

DIGITAL TRANSFORMATION OF INDIAN ARCHIVES AND EPIGRAPHY: THE ROLE OF OCR AND HTR

*Lokesha Naik K.**

DOI: <https://doi.org/10.5281/zenodo.21208831>

ABSTRACT

Artificial Intelligence (AI) has emerged as a transformative force in historical research and archival studies, reshaping traditional methodologies of data analysis, interpretation, and preservation. The rapid digitization of historical sources such as manuscripts, inscriptions, administrative records, and visual materials has created vast datasets that require advanced computational tools for effective analysis. AI technologies, including machine learning, Optical Character Recognition (OCR), and Handwritten Text Recognition (HTR), enable historians and archivists to process, classify, and interpret these sources with greater accuracy and efficiency. In archival studies, AI facilitates automated cataloguing, metadata generation, enhanced retrieval systems, and digital preservation of fragile materials. Furthermore, AI supports interdisciplinary research by integrating historical inquiry with digital humanities and data science. However, the application of AI also raises critical challenges related to algorithmic bias, contextual misinterpretation, ethical concerns, and data authenticity. This study examines the role, potential, and limitations of Artificial Intelligence in historical research and archival practices, arguing that AI, when used responsibly, serves as a powerful complementary tool that enhances scholarly rigor while preserving human interpretative authority in understanding the past.

KEYWORDS: *OCR In Archival Studies, OCR In Indian Archives, OCR In Indian Epigraphy, HTR In Archival Studies, HTR In Indian Archives, HTR In Indian Epigraphy.*

Introduction

Artificial Intelligence (AI) has significantly transformed historical research and archival studies by enabling efficient processing, preservation, and analysis of vast historical data. Technologies such as OCR, HTR, NLP, and machine learning help digitize manuscripts, automate cataloguing, improve translation, and enhance accessibility of archives. AI supports historians and archivists in managing large-scale and multilingual sources while contributing to digital preservation and restoration. As part of the digital humanities, AI encourages interdisciplinary collaboration and modern research methods. However, concerns regarding ethics, bias, and authenticity remain important, making AI a supportive tool rather than a replacement for human expertise.

A. Optical Character Recognition (OCR) in Archival Studies

Optical Character Recognition (OCR) is a

digital technology that converts scanned images and historical documents into machine-readable text. In historical research, OCR plays a vital role in digitizing manuscripts, newspapers, records, and reports, making them searchable and suitable for large-scale analysis. It enables historians to trace ideas, compare texts, and conduct quantitative research efficiently.

Role of Optical Character Recognition (OCR) in Archival Studies

Optical Character Recognition (OCR) plays a central role in modern archival studies by transforming traditional archives into digitally accessible and searchable repositories. Archives contain vast collections of printed records such as government files, reports, registers, newspapers, gazetteers, and administrative correspondence. OCR enables these materials to be converted from static images into machine-readable text, thereby significantly enhancing archival management, access, and preservation.

* **Prof. Lokesha Naik K.**, Assistant Professor, Department of History, Government First Grade College, Davangere.

Digitization of Archival Records: One of the primary roles of OCR in archival studies is to support large-scale digitization. Scanned archival documents are processed through OCR to extract textual content, converting physical records into digital resources. This process helps archives preserve rare and fragile documents while making their content accessible without repeated physical handling.

Improved Access and Discoverability: OCR allows archives to provide full-text search capabilities. Researchers can search for specific names, dates, places, or keywords within millions of pages of archival material. This greatly reduces the time and effort required to locate relevant records and democratizes access to archival resources for scholars, students, and the general public.

Cataloguing and Metadata Creation: In archival management, accurate cataloguing is essential. OCR assists in the automatic extraction of key information from documents, such as titles, subjects, and administrative details. This supports metadata generation and improves archival classification systems. As a result, archivists can manage large collections more efficiently with reduced manual workload.

Preservation and Conservation: OCR contributes indirectly to the preservation of archival materials. By enabling digital access, it minimizes the physical use of fragile documents, reducing wear and tear. OCR also supports digital preservation strategies by creating searchable backups of archival content, ensuring long-term protection against loss, damage, or deterioration.

Optical Character Recognition has become an indispensable tool in archival studies. By improving access, efficiency, and preservation, OCR transforms archives into dynamic and user-friendly knowledge systems. However, its effectiveness depends on careful implementation, quality control, and continued involvement of archivists and scholars to ensure accuracy and contextual integrity.

