AI for Cultural Heritage: Advances in the Digital Reconstruction of Art and Artifacts

M. Ali Khan

Department of Computer Science, University of Engineering and Technology Lahore, Pakistan.
Email: 2023mscs23@student.uet.edu.pk, ORCID https://orcid.org/0009-0000-7927-0684

Received on 02 Nov 2025 : Revised on 25, Nov, 2025 Accepted: 2025

Abstract - Indus Valley Civilization, the world's oldest known urban civilization, left a cryptic script behind which is yet to be deciphered, owing mainly to the fragmented remains and lack of bilingual samples. Deciphering this old script has remained the problem of archaeologists and linguists for years. Recent progress in artificial intelligence, particularly computer vision and machine learning, has been instrumental in the deciphering of these complex symbols and inscriptions during recent years. This paper offers a technical overview of deep learning and image processing methods used in Indus seal and other epigraphic or grapheme-bearing objects studies. While models currently available show potential for promising identification and reconstruction of old texts, they are compromised by the scope and consistency of available data, and the quality and nuance of the inscriptions. Generative AI has the potential in the future to close these gaps and enable better restoration of removed or destroyed symbols and more nuanced understanding of the linguistic and cultural meaning of the Indus script.

Index Terms:

1 INTRODUCTION

Among the world's oldest cities, the Indus Valley Civilization—otherwise referred to as the Harappan Civilization, thrived in the extensive floodplains of the Indus River from 2600 to 1900 BCE [4]. Extended from what is presently Pakistan and northwestern India and possibly even as far as across the now-non-existent Saraswati River [2], the civilization is world-famous for its highly sophisticated urban planning in terms of features such as very advanced drainage, plus standardized architecture utilizing baked brick, and having a uniform system of weights and measures [3]. Extensive excavations have uncovered a rich collection of seals and artefacts inscribed with the still-mysterious Indus script [20]. This still-undeciphered writing system remains one of the most profound linguistic enigmas in human history. Unlike Egyptian and Mesopotamian scripts, which were deciphered years ago, the Indus script remains indecipherable, over 80 years since intellectual interest first fell on it [11].

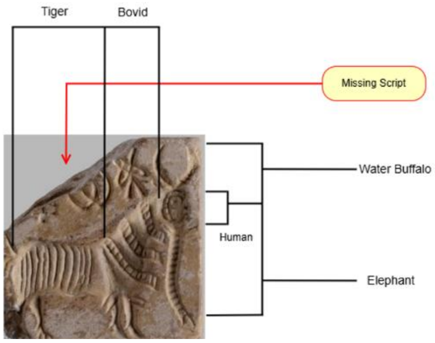

Fig. 1: Illustration of missing script detection in a damaged Indus seal. Highlighted segments show partially lost or eroded graphemes, which present challenges in script reconstruction and interpretation

The undeciphered nature of the Indus script remains an ongoing mystery and a fundamental issue in archaeological research [7]. Researchers have discovered about 70–80% of Indus inscriptions on seals [5], whereas the remaining inscriptions are found on pottery fragments, copper tablets, stone slabs, jeweler, and other artefacts (e.g. Fig. 1). Centuries of harsh weather and poor preservation at excavation sites have caused most artefacts to break, wear down, or become incomplete [6] in such conditions, script reading and symbol identification are very challenging [21].

Deciphering and analyzing the script is essential for unlocking and preserving valuable historical information.

$$Deciphering\ and\ Analysis\ =\ \sum_{i=1}^{5}(C_i + D_i + R_i + H_i + E_i)$$

Where:

C_i = Cultural Heritage Preservation

D_i = Digitalization Efforts

R_i = Access to Historical Research

H_i = Recovery of Lost and Destroyed History

E_i = Educational Resources and Outreach

This article provides a systematic review of recent advancements in Indus script research. We focus on identifying and interpreting epigraphic text and graphemes in the context of damaged seal artefacts. We examine past methods and explore how emerging technologies, machine learning and computer vision—can drive new developments in archaeology and epigraphy research. The following Fig. 2 provides an overview of AI-based methods for interpreting degraded Indus seal scripts and discovering motifs.

