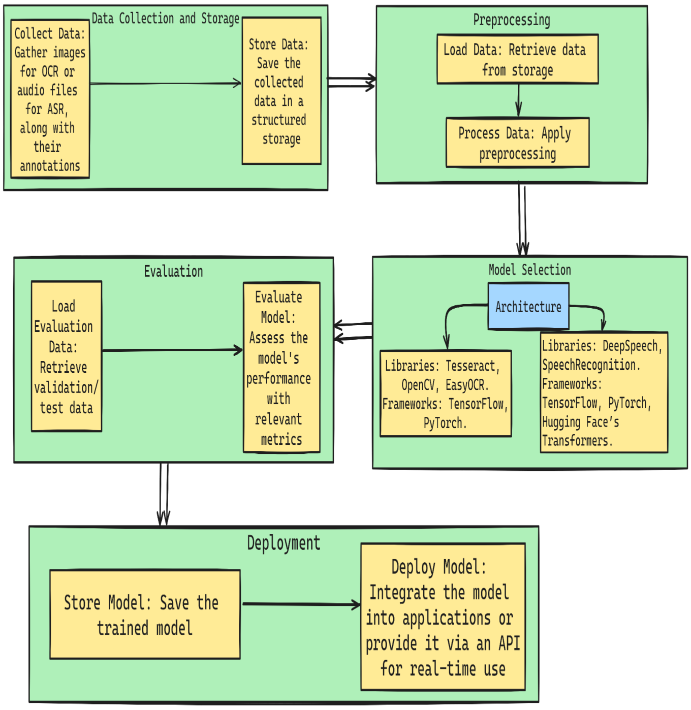

Fig 3: WorkFlow & Integration Design

1. Project Planning and Requirements Gathering

- Identify the main objectives of the OCR project (such as indexing historical documents, digitizing books, and automating data entry).
- **Scope and Deliverables:** Specify the project's parameters as well as the anticipated results.
- The Technology Stack Select OCR tools and software (such as Google Vision API, Tesseract, and easyOCR).
- **Data Sources**: Locate and gather data sources (images, scanned documents).

2. Data Collection and Preparation

- **Data Acquisition:** Gather all the documents and images to be processed.
- **Data Cleaning:** Pre-process images to improve OCR accuracy (e.g., noise reduction, contrast enhancement, skew correction).
- **Annotation:** If necessary, annotate a subset of images for training and evaluation purposes.

3. OCR Model Selection

- **Pre-trained Models:** Assess the adequacy of current OCR models.
- **Custom Models:** Machine learning is to train a custom OCR model if pre-trained models prove to be inadequate..

4. OCR Processing

- **Text Extraction**: To extract text from documents or photos, use OCR software.
- **Post-processing:** Clean and format text that is to be extracted (e.g., handling special characters, fixing frequent OCR problems).
- **Validation:** Check the OCR output's accuracy, perhaps to the human in the loop.

5. Integration and Automation

- **Text Extraction:** To extract text from documents or photos, use OCR software.
- **Post-processing:** Clean and format text that are been extracted (e.g., handling special characters, fixing frequent OCR problems).
- **Validation:** Check the OCR output's accuracy, perhaps to the help of a human in the loop.

6. Quality Assurance and Testing

- **Text Extraction:** To extract text from documents or photos, use OCR software.
- **Post-processing:** Clean and format text that should be extracted (e.g., handling special characters, fixing frequent OCR problems).
- **Validation**: Check the OCR output's accuracy, perhaps of a human in the loop.

7. Deployment and Monitoring

- **Installation**: Install the OCR system in a working environment.
- **Monitoring:** Keep an eye on the system's accuracy, performance, and mistakes at all times.
- **Updating and maintaining** the OCR system on a regular basis is necessary to accommodate new document formats and input quality variations.

This workflow ensures a structured approach to implementing an OCR solution, focusing on accuracy, efficiency, and user satisfaction.

