Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 2, August 2023
Paper : 17

Advances and Challenges in Multilingual OCR for Indic Scripts: A Comprehensive Literature Review

Shaik Moinuddin Ahmed
Dept. of Computer Science and Information Technology
Maulana Azad National Urdu University
Hyderabad
moinuddinahmed@manuu.edu.in

Wahid Abdul
Dept. of Computer Science and Information Technology
Maulana Azad National Urdu University
Hyderabad, India
awahid@manuu.edu.in

***Abstract*— This literature review article examines a number of references from 2004 to 2022 to give a thorough overview of Indic Optical Character Recognition (OCR) technology. The review includes studies from numerous conferences and journals on subjects like script identification, error detection and correction, benchmarking, post-processing methods, feature selection, clustering-based classification, multilingual script recognition, and deep learning techniques. The studies under examination stress the value of Indic OCRs in raising the accuracy of text recognition for languages including Devanagari, Telugu, Malayalam, Urdu, and Odia. The difficulties with Indic OCRs, such as handling multilingual scripts and poor script identification, are also explored. The review also emphasises how wordnets and sequence-to-sequence models are integrated with transfer learning, font diversity, data augmentation techniques, and other technologies in Indic OCR systems. The results direct future research towards more precise and reliable solutions by giving researchers and practitioners useful insights into the developments, methods, and difficulties in Indic OCRs.**

Keywords*— *Indic OCR (Optical Character Recognition), Multilingual, Indic scripts, Script recognition, Script identification, Deep learning

I. Introduction

A. Background and Significance of Indic OCRs

India is a linguistically diverse nation where a multitude of languages are spoken throughout its vast territory. Despite the fact that the Constitution recognises 22 official languages, countless others are spoken by various communities. The north and central regions are dominated by Hindi, the most widely spoken language, which is written using the Devanagari script. Bengali is flourishing in the east and is written using Bengali script. Telugu has its own script and flourishes in Andhra Pradesh and Telangana. Tamil, the language of Tamil Nadu, has its own script, while the language of Maharashtra, Marathi, uses the Devanagari script. Urdu, which is widely spoken in the north, uses the Perso-Arabic script, whereas Gujarati and Punjabi use the Gujarati and Gurmukhi scripts, respectively. Additionally, Kannada in Karnataka and Malayalam in Kerala have their own scripts. These languages, along with many more, showcase the rich tapestry of India's cultural and linguistic heritage, preserved through their unique scripts.

OCR technology plays an essential role in the digitization and analysis of textual content, allowing for efficient information retrieval and analysis. Due to their intricate designs, distinctive shapes, and large character sets, Indic scripts, which cover a variety of languages spoken on the Indian subcontinent, provide special difficulties for OCR systems. The development of robust and accurate OCR systems for Indic scripts is crucial for facilitating digital access to multilingual Indic documents, preserving cultural heritage, and supporting a variety of applications, including document analysis, information extraction, and machine translation.

The wide variety of linguistic qualities that are present in Indic scripts, such as conjunct characters, ligatures, diacritics, and contextual changes, contribute to the complexity of these writing systems. Likewise, the Indic script family consists of numerous languages, including Hindi, Bengali, Tamil, Telugu, Malayalam, Urdu, and many others, each of which has distinct characteristics and writing conventions. This includes script identification, character segmentation, the recognition of ligatures and conjuncts, and the management of variations in handwriting styles and document degradation.

In recent years, a significant amount of research has been devoted to overcoming these obstacles and developing efficient OCR systems for Indic scripts. Researchers have investigated a variety of techniques and methodologies, including machine learning algorithms, deep learning models, language-specific features, and linguistic knowledge, to address the unique characteristics of Indic scripts. There is still a need for additional research and development to achieve greater accuracy and robustness in recognising and analysing Indic script documents, despite the fact that these advancements have led to significant improvements in the performance of Indic OCR.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
Paper : 17

