Deciphering Ancient Inscriptions with Optical Character Recognition

Anitha Julian[1], Devipriya R[2]

{cse.anithajulian@gmail.com[1], devipriyapriya326@gmail.com[2]}

Saveetha Engineering College, Chennai, India[1,2]

Abstract. Archaeologists strive to gain a deeper understanding of historical contexts across various regions by deciphering ancient Tamil inscriptions. However, this manual decoding process demands a considerable amount of labor and time. The inefficiency of this traditional approach has the potential to hinder future archaeological research. To address this issue, a proposed effort is underway to create an Optical Character Recognition (OCR) system specifically designed for the interpretation of medieval Tamil inscriptions. This study primarily focuses on the OCR module, which combines fully convolutional technologies with adaptive neuro-fuzzy inference (ANN). By conducting tests using actual images, the recognition rates of these two micro technologies were compared for the initial training set, preprocessed test data, and test data. Ultimately, the CNN-based OCR module emerged as the superior solution for this purpose.

Keywords: deciphering, inscriptions, CNN, OCR, character recognition

1 Introduction

Numerous inscriptions have come to light in various Asian countries, including the ancient cities of Polonnaruwa and Anuradhapura in Sri Lanka. Among these, we find notable examples like the Golgotha (Stone Book), the Mirror Wall, and Thonigala engravings. These inscriptions hold immense significance as they represent the primary sources of knowledge about ancient Sri Lanka and vital insights into the geographical context, historical circumstances, and linguistic evolution. Presently, the task of translating these inscriptions into contemporary Tamil is undertaken manually by archaeologists with specialized expertise in ancient scripts. The decipherment of these inscriptions is particularly challenging due to the complexity of the Tamil language's historical development. Moreover, the inscriptions may be damaged or partially eroded, posing further difficulties. Another critical issue lies in the scarcity of specialized knowledge and resources dedicated to the interpretation of inscriptions. Researchers currently invest significant effort in unraveling these inscriptions, necessitating the manual interpretation of their content.

IACIDS 2023, November 23-25, Lavasa, India
Copyright © 2024 EAI
DOI 10.4108/eai.23-11-2023.2343235

2 Literature Review

The importance of digitizing ancient images and documents, particularly stone inscriptions, to unveil cultural history and safeguard our national heritage has been emphasized by the authors of [1]. Brahmi, an ancient script used in India, South Asia, and Central Asia, plays a significant role in this context. In [2], the authors introduce an efficient system capable of processing unsegmented Brahmi character arrays, overcoming OCR errors, and employing a word identification scoring method to convert them into coherent Sinhala sentences. Meanwhile, [4] showcases the advantages for epigraphists and archaeological researchers working with the Devanagari Script. In [5], the focus is on accurately translating ancient text from images of stone inscriptions into modern languages, achieving a 90% accuracy rate. [6] presents a novel feature using the Histogram of Oriented Gradients for recognizing characters inscribed on ancient stones, irrespective of the language script used. In [7] and [11], databases are created, featuring alphabets from Kadamba, Chalukya, Hoysala, and Vijayanagara phases, along with their corresponding Kannada counterparts in jpg or bmp format, illustrating the transition of Kannada characters between different forms. [8] introduces an improved optical character recognition technique for ancient Tamil script, prevalent between the 7th and 12th centuries. This technique incorporates Google's text-to-speech voice engine to produce an audio output of the digitized text. The utilization of Convolutional Neural Networks, a common Deep Learning model, is suggested in [9] and [10], demonstrating high performance in image classification. Lastly, [12] proposes the implementation of neural networks and Markov models in the analysis of stone inscriptions. The proposed work creates an Optical Character Recognition system for interpretation os medieval Tamil inscriptions. The system combines convolutional technologies.

3 Methodology

The initial stage involves exploring viable Optical Character Recognition (OCR) systems. It is imperative to commence with the preparation of a comprehensive dataset. In this context, the research methodology is divided into the following three distinct phases:

1. In the first phase, we undertake the creation of a database comprising ancient characters intended for training purposes.
2. The subsequent phase entails the evaluation of various operational OCR systems and the prediction of their respective outcomes. Each OCR solution will follow the character recognition process, encompassing preprocessing, feature extraction, and character recognition steps. Our primary objective here is the development of an OCR method characterized by an enhanced detection rate.
3. Finally, the most effective OCR solution is integrated into the system across all three stages.