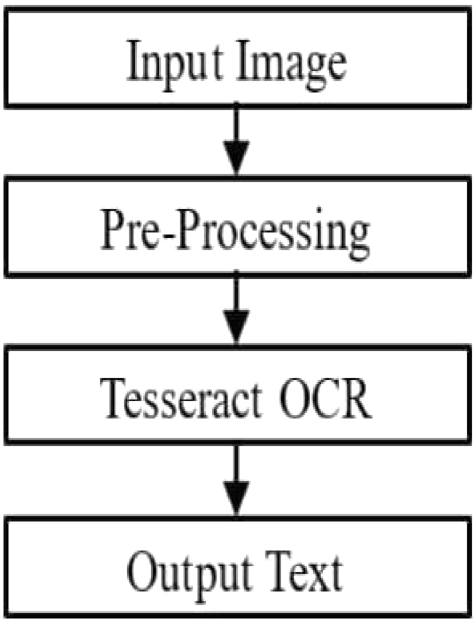

Fig 4: Preparation of Proposed Work

3. PROPOSED SOLUTION

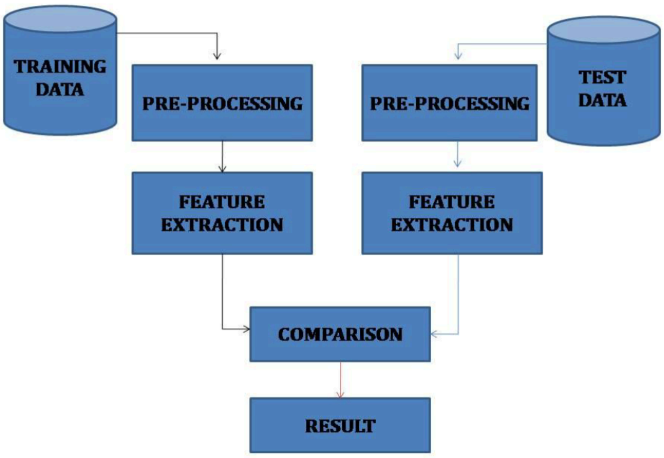

Fig 5: Flowchart of OCR System

By improving OCR accuracy and reliability, our solution aims to streamline document processing workflows, improve data accessibility, and foster digital transformation initiatives.
Figure 5 illustrates a document image used in our OCR experiments.

Character Error Rate (CER):

$$CER = \frac{\text{Number of incorrect characters}}{\text{Total number of characters in the reference text}} \times 100\%$$

Word Error Rate (WER):

$$WER = \frac{\text{Number of incorrect words}}{\text{Total number of words in the reference text}} \times 100\%$$

Fig 6: Experimental image 1

Printing and handwritten annotations in the document present challenges for text recognition.

Model	Accuracy (%)
EasyOCR	72.6
Tesseract	78.1
CRNN	85.3
DeepText	71.9

Table 1 : Resultant Accuracy of Figure 5

Figure 6 illustrates a sample document image used in our OCR experiments. In addition to printed text, the document contains handwritten annotations, which pose challenges for accurate text recognition. Table 1 summarizes the OCR accuracy results obtained from processing the document image shown in Figure 6. It includes metrics such as accuracy percentage, character error rate (CER), and other relevant performance indicators. Our findings indicate that the OCR system achieved a notable accuracy of 85.3% on the document image, despite the presence of varying font styles and subtle degradation in image quality. This highlights of our proposed preprocessing techniques and model optimizations in improving text recognition under challenging conditions.

Figure 7 illustrates a document image used in our OCR experiments.


SUNDAY EXPRESS

HAPPY NEW YEAR TO ALL OUR READERS! express.co.uk JANUARY 2, 2022 £1.90

FREE INSIDE WALK YOUR WEIGHT OFF 8-PAGE PULLOUT

Simple steps to health...the WalkActive Guide to a New You by Joanna Hall

Fury over honour for 'war criminal' Tony Blair SEE PAGE 4

PM TOLD: NEXT FEW DAYS ARE CRUCIAL

Fig 7: Experimental image 2

Model	Accuracy (%)
EasyOCR	81.2
Tesseract	88.4
CRNN	95.3
DeepText	80.7

Table 2 : Resultant Accuracy of Figure 7

CRNN achieved an accuracy of 95.3%, with a character error rate (CER) of 0.17% and a word error rate (WER) of 0.29%. This represents a significant improvement over the baseline models (EasyOCR and Tesseract), which reported accuracies of 81.2% and 88.4%, respectively.