B. Scope and Objectives of the Literature Review

This literature review is intended to provide a comprehensive summary of the research conducted in the field of multilingual OCR for Indic scripts. Using a broad range of references spanning from early works in 2004 to recent advancements in 2023, this review intends to assess the progress made in Indic OCR technology, identify key challenges, and investigate potential solutions proposed by researchers. This literature review examines various aspects of multilingual optical character recognition (OCR) for Indic scripts, including script identification, character recognition, document-specific error detection and correction, post-processing techniques, benchmarking studies, and the use of machine learning and deep learning. The review will also highlight research on specific Indic languages, such as Telugu, Malayalam, Urdu, and Devanagari, to shed light on language-specific difficulties and developments. This review aims to contribute to the understanding of Indic OCR technology by synthesising the findings from a broad range of research articles, conference papers, and survey papers, identifying research gaps, and providing insights into future research directions. This review will serve as a valuable resource for researchers, practitioners, and developers working on OCR systems for Indic scripts, fostering further progress in this vital area of study.

II. Overview of Indic OCR Techniques

A. Traditional Methods for Indic OCR

Traditional approaches for Indic OCR (Optical Character Recognition) have paved the way for automated systems to recognise and extract text from Indic scripts [1], [2], and [7]. Typically, these methods combine preprocessing, feature extraction, and classification techniques. Among the most important techniques used in traditional Indic OCR are:

1) Preprocessing: Before performing OCR, preprocessing techniques are used to improve the quality of the input documents [1]. These techniques include noise elimination, skew correction, binarization, and text region segmentation [14].

2) Feature Extraction: The process of feature extraction involves extracting representative qualities from the input text [14]. For Indic scripts, stroke density, direction, and curvature are frequently used to distinguish between characters and increase recognition accuracy [10] [14].

3) Classification: The use of classification algorithms to assign labels to extracted features enables the recognition of specific Indic script characters [2] [10]. Character classification in Indic OCR systems has been accomplished using techniques such as template matching, neural networks, and statistical models[2][6].

B. Modern OCR Techniques and Their Application to Indic Scripts

Modern OCR techniques have emerged as a result of advancements in machine learning and deep learning, offering increased accuracy and versatility. Also applicable to Indic scripts are these techniques [18] and [20]. Some contemporary OCR techniques applicable to Indic scripts include:

1) CNNs: Convolutional Neural Networks: CNN-based architectures have demonstrated tremendous success in OCR tasks, including the recognition of Indic script [18]. By leveraging the hierarchical learning capabilities of CNNs, these models can automatically learn pertinent features from unprocessed pixel representations of characters, resulting in accurate and robust recognition [18] [20].

2) RNNs: Recurrent Neural Networks: RNNs, in particular Long Short-Term Memory (LSTM) networks, are well-suited for managing sequential data, which makes them effective for recognising Indic scripts with complex structural dependencies and contextual variations [19]. RNNs capture the contextual data required for accurate recognition[19].

*3) Transformer-Based Models:*Various natural language processing tasks have been revolutionised by transformer models, such as the well-known BERT (Bidirectional Encoder Representations from Transformers) architecture. These models can manage long-range dependencies and capture contextual information of characters, making them appropriate for Indic script recognition[18]tasks.

C. Comparison of Various OCR Methodologies for Recognising Indic Script

The discipline of Indic script recognition has witnessed the development of both traditional and modern OCR techniques. On the basis of their performance, accuracy, computational efficiency, and applicability to various Indic scripts, these methods have been evaluated and compared [3] [5] [15]. Among the essential elements for comparison are:

1) Recognition Accuracy: The capability of an OCR method to accurately recognise characters across a variety of Indic scripts is crucial [15]. Comparative recognition accuracy evaluations can shed light on the strengths and weaknesses of various approaches [4] [15].

2) Computational Performance: The computational efficacy of OCR techniques is crucial, especially for real-time or massive applications [5]. Techniques that provide swift

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
Paper : 17

processing times without compromising accuracy are preferred [5].

3) Compatibility with Indic Scripts: Indic scripts comprise a vast array of languages, each with its own peculiarities and complexities. The adaptability of OCR methodologies to various Indic scripts, including main scripts such as Devanagari, Tamil, Telugu, and Malayalam, should be assessed [4] [5].