The primary goal is the development of a system capable of interpreting ancient Tamil texts through the utilization of Optical Character Recognition (OCR) technology. This study places a specific emphasis on the OCR functionality within the application. The OCR tool employs both Artificial Neural Networks (ANN) and advanced Deep Neural Networks (CNN) technologies. A comparative analysis of recognition rates was conducted through experiments involving training data sets, preprocessed test data, and real image test data for these two OCR approaches. Various solutions were evaluated, ultimately leading to the determination that CNN produced the most optimal OCR results. However, a significant research limitation emerged in the form of limited available data, which has a notable impact on OCR accuracy. Consequently, the CNN OCR technology successfully detected nine characters. The Optical Character Recognition (OCR) process is subdivided into three key components: pre-processing, character recognition and post-processing. Apart from this, template matching stands as the fundamental Optical Character Recognition (OCR) method for identifying image regions that closely resemble a template and are of the same size. Feature extraction involves the utilization of various well-established feature extraction techniques such as the Histogram of Oriented Gradients (HOG), Accelerated Substantial Features (SURF), Binary Pattern (LBP), Haar wavelet transforms, Color Histograms, and other extensively employed methods. In the context of template matching, the proposed method adopts the Lengthened Robust Features approach, which is known for its computational efficiency when compared to alternative techniques. After feature extraction, an input character image is juxtaposed against a collection of templates representing each language class.

4 Implementation and results

The implemenation process is shown in Table1. A classification report calculates crucial metrics, including accuracy, recall, precision, and the F1-score (shown in Table 2), all of which rely on values from a confusion matrix.

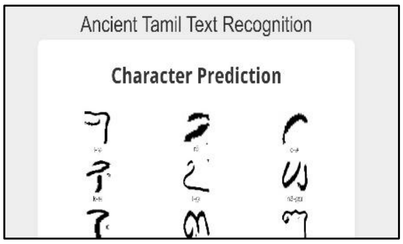
Table 1. Implementation process

Preprocessing	Sanitizes the image
Character recognition	Engraves the image
Post processing	Acknowledges
Image processing	Uses True OCR
Template matching	Uses HOG,SURF,LPG
Convolution layer	Uses Convnet
Preprocessing of Image data	Encodes the image
Preprocessing of image data	Gathers the image

4.1 Experiments and results

The methodology involves a comparative analysis of two approaches, namely the Artificial Neural Network (ANN) and the Convolutional Neural Network (CNN). While we explored character recognition techniques, they proved inadequate due to their low recognition rates and inability to effectively handle complex characters. In the context of template matching, objects are represented as sample images or "templates." The matching process involves comparing the number of pixels within these templates. However, this method also exhibited limitations, characterized by low recognition rates and an incapacity to effectively identify complex characters. These objects are usually represented as sample images or "templates." When a suitable match is detected, the Optical Character Recognition (OCR) system may successfully identify the object within the image. Our findings indicate that a significant portion of the test data is impacted by these limitations, resulting in low recognition rates and an inability to effectively handle complex characters as shown in Figures 1 to 4.

Many text scripts exhibit a fusion of elements from external languages and scripts, necessitating the extraction of valuable information from intricate and context-dependent sources.



Ancient Tamil Text Recognition

Character Prediction

Fig 1. Character prediction

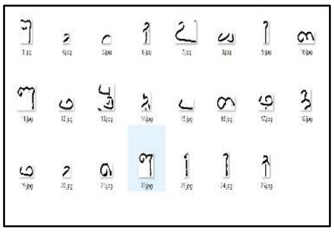


Fig 2. Character recognition

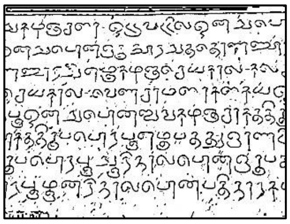


Fig 3. Previse

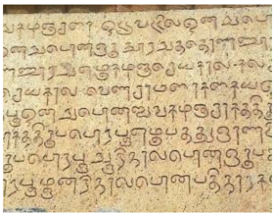


Fig 4. Envision of the stone

In this context, the importance of these sources remains pivotal, delineating a well-defined scope for the exploration of ancient manuscripts engraved in stone. To tackle this challenge, we employ the NGFICA algorithm [3], which harnesses a range of Gaussian techniques, including Sub Gaussian and Super Gaussian, to enable the extraction and analysis of such complex text scripts. The evaluation process entails the thorough analysis and interpretation of data, incorporating various enhancement techniques to present clear and comprehensible results that can be utilized and further improved. This approach offers improved differential propagation and avoids the backpropagation complexities associated with other activation functions. An activation function is a mathematical equation that determines whether another node in a network should be activated. The process involves applying a 5x5 matrix kernel to the Gaussian colored image, allowing for the identification of independent components within colored text through the NGFICA method. The model is identified by,

$$X = A\ Y \tag{1}$$

$A[X * T]$ is the original image,
$B[X * Y]$ is the independent text,
$C[X * X]$ is mixing text matrix,