Optical Character Recognition (OCR) in Indian Archives and Epigraphy

Optical Character Recognition (OCR) has assumed great importance in the Indian context due to the vast, multilingual, and historically rich nature of Indian archival and epigraphically sources. India possesses one of the largest collections of archival records and inscriptions in the world, spanning ancient, medieval, and modern periods and written in multiple scripts and languages. The application of OCR has significantly

contributed to the digitization, preservation, and scholarly analysis of these materials.

OCR in Indian Archives

Indian archives such as the National Archives of India (NAI), State Archives, District Record Rooms, and institutional archives contain enormous quantities of printed records, including colonial administrative files, gazetteers, census reports, court proceedings, legislative debates, newspapers, and official correspondence. OCR plays a crucial role in converting these printed records into searchable digital text. Through OCR-based digitization projects, Indian archives have enhanced public access and research efficiency. Scholars can now search for names, places, policies, and events across thousands of pages, which was previously impractical. OCR has also facilitated large-scale historical research in fields such as colonial governance, economic history, social reform movements, and nationalist politics. In recent years, OCR has been increasingly used in regional language archives, including Kannada, Tamil, Telugu, Marathi, Urdu, and Persian records. While OCR accuracy is relatively high for modern printed texts, historical documents often require customized OCR models and post-correction due to old fonts, degraded paper, and archaic spellings.

OCR in Indian Epigraphy

Epigraphy-the study of inscriptions engraved on stone, metal, and other durable materials-poses unique challenges and opportunities for OCR. Indian inscriptions are written in a wide variety of ancient scripts such as Brahmi, Kharosthi, Gupta, Nagari, Grantha, Nandinagari, Sharada, and Modi, as well as regional scripts like old Kannada and Tamil-Brahmi. OCR in epigraphy focuses primarily on image-based character recognition combined with AI-assisted pattern analysis. High-resolution photographs and 3D scans of inscriptions are processed to enhance legibility, remove surface noise, and identify characters. While traditional OCR struggles with irregular carving styles and weathered surfaces, AI-based approaches have shown promising results in assisting epigraphists. Inscriptions issued by dynasties such as the Mauryas, Satavahanas, Chalukyas, Rashtrakutas, Hoysalas, Cholas, and Vijayanagara rulers are increasingly being digitized and studied using computational tools. OCR-assisted transcription helps scholars compare inscriptions, identify formulaic expressions, royal titles, land-grant patterns, and administrative terminology.

OCR has significantly advanced the digitization and accessibility of Indian archival and epigraphical sources. In Indian archives, OCR

enables efficient management, retrieval, and large-scale analysis of historical records. In epigraphy, OCR-assisted tools support the documentation and study of ancient inscriptions, opening new possibilities for comparative and interdisciplinary research. Despite technical limitations, OCR—when combined with scholarly expertise—plays a transformative role in preserving and interpreting India's rich historical heritage.

B. Handwritten Text Recognition (HTR) in Archival Studies

Handwritten Text Recognition (HTR) is an AI-based technology that converts handwritten archival documents into searchable digital text. In archival studies, HTR improves accessibility, cataloguing, and preservation by enabling large-scale transcription of manuscripts, registers, letters, and historical records. It supports researchers through full-text search and reduces physical handling of fragile documents. HTR is especially valuable in multilingual archival contexts like India, though challenges such as varied handwriting and degraded materials remain. Despite technological advances, HTR complements rather than replaces archivists and historians, as human expertise is essential for accurate interpretation and contextual understanding of archival records.

Role of Handwritten Text Recognition (HTR) in Archival Studies

Handwritten Text Recognition (HTR) is an advanced Artificial Intelligence-based technology that enables the automatic recognition and transcription of handwritten text from scanned images or photographs into machine-readable digital text. Unlike Optical Character Recognition (OCR), which is primarily designed for printed materials, HTR is specifically developed to handle handwritten documents, making it especially significant for archival studies. Since a large proportion of archival records consist of handwritten manuscripts, letters, registers, diaries, land records, and administrative files, HTR has become a transformative tool in modern archival practice.

Digitization of Handwritten Archival Records: Archives preserve vast collections of handwritten documents dating from the pre-modern and colonial periods. Manual transcription of such materials is extremely time-consuming and requires specialized palaeographical expertise. HTR enables large-scale digitization by automatically converting handwritten archival records into searchable digital text. This significantly accelerates the digitization process and supports the long-term preservation of fragile

documents.

Enhanced Access and Searchability:

One of the most important contributions of HTR in archival studies is improved access. HTR allows archives to offer full-text search even for handwritten materials. Researchers can search for personal names, place names, dates, and keywords across thousands of handwritten pages, which was previously possible only through manual reading. This democratizes access to archival collections and expands research possibilities.