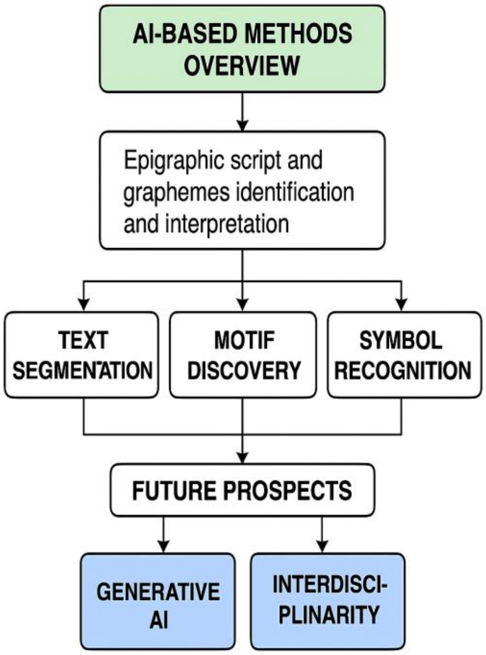

Fig. 2: AI-based methodological framework for the identification and interpretation of epigraphic scripts and graphemes. The workflow encompasses text segmentation, motif discovery, and symbol recognition, with prospects directed towards generative AI.

1.1 Research Objectives

- Read recent research on the Indus script.
- Recognize problems with broken or missing artefacts in script analysis.
- Describe the application of computer vision in grapheme and epigraphic text recognition.
- Explain the potential ways in which machine learning can be used to read, decipher, and interpret the Indus script.

Propose potential future applications of digital techniques for epigraphy in archaeology.

2 LITERATURE REVIEW

Deciphering the Indus script remains one of the most formidable challenges in archaeology and epigraphy. This difficulty mainly arises from two key factors: the poor condition of most artefacts and the lack of bilingual texts for comparison. The script mainly appears on seals and pottery shards and is believed to follow a logo-syllabic structure where individual signs may represent entire words or syllables like the earliest Sumerian or Egyptian writing systems.

According to Ralph et al. [16], the question of how to interpret script from a bilingual text and within a cultural context is a major challenge. Moreover, variations in writing style significantly complicate the process of deciphering the text. To solve this problem, researchers have turned to new technologies—artificial intelligence (AI), natural language processing (NLP), optical character recognition (OCR), and deep learning—to automate and improve the analysis of scripts. Empirical methods for analysis include pattern detection, image processing, semantic analysis, and even generative AI models.

As seen recently by studies such as Mirbahar, Q. et al. [8], the script is consistent with a combination of logographic and syllabic symbols. This is in conformity with the hypothesis that the language the Indus script may derive from may be Dravidian, a hypothesis which have been tested empirically by based on Asko Parpola, based on the rebus principle.

In a comparative analysis of anomalous Indus inscriptions in West Asia, Varun Venkatesh et.al. [19] employed the ICIT corpus to compute perplexity scores for different texts. The results indicated important stylistic contrasts between West Asian inscriptions and Indian subcontinent inscriptions, due to the presence of infrequent sign patterns. They employed Lidstone's model to detect outlier texts in places such as Mesopotamia (67% perplexity) and the Persian Gulf (60%).

Satish et al. [10] propose a deep learning pipeline in which Indus seals are unpacked into graphemic (text) regions, non-text regions and mixed regions to support digital humanities tasks via pattern recognition and classification. The system already detects repetitive symbols (i.e. the jar sign 342), from the Mahadevan corpus.

Reddy et al. [1] from the Florida Institute of Technology created ASR-Net and MIP-Net. The machine learning algorithms assist in detecting graphemes and motifs from annotated Indus seal datasets. They built a high-resolution image database that is supplemented with metadata. A key question asked was if AI models can replicate human-level recognition of symbols, even for broken or partially visible images.

In the field of epigraphy, Padmaprabha et al. [12] introduced an automated segmentation approach using region-based CNNs, leveraging models like InceptionV3 and ResNet50. Their system performed well in separating text from noisy, aged inscriptions. InceptionV3 effectively digitized ancient scripts.

In another pertinent contribution, Ayush et al. [17] investigated a blend of CNNs and conventional approaches such as SIFT + SVM for the interpretation of ancient inscriptions. Their results indicate that CNNs perform better than traditional methods, with 98.22% accuracy. Furthermore, methods such as Contrast Limited Adaptive Histogram Equalization improved image quality prior to classification.

Table 1: Comparison of various Artificial Intelligence approaches on Indus Valley Civilization

Ref.	Model	Dataset	Achievement	Limitation
[9]	Hybrid model	Nine Indus Valley copper, plates including 400 to 700 distinct signs	Pattern recognition in ancient inscription, AI-assisted analysis of script	Small sample size, lack of bilingual text, short length of inscription
[1]	ASR-Net with M-Net and YOLOv3	1,2264 (300+ images for each class)	Comprehensive database with high-resolution images	Fail on broken, stylish seals, complex motif
[15]	CNN model	6,053 images	Train 98%, validation 92%, prediction 70%	Operable on ten Indus signs
[17]	CNN, SIFT + SVM	3,000 images	98.22% accuracy	Lower performance with SIFT + SVM on max key approaches
[18]	n-gram, Markov chain model	ICIT corpus dataset	Amongst five signs prediction accuracy of 57% and amongst ten 63%, missing sign rate prediction 40%, only 35% rate get success	Limited dataset
[12]	Region-based CNN	2,000 regions per test	Effective segmentation of text from noisy ancient epigraphical	Denoised images
[19]	Markov chain language model: Lidstone language	ICIT corpus dataset	67% of the texts from Mesopotamia and 60% from the Persian Gulf	Low generalizability
[14]	CNN model	1,000 images	Pen data achieved 90%, mouse data achieved 80%	Model accuracy decreases by Increasing no. of epochs

