e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

COMPREHENSIVE SURVEY ON OPTICAL CHARACTER RECOGNITION (OCR)

Anshika Singh*1, Atharva Pardeshi*2, Avanti Thakare*3, Rucha Rajmane*4,
Prof. Arundhati A. Chandorkar*5

*1,2,3,4,5Department Of Computer Engineering, Pune Institute Of Computer Technology, Pune, India.

ABSTRACT

The verification and documentation of real estate transactions in the finance sector is a time-consuming and sus-ceptible to errors, traditionally involving the manual extraction of data from property documents to generate a report. The research within this paper looks into how OCR (Optical Char-acter Recognition) is applied to automate extraction, validation, and processing of key data such as property addresses for scanned documents complying with legal requirements. Data verification and intelligent error detection mechanisms through which artificial intelligence and machine learning techniques are integrated into the system ensure a certain level of accuracy and reliability in document generation.

The OCR-based proposed solution reduces processing times significantly, minimizes human error, and enhances overall pro-ductivity in the legal verification workflow. This approach will optimize the creation of compliantly legal documents while automating data extraction and validation for the efficient and accurate processing of real estate transactions in the finance sector.

Keywords: Binerization, Optical Character Recognition, Pattern Matching, Segmentation, Tesseract, Text Extraction.

I. INTRODUCTION

OCR facilitates the conversion of text contained within images, scanned documents, or even PDFs into readable machine data from printed or handwriting. While OCR was specifically designed for the purpose of assisting those who are blind or suffering from any form of visual impairment, upon its invention, this technology quickly became usable in multiple fields, and its first extension comprises academic research. OCR assists in the process of extracting texts without wasting much time in retrieving information, indexing, and later analysis that can be helpful in the digitization and preservation of academic literature. For researchers, OCR can unlock vast amounts of printed material in order to enhance searchability and the manipulation of textual content to great advantage in literature reviews and data collection.

II. RELATED WORK

Author[1] analyses the fundamentals of Optical Character Recognition (OCR). They focus on the main principles and diverse applications, as can be used in banking, health care, or jurisprudence. The authors concentrate how OCR technology is applied to convert text into a digital format for enhanced readability and efficiency. A paper further focuses on the typical issues with precision and usability dependency on font type and quality of the documents for OCR processing. The authors will present an all-inclusive overview, explaining to the practitioners and researchers the status of OCR technology and the future potential advancements.

Author[2] performs a thorough case study of the Tesseract OCR tool in terms of architecture, functionalities, and perfor-mance analysis. It was established that the use of a grayscale image improves recognition significantly more than color images; this is mainly because of low levels of noise. They suggest several pre-processing techniques, such as adaptive thresholding, noise removal, etc., to enhance the quality of input before the process so that the input data is as optimal as possible. They compare Tesseract with other OCR tools in different font types and layout conditions, making their outcome beneficial for developers to design OCR solutions properly.

Author[3] does an in-depth analysis of OCR systems dis-cussing some of the major issues widely encountered in recognizing complex languages like Arabic and Sindhi. They trace the development of OCR technology and identify areas where current systems fall short in achieving high accuracy with complicated scripts. The authors make the case for targeted research on recognition algorithms specific to these languages as well as noting that the much-vaunted potential of multilingual OCR could be more effectively exploited were an ability to easily switch between scripts available. Their study points out the critical aspect for the usability of OCR technology: its capability to be applied more by a greater population of linguistic communities.

Author [4] talks about various OCR methods and concen-trates on the specific problem of degradation in conversion from old books, poor scans, etc., into editable text. It describes how OCR was advanced and neural networks were trained to enhance recognition abilities to learn complex patterns present within degraded images. Character n-grams are stressed by Singh also, considering the contextual relationship for enhanc-ing accuracy in using characters. This paper deals with these challenges and recent innovations, giving valuable insights to efforts made in developing better OCR performances with regards to problems concerning poor-quality inputs.