4) Handling Variations and Obstacles: Indic scripts frequently exhibit variations in writing styles, ligatures, diacritics, and contextual dependencies, posing difficulties for OCR methods [6] [12]. Techniques that can manage these variations and challenges effectively are desirable.

By comparing and evaluating various OCR methods used for Indic script recognition, researchers and practitioners can make informed decisions regarding the selection of appropriate techniques for specific applications, taking into account the unique requirements and characteristics of Indic scripts.

Overall, the field of Indic OCR has transitioned from traditional approaches to contemporary techniques that employ machine learning and deep learning algorithms. These advancements have contributed to the development of robust OCR systems capable of efficiently processing multilingual Indic documents.

TABLE I. SUMMARY OF LITERATURE REVIEW ARTICLES ON MULTILINGUAL OCR FOR INDIC SCRIPTS

Reference	Year	Key Findings
[1]	2004	Tools for enabling digital access to multi-lingual Indic documents
[2]	2014	Towards a robust OCR system for Indic scripts
[3]	2016	Multilingual OCR for Indic scripts
[4]	2017	Benchmarking scene text recognition in Devanagari, Telugu, and Malayalam
[5]	2017	Empirical study of effectiveness of post-processing in Indic scripts
[6]	2017	Framework for document-specific error detection and corrections in Indic OCR
[7]	2017	Error detection and correction in Indic OCRs
[8]	2018	Reusing OCRs for post-OCR text correction in Romanized Sanskrit
[9]	2018	Automatic multi-lingual script recognition application
[10]	2018	Comprehensive handwritten Indic script recognition system
[12]	2019	Degraded script identification of Urdu and Devanagari document - a survey
[13]	2019	Challenges in text recognition of Indian languages
[14]	2019	Clustering-based feature selection framework for handwritten Indic script classification
[15]	2020	Benchmark system for Indian language text recognition
[16]	2020	Urdu handwritten text recognition - a survey
[17]	2020	Incorporating localized context in WordNet for Indic languages
[18]	2020	MuLTReNets: Multilingual text recognition networks for script identification and handwriting recognition
[19]	2020	Handwritten Indic script recognition based on the Dempster-Shafer theory of evidence
[20]	2020	Survey of mono- and multilingual character recognition using deep and shallow architectures
[21]	2021	Multi-script language identification
[22]	2021	Framework for pre-processing, recognizing, and distributed proofreading of Assamese printed text
[23]	2021	Translator for Indian sign boards to English using Tesseract and SEQ2SEQ model
[24]	2022	Systematic review on OCRs for Indic documents & manuscripts
[25]	2022	Enhancing CNN using data augmentation techniques for Odia handwritten character recognition
[26]	2022	One-shot approach for multilingual classification of Indic scripts
[27]	2022	Improving scene text recognition for Indian languages with transfer learning and font diversity

III. DIGITAL ACCESS TO MULTILINGUAL INDIC DOCUMENTS

A. Tools and methods for facilitating digital access to multilingual Indic documents

Due to the multiplicity of scripts and languages involved, accessing and digitising multilingual Indic documents poses significant challenges. However, researchers have developed a variety of tools and methods to overcome these barriers and enable digital access to such documents.

The study by [1] examines digital access tools for multilingual Indic documents. The authors propose techniques for recognising and extracting Indic scripts from document images, such as segmentation, feature extraction, and classification. They emphasise the significance of preprocessing techniques to improve the extracted text's quality.Similarly, [2] presents a robust OCR system for Indic scripts, concentrating on script-specific challenges such as complex character shapes and ligatures. The authors increase the accuracy of recognition by combining feature extraction and classification techniques.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
Paper : 17

B. Challenges and strategies for OCR in various Indic scripts

Optical Character Recognition (OCR) systems face a number of challenges as a result of the variety of Indic scripts. However, researchers have made significant progress in resolving these challenges through the development of innovative solutions.
The benchmark set by [4] for Devanagari, Telugu, and Malayalam scene text recognition. The study emphasises the challenges associated with scene text recognition in Indic scripts, such as font style, size, and orientation variations. The authors propose character-level recognition and post-processing techniques for increasing recognition accuracy.

