IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

OPTICAL CHARACTER RECOGNITION FOR CONVERT TEXT

[1]SONAL PRAJAPATI ,[2]SHUKLA HIMADRI ,[3]RADHIKA GOKANI
[1]STUDENT ,[2]ASSISTANT PROFESSOR ,[3]ASSISTANT PROFESSOR
[1]COMPUTER ENGINEERING ,
[1]SWAMINARAYAN UNIVERSITY , KALOL, GANDHINAGAR, GUJARAT,INDIA

Abstract : Optical Character Recognition (OCR) is a transformative technology that converts printed, handwritten, or scanned text into machine-readable formats. By leveraging advanced algorithms, OCR systems recognize characters, symbols, and structures in images, effectively digitizing content. This paper presents an overview of feature extraction methods for character recognition. Feature extraction method selection is the only most important factor in achieving high recognition performance in character recognition systems.

OCR has applications across industries, including document digitization, automated data entry, and accessibility tools for visually impaired individuals. It is widely used in banking for check processing, retail for receipt scanning, and government for archiving historical documents. Advanced OCR systems also handle multi-lingual text, cursive handwriting, and non-standard fonts with high precision.

INTRODUCTION

Optical character recognition (OCR) is the process of classifying optical patterns contained in a digital image corresponding to alphanumeric or other characters [1]. By analyzing and interpreting the visual patterns of characters in an image, OCR systems can transform static text into digital data that can be processed, searched, and stored.

OCR is a technique of translating handwritten, typewritten, or printed text characters to machine-encoded text [2]. It has many applications such as‒classification and analysis of RADAR signaling, character (letter or number) recognition, and handwriting analysis (_notepad‘ computers)‖. Other applications include bank checks, tablet computers, personal digital assistants (PDAs),Cheque reading, postcode recognition, form processing, and signature verification [3]

The applications of OCR are vast and span multiple industries. It is widely used for automating data entry, digitizing historical archives, enabling accessibility for visually impaired individuals, and facilitating document management in sectors such as banking, healthcare, retail, and education.OCR also plays a crucial role in emerging technologies like automated language translation and real-time text extraction for augmented reality (AR) systems. Optical character recognition has many different practical applications. The main areas where OCR has been of importance are text entry (office automation), data entry (banking environment), and process automation (mail sorting) [4].

Despite its advancements, OCR faces challenges such as low-quality input images, variations in handwriting, and contextual interpretation of text. Continuous innovation in artificial intelligence and natural language processing aims to address these limitations, ensuring OCR remains a cornerstone technology in the era of digital transformation.

Why is OCR important?

Most business workflows involve receiving information from print media. Paper forms, invoices, scanned legal documents, and printed contracts are all part of business processes. These large volumes of paperwork take a lot of time and space to store and manage. Though paperless document management is the way to go, scanning the document into an image creates challenges. The process requires manual intervention and can be tedious and slow.

Moreover, digitizing this document content creates image files with the text hidden within it. Text in images cannot be processed by word processing software in the same way as text documents. OCR technology solves the problem by converting text images into text data that can be analyzed by other business software. You can then use the data to conduct analytics, streamline operations, automate processes, and improve productivity.

How does OCR work?

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

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.

Preprocessing

The OCR software first cleans the image and removes errors to prepare it for reading. These are some of its cleaning techniques:

Deskewing or tilting the scanned document slightly to fix alignment issues during the scan.

Despeckling 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

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.

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.

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.

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.

What is OCR used for?

The following are some common OCR use cases in various industries:

Banking

The banking industry uses OCR to process and verify paperwork for loan documents, deposit checks, and other financial transactions. This verification has improved fraud prevention and enhanced transaction security. For example, BlueVine is a financial technology company that provides financing to small and medium-sized businesses. It used Amazon Textract, a cloud-based OCR service, to develop a product for small businesses in the US to quickly access Paycheck Protection Program (PPP) loans as part of the COVID-19 relief stimulus package. Amazon Textract automatically processed and analyzed tens of thousands of PPP forms per day so that BlueVine could help several thousand businesses get funds, saving over 400,000 jobs in the process.

Healthcare

The healthcare industry uses OCR to process patient records, including treatments, tests, hospital records, and insurance payments. OCR helps to streamline workflow and reduce manual work at hospitals while keeping records up to date. For example, the nib Group provides health and medical insurance to over 1 million Australians and receives thousands of medical claims per day. Its customers can take photos of their medical invoice and submit them through the nib mobile app. Amazon Textract processes these images automatically so that the company can approve claims much faster.

Logistics

Logistics companies use OCR to track package labels, invoices, receipts, and other documents more efficiently. For example, the Foresight Group uses Amazon Textract to automate invoice processing in SAP. Manual entry of these business documents was time-consuming and error-prone because Foresight employees had to enter the data in multiple accounting systems. With Amazon Textract, Foresight software can read characters more accurately across many different layouts, which increases business efficiency.

