A Survey on Optical Character Recognition System

Gopika.D[1], Naasif.M[2], Vanmathi.J[3], Deepika.P[4]
[1]*Assistant Professor, Department of Computer Science, Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore*
[2,3,4]*II M.Sc Computer Science, Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore*

Abstract: ***Optical Character Recognition (OCR) has remained a significant area of academic and industrial interest for several decades. It is fundamentally defined as the technological process of converting document images into a format of individual characters that a machine can process. While extensive research has been conducted, achieving human-level accuracy in character recognition remains a persistent challenge in the field. This complexity has led to a surge in the number of academic laboratories and corporate entities dedicated to refining these systems. The primary objective of this research is to synthesize current developments within the OCR landscape. This paper provides an extensive overview of OCR methodologies and examines various technical proposals designed to mitigate existing system limitations.***

Keywords: Optical character Recognition (OCR), Document Image Processing, Character Recognition, Text Extraction

I.INTRODUCTION

Optical Character Recognition (OCR) serves as a critical software bridge that transforms physical text or images into digital data, allowing machines to manipulate and store the information. A fundamental hurdle in this field is that machines lack the innate human capability to instantly perceive and interpret visual characters. Consequently, researchers have focused on developing algorithms that can translate document images into a machine-readable syntax.

The OCR process is inherently difficult due to the vast diversity of global languages, varying font styles, and the intricate grammatical rules governing different scripts. To address these hurdles, the field integrates techniques from image processing, pattern classification, and natural language processing.

This article intends to guide the reader through the historical evolution, diverse applications, technical challenges, and the core phases involved in modern OCR development.

II.LITERATURE REVIEW

The history of character recognition precedes the digital age, with its beginnings found in mechanical systems developed before the modern computer. These initial OCR attempts relied on mechanical hardware to identify symbols, though they were limited by significant latency and minimal accuracy. A notable milestone occurred in 1951 when M. Sheppard developed GISMO, an early robotic reader. While GISMO could interpret musical symbols and printed words, its recognition capability was restricted to just 23 characters. Shortly after, in 1954, J. Rainbow introduced a device capable of identifying uppercase English letters at a rate of one character per minute. During the 1960s and 1970s, research momentum slowed as critics pointed to high error rates and sluggish performance. Development during this era was largely confined to high-budget organizations like banks and government departments. To improve reliability, the industry introduced standardized fonts, specifically OCRA and OCRB, which were established by ANSI and EMCA in 1970

III.TYPES OF OPTICAL CHARACTER RECOGNITION SYSTEMS

OCR research has branched into various specialized domains based on the mode of data acquisition, character connection styles, and font constraints.

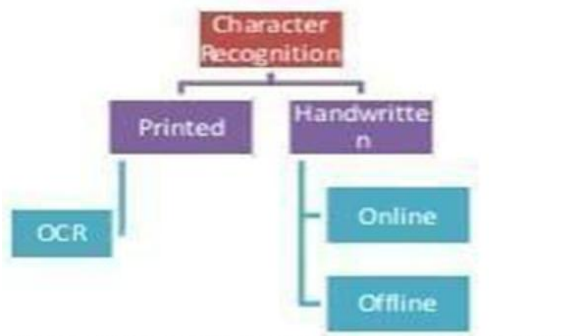

Figure.1: Types of character recognition system

As shown in the classification above, systems are primarily categorized by the nature of the input.

Machine Printed Character Recognition: This is considered a less complex task because the characters typically have standardized dimensions and predictable positioning on the page. Handwriting Character Recognition: This presents a much higher level of difficulty due to the vast differences in individual writing styles and varying pen strokes for the same character. On-line Systems: These systems capture data in real-time as the user writes. They are generally easier to implement because they record temporal data, such as stroke speed, velocity, and direction, which simplifies the recognition process. Off-line Systems: These operate on static, pre-recorded images or bitmaps

IV.APPLICATIONS OF OCR

Optical Character Recognition facilitates a wide array of practical utilities across various sectors. In its early implementation, the technology was primarily utilized for sorting mail, processing bank cheques, and verifying signatures. Modern organizations now leverage OCR for automated form processing to efficiently manage massive datasets that exist in printed material. Other critical applications includes Government and Security, Finance, Accessibility.