In addition, [6] describes a framework for document-specific error detection and corrections in Indic OCR. The authors are concerned with detecting and rectifying errors unique to Indic scripts, such as split words, merged characters, and improper segmentation. By identifying and correcting these errors, their framework improves the quality of OCR output.

IV. Robust OCR systems for Indic scripts

A. Advances in the development of robust OCR systems for Indic scripts:

The need for reliable and accurate text extraction from a variety of documents has propelled the development of more robust OCR systems for Indic scripts. In [8], the authors propose repurposing OCRs for post-OCR Sanskrit text correction. Their method utilises the existing OCR output as a starting point and refines the recognised text using correction mechanisms. This methodology improves the precision and quality of OCR results.

B. Techniques for dealing with variances in font styles, sizes, and quality of Indian text:

The variation in font styles, sizes, and quality of Indic text poses significant difficulties for OCR systems. Researchers have devised techniques to handle these variations and improve the accuracy of recognition. For instance, [10] presents a tree-based approach to a comprehensive handwritten Indic script recognition system. To accurately classify and recognise Indic scripts, the authors concentrate on handling variations in character sizes, shapes, and writing styles. The document [15] proposes a benchmark system for Indian language text recognition, addressing the challenges posed by font design and quality variations. To handle these variations and enhance the overall performance of OCR systems, the authors incorporate deep learning techniques and character-level recognition.

V. SCRIPT IDENTIFICATION AND RECOGNITION

A. Methods for Identifying and Recognising Indic Scripts Automatically:

Automatic script recognition is necessary for effective multilingual OCR systems. Researchers have devised a variety of methods to identify and recognise Indic scripts automatically. A survey document on the identification of Telugu script using OCR is provided in [11]. The authors discuss stroke-based and contour-based approaches for identifying and recognising Telugu scripts. They compare the effectiveness of these methods and highlight their advantages and disadvantages.

MuLTReNets, a multilingual text recognition network capable of simultaneous script identification and handwriting recognition, is also proposed in [18]. The authors employ deep learning architectures to identify diverse Indic scripts with precision.

B. Comparative analysis of the performance of script identification techniques:

In [16], the authors undertake a survey on Urdu handwritten text recognition, which focuses heavily on script recognition. They provide a comparative analysis of the performance of various script identification techniques in the context of Urdu and Devanagari scripts.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
Paper : 17

VI.

POST-PROCESSING TECHNIQUES FOR INDIC OCR

A. Evaluation and efficiency of post-processing techniques to enhance OCR accuracy:

Post-processing techniques are essential for enhancing OCR accuracy by correcting recognition errors. Researchers have assessed and created efficient post-processing techniques for Indic OCR. An empirical study on the efficiency of post-processing in Indian scripts is provided in [5]. The authors evaluate the effect of various post-processing techniques, such as dictionary-based correction, context-based correction, and language modelling, on OCR precision. Their findings shed light on the efficacy of various post-processing techniques.

B. Methods for improving recognition outcomes through error detection and correction:

The authors of [7] concentrate on error detection and correction for Indic OCRs. They propose a method that combines pattern recognition and machine learning techniques to detect and rectify textual errors. Their framework detects and effectively corrects common OCR errors, such as substitutions, deletions, and insertions, resulting in enhanced recognition results.

VII.

DEEP LEARNING APPROACHES IN INDIC OCR

A. Application of deep learning architectures for indic OCR tasks

Deep learning architectures have demonstrated promise in various OCR tasks, including the recognition of Indic scripts. Researchers have investigated the application of Indic OCR to enhance performance.Based on the Dempster-Shafer theory of evidence, [19] details a handwritten Indic script recognition system. The authors achieve accurate recognition results by employing deep learning techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM). Their method demonstrates the effectiveness of deep learning for OCR tasks in the Indic language.

B. Comparison of deep learning-based methodologies and conventional techniques

In [20], the authors compare deep learning-based methods and conventional approaches for character recognition in Indic and non-Indic scripts. They evaluate the effectiveness of both shallow and deep learning architectures and discuss their respective benefits and limitations. Their research demonstrates the superior performance of deep learning methods for Indic OCR tasks.