Support for Archival Cataloguing and Metadata Creation: HTR assists archivists in generating metadata by extracting key information from handwritten documents. Registers, correspondence files, and administrative records can be automatically analyzed to identify titles, subjects, authors, and chronological details. This reduces the burden of manual cataloguing and improves archival organization and retrieval systems.

Preservation and Conservation Benefits:

By enabling digital access to handwritten records, HTR reduces the physical handling of fragile archival materials, thereby supporting conservation efforts. Digital transcriptions created through HTR also serve as backups, protecting archival content from loss due to deterioration, disasters, or mishandling.

Research and Analytical Applications:

HTR-processed archival texts enable advanced research methods such as text mining, network analysis, and quantitative historical analysis. Historians can study administrative practices, social relations, legal systems, and everyday life by analyzing large bodies of handwritten archival texts. This has particular relevance for social history, microhistory, and institutional history.

Role in Multilingual and Regional Archives: In countries like India, archival records are written in multiple languages and scripts, including Persian, Modi, old Kannada, Telugu, Tamil, and Devanagari. HTR models can be trained for specific scripts and handwriting styles, making it possible to process regional and colonial records more effectively. Although accuracy varies, HTR has opened new possibilities for accessing non-standard and regional archival materials.

Challenges and Limitations: Despite its advantages, HTR faces significant challenges. Variations in handwriting styles, ink quality, page layout, and document condition affect accuracy. Historical handwriting, abbreviations, and obsolete vocabulary further complicate recognition. Therefore, HTR output requires

human verification, especially for scholarly and legal use.

Ethical and Scholarly Considerations:

Over-reliance on HTR may lead to misinterpretation if errors are not carefully corrected. In archival studies, accuracy and authenticity are critical. HTR should be viewed as an assistive technology that complements, rather than replaces, the expertise of archivists, historians, and paleographers.

Handwritten Text Recognition has significantly transformed archival studies by enabling efficient digitization, improved access, and advanced analysis of handwritten records. While technical and linguistic challenges remain, HTR-when combined with human expertise-enhances archival preservation and research. In the digital age, HTR represents a crucial bridge between traditional archival materials and modern historical scholarship.

Handwritten Text Recognition (HTR) in Indian Archives and Epigraphy

Handwritten Text Recognition (HTR) is an Artificial Intelligence-based technology that enables the automatic transcription of handwritten texts from scanned images and photographs into machine-readable digital text. In the Indian context, HTR has immense scholarly significance due to the predominance of handwritten archival records and epigraphical materials written in diverse scripts and languages across different historical periods. Indian archives and epigraphy together represent one of the richest yet most challenging domains for the application of HTR.

1. HTR in Indian Archives

Indian archival repositories such as the National Archives of India (NAI), State Archives, District Record Rooms, Temple Archives, and Private Manuscript Collections preserve enormous volumes of handwritten records. These include royal farmāns, revenue registers, judicial proceedings, correspondence, palm-leaf manuscripts, diaries, and colonial administrative files. Manual transcription of such records has traditionally required paleographical expertise and considerable time. HTR has significantly accelerated the digitization of these handwritten materials by enabling large-scale transcription. By converting handwritten records into searchable text, HTR enhances accessibility, discoverability, and research efficiency. Scholars can now search across vast archival collections for personal names, place names, land grants, taxation details, and administrative terms, thereby opening new possibilities for social, economic, and institutional history. In multilingual archival contexts-where

documents are written in Persian, Modi, Devanagari, Urdu, old Kannada, Telugu, and Tamil scripts-HTR models can be trained for specific scripts and writing styles. Although accuracy varies, especially for older documents, HTR has made previously under-utilized regional and vernacular archives more accessible to researchers.

2. HTR in Indian Epigraphy

Epigraphy, the study of inscriptions engraved on stone, copper plates, temple walls, and monuments, presents a distinct set of challenges for HTR. Indian inscriptions are often written in ancient and medieval scripts such as Brahmi, Gupta, Nagari, Grantha, Nandinagari, Sharada, Tamil-Brahmi, and old Kannada, with significant variation in letter forms and orthography. HTR in epigraphy primarily functions as an assistive tool rather than a fully automated solution. High-resolution photography, reflectance transformation imaging (RTI), and 3D scanning are combined with AI-based recognition techniques to enhance legibility and identify characters. HTR-assisted transcription helps epigraphists compare inscriptions, recognize recurring formulae, royal titles, donative patterns, and administrative terminology across regions and dynasties such as the Mauryas, Satavahanas, Chalukyas, Rashtrakutas, Hoysalas, Cholas, and Vijayanagara rulers. While complete automation is difficult due to weathering, irregular carving, and script evolution, HTR significantly reduces the initial workload and supports systematic documentation of inscriptions.