V.MAJOR PHASES OF OCR

The OCR workflow is an integrated process consisting of several sequential stages designed to transform an image into machine-readable text.

Image acquisition:

The initial phase involves obtaining a digital representation of a document through external hardware such as scanners or digital cameras. This stage includes digitization, quantization, and potentially image compression to facilitate easier processing. Binarization is a specialized form of quantization that reduces the image to just two levels—black and white—which is often sufficient for character identification.

Pre processing:

Following acquisition, pre-processing techniques are applied to refine image quality and remove artifacts that might hinder recognition.

Thresholding:

This involves binarizing the image based on a specific value, which can be determined at a global or local level.

Character segmentation:
In this stage, the document image is divided into its individual character components before being analyzed by the classification engine. Segmentation can be performed explicitly as a standalone step or implicitly as a byproduct of the classification process. Common techniques include projection profiles and connected component analysis; however, advanced methods are required when characters are overlapping, broken, or obscured by noise.

Feature extraction:
This phase involves identifying and extracting unique characteristics that distinguish one character from another. The selection of appropriate features is a vital research challenge.

Geometrical Features:
These include physical attributes such as loops, strokes, and corner points.

VI.CONCLUSION

This paper has presented a comprehensive overview of the various methodologies and techniques currently utilized in the field of Optical Character Recognition (OCR). It is established that OCR is not a single, atomic operation but rather a composite workflow

consisting of critical phases: acquisition, pre-processing, segmentation, feature extraction, classification, and post-processing. A detailed examination of each stage reveals that the successful integration of these techniques is essential for developing high-performance recognition systems. Such systems find vital utility in diverse real-time applications, including smart library management and automated vehicle number-plate recognition.

REFERENCES

[1] Satti, D.A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan. p. 141.

[2] Mahmoud, S.A., & Al-Badr, B., 1995, Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), 49-77.

[3] Bhansali, M., & Kumar, P, 2013, An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2(1), 211- 217.

[4] Qadri, M.T., & Asif, M, 2009, Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition presented at International Conference on Education Technology and Computer, Singapore, 2009. Singapore: IEEE.

[5] Shen, H., & Coughlan, J.M, 2012, Towards A Real Time System for Finding and Reading Signs for Visually Impaired Users. Computers Helping People with Special Needs. Linz, Austria: Springer International Publishing.

[6] Bhavani, S., & Thanushkodi, K, 2010, A Survey On Coding Algorithms In Medical Image Compression. International Journal on Computer Science and Engineering, 2(5), 1429-1434.

[7] Bhammar, M.B., & Mehta, K.A, 2012, Survey of various image compression techniques. International Journal on Darshan Institute of Engineering Research & Emerging Technologies, 1(1), 85-90.

[8] Lazaro, J., Martín, J.L, Arias, J., Astarloa, A., & Cuadrado, C, 2010, Neuro semantic thresholding using OCR software for high precision OCR applications. Image and Vision Computing, 28(4), 571-578.

[9] Lund, W.B., Kennard, D.J., & Ringger, E.K. (2013). Combining Multiple Thresholding Binarization Values to Improve OCR Output presented in Document Recognition and Retrieval XX Conference 2013, California, USA, 2013. USA: SPIE

[10] Shaikh, N.A., & Shaikh, Z.A, 2005, A generalized thinning algorithm for cursive and non-cursive language scripts presented in 9th International Multitopic Conference IEEE INMIC, Pakistan, 2005. Pakistan: IEEE

[11] Shaikh, N.A., Shaikh, Z.A., & Ali, G, 2008, Segmentation of Arabic text into characters for recognition presented in International Multi Topic Conference, IMTIC, Jamshoro, Pakistan, 2008. Pakistan: Springer