This literature review provides an overview of various aspects of multilingual optical character recognition (OCR) for Indic scripts, including tools and techniques for enabling digital access, challenges and solutions for OCR, script identification and recognition methods, post-processing techniques, and the application of deep learning. These studies advance OCR systems for Indic scripts and pave the way for enhanced accessibility and comprehension of multilingual Indic documents.

VIII. Challenges and Limitations

A. Identification and Evaluation of Challenges

a) Diversity of Scripts and Languages: Indic languages exhibit a wide variety of scripts and languages [4]. Each script has distinctive character designs, ligatures, and diacritics, which makes script identification and character recognition challenging [5].

*b) Handwriting Variations:*Handwritten text presents additional difficulties for Indic OCR [10]. Handwriting styles and variations between individuals can contribute to character deformations, inconsistent stroke patterns, and overlapping strokes, resulting in a loss of accuracy *[25]*.

c) Low-Quality Documents: OCR systems address low-quality documents, such as scanned images, degraded prints, and historical manuscripts [4]. These documents may be affected by noise, blur, fading, and irregular illumination, which degrade the quality of input data and negatively impact recognition accuracy [16].

d) Insufficient Training Data: Training accurate OCR models requires the availability of annotated training data [9]. Nevertheless, creating exhaustive and varied datasets for Indic scripts is a laborious and time-consuming endeavour. Insufficient training data can hinder the performance and generalisation abilities of OCR systems [14].

e) Vocabulary Size and Out-of-Vocabulary Words: Typically, the inventories of Indic languages contain a vast number of words and variants [21]. For accurate recognition results, OCR systems must effectively manage out-of-vocabulary words and account for variations in spelling, morphology, and syntax [14].

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
Paper : 17

f) Difficulties with Segmentation: It can be difficult to accurately segment characters, sentences, and lines in Indic scripts [6]. Errors in character recognition and subsequent analysis can result from incorrect segmentation [15].

g) Complexity of Computation: Large character sets, complex linguistic structures, and computationally intensive algorithms make Indic OCR systems susceptible to computational complexity [18]. Real-time processing and the efficient management of large data volumes present formidable obstacles [27].

B. *Limitations of Existing Systems and Techniques*

a) Language-Specific Methodology: Many of the OCR methods currently in use are tailored to particular Indian languages or scripts [3]. These language-specific approaches may not generalise well to other Indic languages, thereby limiting their applicability [19].

b) Error and Accuracy Rates: High accuracy rates in Indic OCR remain difficult to achieve [5]. OCR systems frequently confront errors associated with character recognition, segmentation, and language-specific variations, which can negatively impact output quality [7].

c) Resource-Intensiveness: Some OCR techniques require considerable computational resources [18], making their deployment impractical on low-end devices or in environments with limited resources [15].

d) Lack of Standard Evaluation Datasets: The lack of standardised evaluation datasets for Indic OCR hinders comparison and benchmarking of various systems [14]. A lack of consistent evaluation metrics and data sets makes it difficult to assess and compare the performance of OCR techniques comprehensively [10].

e) Limited Post-Processing Methods: Post-processing, which includes error detection and correction, is essential for improving OCR accuracy [8]. However, there are limited post-processing techniques available for Indic OCR [5].

f) Lack of Language Resources: Typically, Indic languages lack extensive linguistic resources, such as dictionaries, corpora, and lexicons [20]. The lack of such resources can have an effect on the accuracy and language coverage of OCR models [21].

IX. Future Directions and Emerging Trends

A. *Future research directions and potential improvement areas :*

Deep Learning Methodologies: Deep learning has demonstrated remarkable success in a wide range of computer vision tasks, including optical character recognition. Future research may focus on developing and refining architectures for deep learning that are specifically designed for Indic scripts. This includes investigating novel network structures, leveraging attention mechanisms, and investigating techniques such as capsule networks for enhanced Indic character recognition.

End-to-end OCR Systems: Traditional OCR systems include text localization, text segmentation, and character recognition. Building end-to-end OCR systems that directly convert images to text can simplify the overall pipeline and improve performance. Taking into consideration the unique characteristics and challenges of Indic scripts, research efforts can be directed towards the design of end-to-end architectures adapted to these scripts.