Ali et al. [18] utilized the application of n-gram Markov chain models to simulate the positional patterns of symbols, with 63% accuracy on filling in missing signs for over 100 texts. Not only is the process utilized to fill gaps in the corpus of the Indus script, but also theories for placing symbols and building sentences can be further explained.

With the lack of adequately preserved instances, Yasufumi et al. [14], [15] overcame data limitations by constructing a crowdsourced dataset of hand-drawn Indus symbols. By using a web-based platform, IndusDraw, 36 volunteers provided drawings of 10 symbols. The researchers employed them to train classification models using TensorFlow and PyTorch to achieve approximately 70% accuracy, a promising result that improves when more volunteers are used.

Florian et al. [9] decode nine inscriptions in Cracking the Code of the Indus Valley Civilization Indus copper plates through a multi- Indus copper plates through a multi-disciplinary approach combining symbol frequency analysis, AI-based pattern matching, comparative mythology, and proto-linguistic reconstruction. Using techniques including symbol segmentation, computer vision, machine learning, clustering, and a translation matrix, the study finds symbolic parallels with other ancient languages and proposes meanings based on these parallels. Contrary to administrative purpose, the findings favor the notion that, like Mesopotamian and Egyptian systems, the Indus script communicated theological and philosophical concepts. Attempts are still being made to explore further the cultural context of the script and streamline AI-based decoding processes.

Rajesh et al. [13] used statistical models to examine sequential dependencies within the script. According to their analysis, some signs regularly appear at the end or beginning of messages, reflecting syntactic roles—observations similarly voiced by other researchers.

3 COMPARATIVE ANALYSIS

The paper contrasts prominent works in the study of the Indus Valley script that describe how they approached them, their most insightful findings, and limitations. They employ varying sets of interdisciplinary approaches such as deep learning, image processing, language analysis, and statistical models for segmenting, interpreting, and assisting the decipherment of the script. Collectively, they provide significant contributions to the subject. For more information, refer to Table 1. Despite significant progress, the wide range of methods used highlights the complexity of the problem. Differences in data availability, image quality, interpretive practices, and computational models reveal underlying technical challenges and methodological lacunae. These multifaceted efforts extend the boundaries of research, but they also underscore a definite requirement for more harmonized, integrated approaches that can tackle these challenges holistically.

4 CURRENT CHALLENGES

There are a few technical and research challenges that continue to impede Indus Valley script research progress. Machine learning models like ASR-Net [1] and CNN-based [12] struggle with working effectively with broken or partially occluded symbols. Their low noise robustness lowers effectiveness on worn-out or incomplete ancient inscriptions. Statistical models [18] attempt to identify symbol patterns, yet without similar bilingual texts, the patterns cannot be simply tested. Lack of annotated data also hinders model training and influences prediction performance, as noted in the report [15] where 70% accuracy was attained through crowd-sourced data.

The interpretation of symbols is made complex by the nature of context [8], whose interpretation is therefore a challenge with current OCR systems, which interpret characters independently. On these front multimodal models that derive visual and relational data from the script can be beneficial [13]. Metrics of human judgement are required to help close the gap between computational efficiency and human perception of understanding the text, a key pre-requisite to the effective preservation of the script [16].

5 CONCLUSION

Advanced machine learning and artificial intelligence (AI) technology has made the study of the Indus Valley script completely rethink able. Technology is solving, for the first time, the problems hampered by damaged artefacts and a shortage of well-studied bilingual specimens from ancient times. We have learned how to break texts into parts with deep learning and to discern patterns that point to a likely logographic character, notably in the script. Few signs can be made to recognize them, and large amounts of data remain incomplete, however, giving us much hope of cracking down on the script to understand its inner culture.

6 FUTURE WORK

Researchers also have to generate more large and larger datasets e. g. high-resolution images and metadata to conduct model training. With larger datasets, advanced methods can be used to better predict missing signs in incomplete inscriptions. Combining computational approaches like neural networks and symbolic reasoning can help produce a uniform distribution of results in script analysis. As generative methods develop, these models can also be used to predict missing symbols and to reveal new linguistic and cultural insights concerning the Indus script.

References

[1]. D. M. Atturu, *Deep Learning in Indus Valley Script Digitization*, Melbourne, FL: Florida Institute of Technology, 2024.