C. OCR vs HTR in Archives and Epigraphy

With the digitization of historical and archival materials, modern technology has introduced advanced tools to facilitate the transcription, preservation, and analysis of textual records. Among these tools, Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) play pivotal roles, especially in the fields of archival studies and epigraphy. While both technologies aim to convert analog documents into machine-readable formats, their applications, methodologies, and challenges differ significantly.

OCR is primarily designed for printed text. It identifies and converts clearly typed or printed characters into digital text, enabling automated indexing, full-text search, and metadata generation. In archival contexts, OCR is widely used for colonial administrative records, newspapers, gazetteers, and printed manuscripts, making large-scale digitization efficient. However, OCR's effectiveness declines when dealing with handwritten or irregularly formatted records, and it is largely unsuitable for ancient inscriptions with worn or

uneven surfaces.

HTR, on the other hand, is specifically developed to recognize and transcribe handwritten documents, including letters, diaries, palm-leaf manuscripts, revenue registers, and inscriptions. Using machine learning and AI-based algorithms, HTR can handle variations in handwriting, cursive scripts, and complex ligatures, making it indispensable for pre-modern and regional archival collections. In epigraphy, HTR assists in transcribing inscriptions written in diverse scripts and languages, such as Brahmi, Tamil-Brahmi, old Kannada, and Persian, thereby enabling systematic analysis of handwritten or carved texts.

Comparatively, OCR is faster and more accurate for printed and uniform texts, whereas HTR is more versatile for handwritten, variable, and historical scripts. In Indian archives and epigraphy, where materials are multilingual, handwritten, and often damaged or eroded, HTR has emerged as a critical technology, complementing OCR in the digitization and study of historical sources. Together, these technologies have transformed archival practices and epigraphical research by improving accessibility, discoverability, and long-term preservation of textual heritage, while still requiring expert validation to ensure historical accuracy.

Aspect	OCR	HTR
Primary Use	Printed text	Handwritten text
Archives	Books, gazetteers, newspapers, administrative files	Manuscripts, letters, diaries, registers
Epigraphy	Clear inscriptions only	Manuscripts, palm-leaf, temple inscriptions
Handling Damage	Poor	Moderate; AI-assisted reconstruction possible
Script Variability	Limited	Handles regional and variable scripts
Accuracy	High for printed text	Requires training; moderate to high for trained scripts
Role	Mechanical transcription	Transcription + interpretative aid

Conclusion

OCR and HTR are vital for digitizing and preserving archival and epigraphic materials. OCR efficiently processes printed texts, while HTR interprets handwritten and damaged historical records, manuscripts, and inscriptions. Together, they improve accessibility, research, and preservation of India's multilingual heritage.

Despite their advantages, both technologies require expert human validation to ensure historical accuracy and contextual integrity.

References:

1. Gaikwad, S., Kachhoria, R. & Yadav, G. (2025) - AI-Based OCR for Digitizing Ancient Indian Texts: Preserving Linguistic Heritage and Overcoming Script Challenges. *International Journal of Linguistics Applied Psychology and Technology*.
2. Sharma, V., Verma, R. & Saluja, R. (2025) - AnciDev: A Dataset for High-Accuracy Handwritten Text Recognition of Ancient Devanagari Manuscripts. In *Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)*, Mumbai, India. Presents a publicly available dataset for improving HTR on ancient manuscript collections, advancing digital preservation efforts.
3. Sivankalai, S. & Balachandran, S. (2025) - Transforming Physical Archives into Searchable Digital Libraries with Optical Character Recognition. *De Gruyter*. - Explores an integrated machine learning OCR pipeline for multilingual documents including complex agency preprocessing and contextual correction.
4. Griffith, J. & Others (2022) - Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*, 22, 367–392. - A systematic review of HTR adoption in archive digitization, assessing how AI aids full-text search and transcription of handwritten sources.
5. Agrawal, Y., Balasubramanian, S., Meena, R., Alam, R., et al. (2024) - Optical Character Recognition using Convolutional Neural Networks for Ashokan Brahmi Inscriptions. *arXiv preprint*. - Demonstrates application of CNN-based OCR for ancient epigraphical scripts such as Brahmi, indicating potential for inscription digitization.
6. Archaeological Survey of India (ASI) - Digital initiatives such as the Bharat Shared Repository of Inscriptions (Bharat SHRI) aim to digitize and provide searchable access to India's epigraphic records, including e-estampages and metadata.