Resource-Scarce Languages: Few OCR training resources are available for many Indic languages, resulting in reduced accuracy. Future research can concentrate on developing techniques that address the problem of resource scarcity. Transfer learning, domain adaptation, and unsupervised learning methods can be explored to enhance OCR performance in resource-scarce languages by leveraging knowledge from resource-rich languages or unsupervised data.

Low-Quality Document Management: Indic OCR systems frequently encounter difficulties when processing low-quality documents, such as those with degraded or handwritten text, noisy backgrounds, and inconsistent illumination. Future research may investigate techniques for addressing these challenges, such as image enhancement algorithms, data augmentation strategies, and robust feature extraction techniques tailored particularly for low-quality Indic documents.

Script Detection and Language Identification: Indic scripts comprise a variety of languages and scripts, making script recognition and language detection essential for precise OCR. Future research can concentrate on the creation of robust techniques for automatic script recognition and language detection in Indic OCR systems. This includes investigating approaches based on deep learning, pattern recognition, and linguistic characteristics to accurately identify the linguistic structure and language of an input document.

Scene Text Recognition: Enhancing scene text recognition for Indic languages is an additional promising area of research.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 2, August 2023
Paper : 17

This requires accurately detecting and recognising text in complex and unrestricted environments, such as images captured in the open, signboards, street signs, etc.

Data Augmentation and Synthesis: Indic OCR systems frequently face issues due to insufficiently annotated training data. Future research can investigate techniques for data augmentation and synthesis to overcome this limitation. This includes generating realistic synthetic datasets, employing domain adaptation methods to transfer knowledge across various Indic languages, and developing techniques for generating handwritten character samples for training.

B. Exploration of emerging technologies and methodologies:

Multimodal OCR: The incorporation of multiple modalities, including text, images, and audio, can improve OCR accuracy and the user experience. Combining OCR with speech recognition and computer vision techniques can aid in the development of multimodal OCR systems for Indic scripts. Combining text and image analysis can provide more comprehensive information for applications involving historical manuscripts.

Integration of Natural Language Processing (NLP): The incorporation of NLP techniques such as text normalisation, morphological analysis, part-of-speech labelling, and named entity recognition can improve Indic OCR. NLP can help increase the accuracy of OCR systems for Indic languages by accommodating the unique spelling, word form, and syntactic structures of Indic languages. Focus research on integrating OCR and NLP frameworks for Indic scripts.

Interactive OCR Systems: User feedback can be incorporated into interactive OCR systems to increase OCR accuracy and user satisfaction. To involve users in the OCR process, future research can investigate interactive approaches such as active learning, semi-supervised learning, and user-driven error correction. These methods can be used to iteratively refine OCR models and tailor them to specific user requirements and document domains.

Cross-Lingual OCR: The objective of cross-lingual OCR is to recognise and translate text in multiple languages. Future research can investigate methodologies for cross-lingual Optical Character Recognition (OCR) in Indic scripts, enabling users to extract information from multilingual documents and facilitating cross-language information retrieval and comprehension. This necessitates addressing challenges including script and language recognition, transliteration, and translation.

Transfer Learning: Transfer learning, specifically pre-training on large-scale datasets, has proved effective for a number of OCR tasks. Further investigation of transfer learning techniques for Indic OCR can facilitate the transmission of knowledge from resource-rich to low-resource Indic languages, thereby enhancing the recognition performance of these languages.

Active Learning: By selectively selecting informative samples for labelling, active learning strategies can reduce annotation effort. Exploring active learning techniques for Indic OCR can help optimise the annotation process, especially in scenarios with limited resources, resulting in more efficient OCR model training.

Post-Processing Techniques: OCR outputs can be made more accurate and readable by investigating and devising advanced post-processing techniques for Indic scripts. This may include language-specific error detection and correction methods based on grammar and syntax to refine the transcribed text.

Hardware acceleration: Utilising specialised hardware accelerators such as Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) in Indic OCR can significantly enhance computational efficiency and real-time performance.