The de mixing text scripts is processed as,

$$\begin{bmatrix} a1 \\ a2 \\ a3 \end{bmatrix} = \begin{bmatrix} s1 & s2 & s3 \\ t1 & t2 & t3 \\ u1 & u2 & u3 \end{bmatrix} * \begin{bmatrix} b1 \\ b2 \\ b3 \end{bmatrix} \tag{2}$$

$\boldsymbol{Z}[\boldsymbol{n} * \boldsymbol{T}]$ is the separated sources and $\boldsymbol{W}\ [\boldsymbol{n} * \boldsymbol{n}]$ be the de-mixing matrix. Then,

$$\boldsymbol{\nabla L}\ (\boldsymbol{W}) = \boldsymbol{w} - \boldsymbol{T} - \boldsymbol{E}\ [\boldsymbol{g}\ (\boldsymbol{Y})\ \boldsymbol{XT}] \tag{3}$$

$$\boldsymbol{g}\ (\boldsymbol{Y}) = \ -\boldsymbol{d}\ /\boldsymbol{dY}\ (\mathbf{log}\, \boldsymbol{P}(\boldsymbol{Y})) \tag{4}$$

The different text gradient formula is given by the Euclidean

$$\nabla L\ (W) = \ (I - E\ [g\ (Y)\ YT])\ W \tag{5}$$

To reduce or lower the output components L(W) has to be reduced. As proved in (2) gradient method is speediest technique for setting set of text. The performance of the proposed system is shown in Table 2.

Table 2. Inference of results

	Before OCR processing.			After OCR processing		
Image	Number	OCR	Accuracy	11	44.5	76.1%
Text	555	57	11%	33	2235	87.4%
Character	2578	835	33%	11	44.5	76.1%

.

5 Conclusion

According to the preceding interpretation and results, CNN OCR engine outperformed ANN OCR engine in the accuracy department. Iso systems primarily structured according to the effect of analytical load on engine performance as well as the overall recognition subassembly. Our version of CNN (with a 5 percent on average error rate) can be incorporated in 2-4 hours, obviously it depends on how familiar one is with it. As a result, the OCR engine's training time will be quite adaptable. The ancient Tamil early byzantine character segmentation system developed with more huge datasets could benefit from a CNN-based OCR engine.

References

[1] R. Vijayalakshmi et al., "A Review on Character Recognition and Information Retrieval from Ancient Inscriptions," *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, Chennai, India, 2022, pp. 1-7.

[2] S. Wickramarathna et al., "Data Driven Approach to Brahmi OCR Error Correction and Sinhala Meaning Generation from Brahmi Character Array," *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 2019, pp. 1-6.

[3] Indu Sreedevi et al., „NGFICA Based Digitization of Historic Inscription Images“, Hindawi, Volume 2013, Article ID 735857, pp. 1-7.

[4] B. S. Babu et al., "Temple Inscriptions Recognition and Transliteration in Devanagari Script," *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, Vellore, India, 2023, pp. 1-6.

[5] M. P. Kurapati et al., "A Review: Text Extraction from Stone Inscriptions and Translating to Modern Language," *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2023, pp. 1515-1519.

[6] G.Bhuvaneswari,, G.Manikandan, “Recognition Of Ancient Stone Inscription Characters Using Histogram of Oriented Gradients”, SSRN Electronic Journal, 2019, pp. 1-6.

[7] Puneeth P et al., “Characterization and Recognition of Stone Inscription”, International Journal of Research In Electronics and Computer Engineering (IJRECE) Vol. 7 Issue 2, 2019, pp. 1087-1090.

[8] Giridhar, L et al., “A Novel Approach to OCR using Image Recognition based Classification for Ancient Tamil Inscriptions in Temples”. ArXiv, abs/1907.04917, 2019.

[9] M. Merline Magrina, “Convolution Neural Network based Ancient Tamil Character Recognition from Epigraphical Inscriptions”, International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 04, 2020, pp. 6130-6143.

[10] David Bouchin, “Character Recognition using Convolution Neural Networks”, Statistical Learning Theory, 2018, pp. 1-9.

[11] Imran Khan et al., “Read and Recognition of old Kannada Stone Inscriptions Characters using MSDD Algorithm”, International Journal of Engineering Research & Technology (IJERT), RTESIT - 2019 Conference Proceedings, Volume 7, Issue 08, pp. 1-5.

[12] Thantilage, D.K, “Ancient Sinhala Inscription Character Recognition using Deep Learning, Informatics Institute of Technology, Digital Dissertation Repository, 2021.