Author[5] discussed many OCR algorithms and techniques. Much knowledge regarding the basic components which result in effectively working OCR systems is provided by them. Their evaluation relates to a vast number of techniques, both classical and recent methods developed with machine learning approaches for achieving even higher recognition rates. Some themes regarding the improvement of overall performance re-late to classification methods, preprocessing steps, and feature extraction. This work presents an invaluable overview of such algorithms for those researching or trying to apply or improve OCR technologies in various applications.

The author[6] provides an overview of the development history of OCR technology and points out the problems that would still prejudice it up to this date. They point out three critical phases of OCR: image acquisition, character recognition, and post-processing and state that each of these phases is essential for obtaining clear-cut results. Talking about the growing applications of OCR in different fields, the authors appeal for more research in countering the issues, including difference in fonts and poor quality documents. The work deals with the requirements for development that can enhance the reliability of OCR and usability in real-world environments.

Author [7] talks about the recent advances in OCR through a review of different recognition approaches and searching into the challenges pursued even after much advancement in handwriting character recognition. They represent the in-efficiencies of the prevailing systems, which usually fail to interpret various writing styles rightly and emphasize the ne-cessity of stronger algorithms that are capable of coping with such differences. The authors mention some areas of ongoing research with respect to enhancing the accuracy and robustness of OCR systems, especially for historical documents and low-quality inputs. Such work leads to the development of more potent and flexible OCR solutions as key areas for further exploration are identified.

III. CHALLENGES

Challenges in character recognition accuracy for OCR sys-tems include the impact of noise, geometric distortions, and scanning artifacts, which can significantly disrupt the feature extraction process. These factors often lead to misclassification of characters, particularly in complex legal documents where precision is essential. Such inaccuracies not only compromise the integrity of the extracted text but also pose risks for legal interpretation and compliance, highlighting the need for robust preprocessing techniques to improve input quality before recognition.

A. Character Recognition Accuracy

Character recognition accuracy in OCR systems is compro-mised by noise, geometric distortions, and scanning artifacts, which disrupt feature extraction and lead to misclassification. In legal documents, such inaccuracies can result in critical errors affecting legal interpretation and compliance, neces-sitating advanced preprocessing techniques to enhance input quality.

B. Handwritten Text Recognition

Handwritten text recognition poses a significant challenge for OCR systems because traditional algorithms are typically designed for printed text, lacking the robustness needed to ac-commodate the wide variability in handwriting styles. Factors such as inconsistent letter formation, slant, and spacing can lead to high error rates in character recognition. Additionally, the contextual nuances and ligatures often present in hand-written text complicate the segmentation and classification processes. As a result, achieving accurate recognition of hand-written annotations in legal documents requires specialized algorithms and training datasets that can effectively learn from diverse handwriting examples.

C. Input Document Quality Variability

Input document quality variability is a critical challenge for OCR systems, as the performance heavily relies on the clarity and fidelity of scanned images. Lower-quality scans may exhibit issues such as blurriness, skewness, and com-pression artifacts, which can obscure text and disrupt the uniformity of character shapes. These artifacts hinder the efficacy of feature extraction algorithms, leading to inaccurate segmentation and classification of characters. Consequently, ensuring high-quality input documents is essential for main-taining OCR accuracy, particularly in sensitive applications like legal document processing where precision is paramount.

D. Data Privacy and Compliance

Data privacy and compliance are paramount when process-ing sensitive legal information through OCR systems, as they must adhere to strict regulations like GDPR and HIPAA. Any unauthorized access or data breaches can result in severe legal consequences and damage to client trust, necessitating robust security measures, such as encryption and access controls, to protect confidential information.

E. Multilingual Recognition Challenges

Multilingual recognition challenges arise in OCR due to the need for distinct models and preprocessing techniques for dif-ferent languages and scripts. This complexity increases the risk of errors, as variations in character sets and writing systems can hinder accurate recognition. Effective multilingual OCR requires specialized training and adaptation to accommodate diverse linguistic features.