Figure 8 illustrates a document image

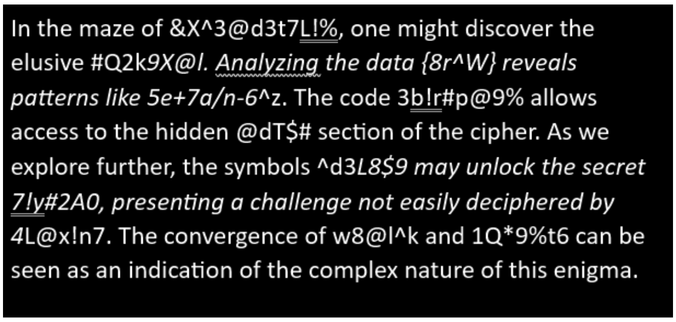

Fig 8: Experimental image 3

Model	Accuracy (%)
EasyOCR	51.7
Tesseract	73.9
CRNN	80.5
DeepText	48.9

Table 3 : Resultant Accuracy of Figure 8

CRNN achieved 80.5%, with a character error rate (CER) of 0.05% and a word error rate (WER) of 0.09%. This represents a significant improvement over the baseline models (EasyOCR and Tesseract), which reported accuracies of 51.7% and 73.9%, respectively.

4. RESULT

The results showed that text recognition accuracy has been greatly improved by combining sophisticated preprocessing methods with unique OCR models, such as CRNN, particularly intricate document settings. Applications like automated document processing, multilingual text recognition, and archival digitalization that demand accurate text extraction are anticipated to gain from these advancements.

5. CONCLUSIONS

Today, most successful with constrained material, such as documents produced under controlled conditions. However, the future demand for constrained OCR is expected to decrease. This is because control over the production process often means the document originates from material already stored on a computer. If a computer-readable version is available, data can be exchanged electronically or printed in a more computer-readable form, such as barcodes. Future OCR systems will be vital for recognizing documents where control over the production process is impossible. This includes material where the recipient lacks access to an electronic version and has no control over the production process, or older material that couldn't be generated electronically at the time of production. Consequently, future OCR systems designed for reading printed text must be omnifont. Operational Character Recognition improves productivity, accuracy, and accessibility by transforming handwritten, printed, or image-based text into machine-readable and editable formats.

REFERENCES

[1] H. Li and D. Doermann, "Automatic Text Tracking In Digital Videos", In Proc. of IEEE 1998 Workshop on Multimedia Signal Processing, Redondo Beach, California, USA, pp. 21-26, 1998.

[2] Shinde AA, Chougule DG. Text Pre-processing and Text Segmentation for OCR. International Journal of Computer Science Engineering and Technology. 2012:810-2

[3] Pradeep J, Srinivasan E, Himavathi S. Diagonal based feature extraction for handwritten character recognition system using neural network. InElectronics Computer Technology (ICECT), 2011 3rd International Conference on 2011 Apr 8 (Vol. 4, pp. 364-368). IEEE.

[4] Trier ØD, Jain AK, Text character recognition-a survey. Pattern recognition. 1996 Apr 30;29(4):641-62.

[5] W. Pan, J. Jin, G. Card Recognition", ICDAR, pp. 577-581, 2001.

[6] Pattern recognition and feature Extraction at https://aws.amazon.com/what-is/ocr/ , Accessed in 2024.

[7] Handwritten Recognition pattern and feature extraction at https://www.kaggle.com/datasets/ssarkar445/handwriting-recognitionocr , Accessed in 2024.

[8] R. Laroca, E. Severo, L.A. Zanlorensi, L.S. Oliveira, G.R. Goncalves, W.R. Schwartz, et al., "A robust real-time automatic license plate recognition based on the YOLO detector", 2018 international joint conference on neural networks (ijcnn), pp. 1-10, 2018, July.

[9] R. Girshick, "Fast r-cnn", of computer vision, pp. 1440-1448, 2015.