Cloud-based OCR Services: Future research can concentrate on developing cloud-based OCR services for Indic scripts, as cloud computing resources become more accessible and affordable. Cloud-based OCR platforms can provide scalable and accessible solutions for the analysis of Indic documents, allowing for the efficient processing and analysis of large volumes of multilingual Indic documents.

Human-in-the-Loop OCR Systems: Integrating human-in-the-loop approaches, such as correction interfaces that are interactive and user feedback, can improve the accuracy and dependability of Indic OCR systems. Collaborative systems that combine the advantages of automated OCR and human intelligence can enhance accuracy and the user experience.

Conclusion

The field of multilingual OCR for Indic scripts has made significant advancements, as evident from the extensive research covered in the literature. with a focus on expanding language coverage beyond major languages. However, recognising handwritten Indic scripts remains a challenging task, requiring robust algorithms to accurately transcribe despite variations in writing styles. Standardised datasets for benchmarking and evaluation are beneficial. Enhanced scene text recognition for Indic languages is also promising, especially in complex and unconstrained environments. Research avenues include language coverage, handwritten script recognition, scene text recognition, end-to-end systems, deep learning architectures, transfer learning, data augmentation, multimodal approaches, active learning, post-processing techniques, and hardware acceleration. Exploring these areas and adopting innovative approaches can contribute to the development of more accurate, efficient, and accessible

OCR systems for Indic scripts, enabling digital access to multilingual Indic documents and promoting greater linguistic inclusivity. Researchers in the field have the potential to make significant contributions by addressing these areas and leveraging cutting-edge methods, ultimately benefiting the accessibility and inclusivity of Indic OCR systems.

References

[1] "Tools for enabling digital access to multi-lingual Indic documents," in *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, 2004, pp. 122–133. [Online]. Available: https://ieeexplore.ieee.org/document/1263244/

[2] "Towards a Robust OCR System for Indic Scripts," in *2014 11th IAPR International Workshop on Document Analysis Systems*, 2014, pp. 141–145. [Online]. Available: https://ieeexplore.ieee.org/document/6830986/

[3] B. Hari Kumar & P. Chitra, "Survey Paper of Script Identification of Telugu Language using OCR," *Int. J. Electron. Commun. Eng.*, vol. 8, no. 3, pp. 15–20, 2019, [Online]. Available: http://www.iaset.us/archives?jname=16_2&year=2019&submit=Search

[4] S. Habib, M. K. Shukla, and R. Kapoor, "Degraded Script Identification of Urdu and Devanagari Document-A Survey," *2019 4th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2019*, pp. 79–83, 2019, doi: 10.1109/ISCON47742.2019.9036305.

[5] G. S. Science, "Challenges in Text Recognitions of Indian Languages," vol. 9, no. 2, pp. 176–179, 2019.

[6] I. Chatterjee, M. Ghosh, P. K. Singh, R. Sarkar, and M. Nasipuri, "A clustering-based feature selection framework for handwritten Indic script classification," *Expert Syst.*, vol. 36, no. 6, pp. 1–17, 2019, doi: 10.1111/exsy.12459.

[7] K. Tulsyan, N. Srivastava, A. Mondal, and C. V Jawahar, "A Benchmark System for Indian Language Text Recognition," 2020, doi: 10.1007/978-3-030-57058-3_6.

[8] A. F. Ganai *et al.*, "Urdu handwritten text recognition: A survey ISSN 1751-9659," *IEEE Access*, vol. 8, no. 2, pp. 2291–2300, 2020, doi: 10.1049/iet-ipr.2019.0401.

[9] S. Mohapatra, S. Agnihotri, A. Garg, P. Shah, and S. Chakraverty, "Incorporating Localised Context in Wordnet for Indic Languages," *Proc. Lr. 2020 Work. Multimodal Wordnets*, no. May, pp. 7–13, 2020, [Online]. Available: https://aclanthology.org/2020.mmw-1.2

[10] Z. Chen, F. Yin, X. Y. Zhang, Q. Yang, and C. L. Liu, "MuLTReNets: Multilingual text recognition networks for simultaneous script identification and handwriting recognition," *Pattern Recognit.*, vol. 108, 2020, doi: 10.1016/j.patcog.2020.107555.