IV. APPLICATIONS OF OCR

Optical Character Recognition is widely applied in different fields:

A. Document Digitization and Archiving

OCR is widely used for digitizing printed documents, en-abling efficient archiving, searching, and retrieval of informa-tion. It has become essential in preserving historical documents and converting academic papers, books, and legal documents into editable and searchable formats.

B. Banking and Finance

In the financial sector, OCR automates the extraction of key information from checks, invoices, receipts, and other documents, improving workflow efficiency and reducing hu-man errors. OCR is also used for processing forms and KYC documents in banking.

C. Healthcare

OCR is used to digitize medical records, prescriptions, and billing information, facilitating quicker retrieval of patient data, reducing manual paperwork, and ensuring better patient care by minimizing errors in transcription.

D. Automated Data Entry

OCR streamlines data entry processes by extracting and pro-cessing text from forms, invoices, and applications, reducing time spent on manual data entry while improving accuracy.

E. License Plate Recognition

OCR is applied in intelligent transportation systems to recognize and process vehicle license plates for toll collection, traffic enforcement, and parking management systems.

F. Assistive Technology

OCR assists individuals with visual impairments by con-verting printed or handwritten text into speech or Braille, providing better accessibility to books, articles, and other textual materials.

G. Translation and Text Analysis

OCR systems are integrated with translation tools to trans-late text from images, such as signs, menus, or scanned docu-ments, into multiple languages, facilitating communication in multilingual environments.

These applications highlight OCR’s role in automating tasks, enhancing productivity, and improving accessibility in various sectors.

V. CONCLUSION

The survey provides a comprehensive examination of the advancements and enduring challenges in Optical Character Recognition (OCR) technology, particularly within the context of automated legal document generation. While significant strides have been made in improving recognition accuracy and processing efficiency, critical challenges such as character recognition fidelity, variability in input document quality, the intricacies of handwritten text recognition, data privacy com-pliance, and the complexities of multilingual support remain formidable barriers. Overcoming these obstacles necessitates a concerted effort toward the development of sophisticated, domain-specific models that leverage advanced machine learn-ing techniques, enhanced preprocessing strategies, and rig-orous security frameworks. By addressing these multifaceted challenges, we can significantly elevate the efficacy of OCR systems in legal applications, thereby fostering greater opera-tional efficiency, ensuring adherence to regulatory standards, and ultimately enhancing access to legal information in a dynamic and increasingly digital environment. This holistic approach not only underscores the potential of OCR technol-ogy but also highlights the imperative for ongoing research and collaboration in this vital domain.

VI. REFERENCES

[1] Ravina Mithe, Supriya Indalkar, Nilam Divekar on Optical Character Recognition, International Journal of Recent Technology and Engineer-ing (IJRTE), Volume-2, Issue-1, March 2013.

[2] Chirag Patel, Atul Patel, Dharmendra Patel on Optical Character Recog-nition by Open Source OCR Tool Tesseract: A Case Study, International Journal of Computer Applications (0975 – 8887), Vol. 55– No.10, October 2012.

[3] Noman Islam, Zeeshan Islam, Nazia Noor on A Survey on Optical Char-acter Recognition System, Journal of Information and Communication Technology-JICT Vol. 10 Issue. 2, December 2016.

[4] Sukhpreet Singh on Optical Character Recognition Techniques: A sur-vey, International Journal of Advanced Research in Computer Engineer-ing and Technology (IJARCET), Vol. 2, Issue 6, June 2013.

[5] Venkata Rao, Dr. A.S.C.S.Sastry, A.S.N.Chakravarthy, Kalyanchakravarthi P on Optical Character Recognition Techniques Algorithms, Journal of Theoretical and Applied Information Technology, 20th January 2016, Vol.83, No.2.

[6] Karez Abdulwahhab Hamad, Mehmet Kaya on A Detailed Analysis of Optical Character Recognition Technology, IJAMEC, 2016, 4(Special Issue), 244–249.

[7] Sushant Chandra, Saurav Sisodia, Preeti Gupta on Optical Character Recognition – A Review, International Research Journal of Engineering and Technology (IRJET), Vol.07, Issue: 04 — Apr 2020