RESEARCH METHODOLOGY

OCR systems follow a structured pipeline to process and recognize text from images or scanned documents. This process involves multiple methodologies, each addressing a specific stage of text recognition. Below are the key methodologies employed in OCR systems :

1. Image Preprocessing

Before recognition, the input image is prepared to improve its quality and suitability for text extraction. Preprocessing ensures that the OCR system works with a cleaner and more uniform input. Common techniques include:

Skew correction: Aligning tilted images by detecting and correcting orientation.

Resizing and Scaling: Normalizing image dimensions for consistent processing.

Edge Detection: Highlighting the boundaries of text regions.

2. **Text Detection and Localization**

This step involves identifying and isolating regions of the image that contain text. Techniques include:

Connected component analysis: Groups pixels to identify text regions.

Region-Based Methods: Detect text blocks using bounding boxes or contours.

Deep Learning Models: Convolutional Neural Networks (CNNs) and region-based frameworks (e.g., Faster R-CNN) are used to detect and extract text areas more accurately.

3. **Segmentation**

Segmentation divides the identified text regions into smaller components, such as lines, words, and individual characters. Methods include:

Line Segmentation: Separates lines of text using horizontal projections.

Word Segmentation: Splits lines into words using spacing thresholds.

Character Segmentation: Breaks words into individual characters using connected components or contour analysis.

4. **Feature Extraction**

In this step, the system extracts meaningful features from segmented characters to differentiate them. These features can include:

Structural Features: Strokes, curves, loops, intersections, and endpoints.

Statistical Features: Histograms, pixel intensities, and texture patterns.

Gradient Features: Edge direction and magnitude for better character distinction.

5. **Classification and Recognition**

The extracted features are analyzed to classify each character. This step involves matching the features to predefined patterns or learning models:

Template Matching: Compares characters to stored templates.

Machine Learning Models: Uses algorithms like Support Vector Machines (SVMs) or Random Forests for recognition.

Deep Learning Models: Employs advanced architectures like CNNs, Long Short-Term Memory (LSTM), or Transformer-based networks for accurate recognition.

6. **Post-Processing**

Post-processing enhances the output text's accuracy and readability. This involves:

Spell Checking and Correction: Fixing recognition errors based on dictionary lookups or context.

Natural Language Processing (NLP): Understanding context to refine ambiguous characters or words.

Formatting: Retaining document structure, such as bold text, italics, or tables.

7. **Multilingual and Handwriting Recognition**

Modern OCR systems are designed to handle multiple languages and handwriting. Techniques include:

Language Models: Leveraging linguistic rules and context to improve accuracy.

Handwriting Analysis: Combining segmentation-free methods (like HMMs) with neural networks to recognize cursive or freehand text.

8. **Performance Optimization**

To ensure real-time performance and scalability, OCR systems integrate:

GPU Acceleration: Speeds up deep learning-based models.

Parallel Processing: Distributes tasks across multiple cores or systems.

Compression Techniques: Reduces the size of processed data for faster handling.

By combining these methodologies, OCR systems effectively transform printed or handwritten text into digital formats, ensuring high accuracy and adaptability to diverse applications. Continuous advancements in machine learning and artificial intelligence are further refining these methodologies for improved results.

RECOGNITION AND CLASSIFICATION TECHNIQUES

An application of neural networks in optical character recognition (OCR) is presented. The concept of learning in neural networks is utilized to a large extent in developing an OCR system to recognize characters of various fonts and sizes and handwritten characters. Parallel computational capability helps reduce recognition time which is crucial in a commercial context. The sensitivity of the network is such that small variations in the input do not affect the output and this results in an improvement in the recognition rate of characters with slight variations in structure, linearity, and orientation.

CONCLUSION

In this paper, we have discussed a survey of feature extraction and classification techniques for optical character recognition. A lot of research has been done in this field. Still, the work is going on to improve the accuracy of feature extraction and classification techniques.

This study emphasizes that OCR technology continues to play a pivotal role in various domains, including digitizing historical archives, automating business processes, and enabling accessibility for visually impaired individuals. However, there is still room for improvement, particularly in low-resource languages and real-time document processing.

Future research should focus on developing more robust OCR systems that can adapt to complex scenarios, including multilingual documents and dynamic environments. Furthermore, advancements in hardware acceleration and cloud-based processing could make OCR solutions faster and more accessible globally.

References

1. Bunke, H., Wang, P. S. P. (Editors), Handbook of Character Recognition and Document Image Analysis, World Scientific, 1997.

2. Rice, S. V., Nagy, G., Nartker, T. A., Optical Character Recognition: An Illustrated Guide to the Frontier, The Springer International Series in Engineering and Computer Science, Springer US, 1999.

3. R.J.SCHALKOFF, Artificial Neural Networks, The McGraw-Hill Companies Inc., New York, 1997.

4. Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin, — " A Survey of OCR Applications"International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012

5. Jagruti Chandarana, Mayank Kapadia, - "Review on Optical Character Recognition" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December - 2013