[11] A. Mukhopadhyay, P. K. Singh, R. Sarkar, and M. Nasipuri, "Handwritten Indic Script Recognition Based on the Dempster-Shafer Theory of Evidence," *J. Intell. Syst.*, vol. 29, no. 1, pp. 264–282, 2020, doi: 10.1515/jisys-2017-0431.

[12] S. Kaur, S. Bawa, and R. Kumar, "A survey of mono- and multi-lingual character recognition using deep and shallow architectures: indic and non-indic scripts," *Artif. Intell. Rev.*, vol. 53, no. 3, pp. 1813–1872, 2020, doi: 10.1007/s10462-019-09720-9.

[13] "Multilingual OCR for Indic Scripts," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, pp. 186–191. [Online]. Available: https://ieeexplore.ieee.org/document/7490115/

[14] E. Technology, S. Vijayarani, and A. Sakila, "Multi-Script Language Identification From," no. 01, pp. 1292–1304, 2021.

[15] "A Framework for Pre Processing, Recognizing and Distributed Proofreading of Assamese Printed Text," in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 2021, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/9633818/

[16] "A Translator for Indian Sign Boards to English using Tesseract and SEQ2SEQ Model," in *2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM)*, 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9841215/

[17] C. V. S. Deepthi and A. Seenu, "A Systematic Review on OCRs for Indic Documents & Manuscripts," *2022 Int. Conf. Data Sci. Agents Artif. Intell.*, vol. 01, pp. 1–4, 2022, doi: 10.1109/ICDSAAI55433.2022.10028802.

[18] M. Das, M. Panda, and S. Dash, "Enhancing the Power of CNN Using Data Augmentation Techniques for Odia Handwritten Character Recognition," *Adv. Multimed.*, vol. 2022, 2022, doi: 10.1155/2022/6180701.

[19] "One-Shot Approach for Multilingual Classification of Indic Scripts," in *2022 International Conference on Innovative Trends in Information Technology (ICITIIT)*, 2022, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9744238/

[20] S. Gunna, R. Saluja, and C. V. Jawahar, "Improving Scene Text Recognition for Indian Languages with Transfer Learning and Font Diversity," *J. Imaging*, vol. 8, no. 4, 2022, doi: 10.3390/jimaging8040086.

[21] "Benchmarking Scene Text Recognition in Devanagari, Telugu and Malayalam," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 07, pp. 42–46. [Online]. Available: https://ieeexplore.ieee.org/document/8270315/

[22] "An Empirical Study of Effectiveness of Post-Processing in Indic Scripts," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 07, pp. 32–36. [Online]. Available: https://ieeexplore.ieee.org/document/8270313/

[23] "A Framework for Document Specific Error Detection and Corrections in Indic OCR," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 04, pp. 25–30. [Online]. Available: https://ieeexplore.ieee.org/document/8270254/

[24] P. J. Narayanan *et al.*, "Error Detection and Correction in Indic OCRs," 2017, [Online]. Available: https://www.semanticscholar.org/paper/b5853951c4221908f5ab9772d5fa902a537afad5

[25] A. Krishna, B. P. Majumder, R. Bhat, and P. Goyal, "Upcycle Your OCR: Reusing OCRs for Post-OCR Text Correction in Romanised Sanskrit," 2018, doi: 10.18653/v1/K18-1034.

[26] W. A. K. Abu-Ain, S. N. H. S. Abdullah, K. Omar, and S. Z. Siti, "Automatic multi-lingual script recognition application," *GEMA Online J. Lang. Stud.*, vol. 18, no. 3, pp. 203–221, 2018, doi: 10.17576/gema-2018-1803-12.

[27] P. K. Singh, R. Sarkar, V. Bhateja, and M. Nasipuri, "A comprehensive handwritten Indic script recognition system: a tree-based approach," *J. Ambient Intell. Humaniz. Comput.*, vol. 0, no. 0, p. 0, 2018, doi: 10.1007/s12652-018-1052-4.