[2]. M. A. Fitzsimons, "The Indus Valley Civilization," The History Teacher, pp. 9–22, 1970.

[3]. K. Jonathan Mark, "The Origin, Context and Function of the Indus Script: Recent Insights from Harappa," in Proceedings of the Pre-symposium of RIHN and 7th ESCA Harvard-Kyoto Roundtable, Kyoto: Research Institute for Humanity and Nature (RIHN), 2006, pp. 9–27.

[4]. J. M. Kenoyer, "Indus Valley Civilization," in Encyclopedia of India, Chicago, IL: Fitzroy Dearborn Publishers, 2006, pp. 258–266.

[5]. J. M. Kenoyer, *Walking with the Unicorn*, Private Publication, 2017.

[6]. J. M. Kenoyer, "The Origin and Development of the Indus Script: Insights from Harappa and Other Sites," in *Studies on Indus Script*, S. O. Script, Ed. Tokyo: Research Institute for Humanity and Nature (RIHN), 2020.

[7]. A. Lawler, "The Indus Script—Write or Wrong?," *Science*, vol. 306, no. 5704, p. 2026, 2004, doi: 10.1126/science.306.5704.2026.

[8]. Q. Mirbahar, "Indus Script: Complexities and Deciphering Challenges," *Journal of History, Art and Archaeology*, pp. 19–27, 2024.

[9]. F. Neukart, "Cracking the Code of the Indus Valley Civilization: A Computational Approach to Lost Knowledge," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5141753.

[10]. S. Palaniappan and R. Adhikari, "Deep Learning the Indus Script," *arXiv* preprint *arXiv:1702.00523*, 2017.

[11]. A. Parpola, "The Indus Script: A Challenging Puzzle," *World Archaeology*, vol. 17, no. 3, pp. 399–419, 1986, doi: 10.1080/00438243.1986.9979979.

[12]. P. Preethi and H. R. Mamatha, "Region-Based Convolutional Neural Network for Segmenting Text in Epigraphical Images," *Artificial Intelligence and Applications*, pp. 119–127, 2023, doi: 10.47852/bonviewAIA2202293.

[13]. R. P. Rao, N. Yadav, M. N. Vahia, H. Joglekar, R. Adhikari, and I. Mahadevan, "A Markov Model of the Indus Script," *Proceedings of the National Academy of Sciences*, vol. 106, no. 33, pp. 13685–13690, 2009, doi: 10.1073/pnas.0906237106.

[14]. S. Saini, H. Shibata, and Y. Takama, "Toward Construction of Handwritten Indus Signs Dataset," in *Proc. 10th Int. Symp. on Computational Intelligence and Industrial Applications (ISCIIA 2022)*, 2022.

[15]. S. Saini, H. Shibata, and Y. Takama, "Construction of Handwritten Indus Signs Dataset Employing Social Approach," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 28, no. 1, pp. 122–128, 2024.

[16]. R. Shad, A. Egon, and K. Potter, "Application of Artificial Intelligence in Deciphering Ancient Scripts and Languages," *EasyChair Preprint*, 2024.

[17]. C. Sindhu, R. Vasanth, and A. Dalara, "Epigraphy Analysis Using Image Processing," *AIP Conference Proceedings*, vol. 2782, no. 1, Art. no. 020045, Melville, NY: AIP Publishing, 2023.

[18]. V. Venkatesh and A. Farghaly, "Deciphering the Indus Script: Decoding Missing and Unclear Indus Signs and Identifying Anomalous Indus Texts from West Asia Using Markov Chain Language Models," in *Proc. 2023 IEEE Integrated STEM Education Conference (ISEC)*, pp. 233–233, 2023, doi: 10.1109/ISEC57411.2023.10402308.

[19]. V. Venkatesh and A. Farghaly, "Identifying Anomalous Indus Texts from West Asia Using Markov Chain Language Models," in *Proc. 2023 14th Int. Conf. on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7, 2023, doi: 10.1109/ICCCNT58801.2023.10451706.

[20]. B. K. Wells, *The Archaeology and Epigraphy of Indus Writing*, Oxford, U.K.: Archaeopress, 2015.

[21]. N. Yadav and M. N. Vahia, "Indus Script: A Study of Its Sign Design," *SCRIPTA: International Journal of Writing Systems*, pp. 133–172, 2011.

Authors' Profile

Muzaffar Ali Khan earned his Master of Science in Computer Science from the University of Engineering and Technology (UET), Lahore, Pakistan. He received his bachelor's in computer science from the Virtual University of Pakistan. His research areas of interest are computer vision, natural language processing, and applications of artificial intelligence in archaeology and the conservation of cultural heritage. He is in the process of employing generative AI and deep learning methods for the restoration and interpretation of ancient artwork and scripts.