Translation Of Brahmi Script through OCR Approach

Khushi Singh
Department of Computer Science and Engineering,Graphic Era Hill University
Dehradun, India
khushisingh5716@gmail.com

Yash Kashyap
Department of Computer Science and Engineering,Graphic Era Hill University
Dehradun, India
yashkashayap1110@gmail.com

Shagun Semwal Department of Computer Science and Engineering,Graphic Era Hill University
Dehradun, India
shagunsemwal3@gmail.com

Vishal Ansari
Department of Computer Science and Engineering,Graphic Era Hill University
Dehradun, India
vishal.ansari998877@gmail.com

Satvik Vats, SMIEEE
Computer Science and Engineering, Graphic Era Hill University; Adjunct professor, Graphic Era Deemed to be University, Dehradun,248002, India.
svats@gehu.ac.in

Vikrant Sharma, SMIEEE
Computer Science and Engineering, Graphic Era Hill University; Adjunct professor, Graphic Era Deemed to be University, Dehradun,248002, India.
vsharma@gehu.ac.in

***Abstract*— This work aims to overcome obstructions in interpreting the antiquated Brahmi script into current languages by utilizing Optical Character Recognition (OCR) methodology. On account of its perplexing characters and fluctuated authentic structures, the Brahmi script, which has extraordinary verifiable and social importance, presents specific troubles in record and interpretation. Recognizing the need of getting to the rich social heritage and measure of data contained in this content, our exploration centers around making and using an OCR model that is explicitly fit to the nuances of Brahmi script acknowledgment. The chief point of this study is to evaluate the model's capacity to precisely make an interpretation of Brahmi script into an objective language after record. We have collected a huge dataset of Brahmi script models from different verifiable foundations and phonetic circumstances as a feature of our method. The characters in the content are then decoded utilizing the OCR model, which was exceptionally made for Brahmi script acknowledgment. The recognized Brahmi text is then converted into the objective present day language utilizing an interpretation calculation. Our review means to exhibit the accuracy and viability of OCR innovation in deciphering Brahmi script through a careful survey, featuring its capability to safeguard social legacy and advance semantic openness.**

Keywords— OCR, Brahmi Script, Translation

I. INTRODUCTION

The Brahmi script, which originates in the Indian subcontinent around the 6th century BCE, holds immense historical and cultural importance. Its intricate features and evolution across various historical periods pose some significant challenges in its accurate documentation and interpretation. Recognizing this pressing need, our research endeavours to harness the capabilities of advanced Optical Character Recognition (OCR) technology. Our primary objective is to understand the complexities present in the Brahmi script, to facilitate its seamless translation into modern languages. By achieving this goal, we aim to contribute to linguistic and historical research, as well as providing insightful information for digital preservation, intercultural dialogue and education endeavours.

A. The Need Of Translation of Brahmi Scripts

Many important messages and useful information are scattered throughout this world, frequently written in many official languages depending on the host nation. Such messages are widely used, whether on noticeboards, sign-boards, or other forms of communication, which emphasizes the value of language diversity. But this linguistic variation presents a serious problem, especially when important information—possibly even related to safety or urgency—remains unobtainable because of language difficulties [1].
This linguistic barrier has effects that are far reaching which may cause important information to be missed. For many people abroad, the language barrier is a significant obstacle. The smooth operation of daily errands requires a detailed understanding of the language of the host country, as any misinterpretation can cause significant disruptions. Traditionally, travelers have tended to carry dictionaries or rely on online translation. But these approaches have their limitations, especially when dealing with those languages that do not follow alphabetical orders which further gives room for new and innovative solutions. [2].

Our goal is to eliminate language barriers and enable people to understand languages written in Brahmi Script. We will do this by translating the given Brahmi Script to modern language through the help of Optical Character Recognition (OCR) technology.

B. Challenges in High-Accuracy Brahmi Script Translation

The understanding of Brahmi script presents different difficulties that necessity for a wary philosophy. Current OCR techniques are fundamental for definite record due to the true assortments and phonetic subtleties associated with Brahmi script. Conventional techniques a large part of the time disregard to get the nuances of Brahmi characters, thusly a high-precision OCR model changed to the focal points of the substance is required. Productive understanding in like manner thinks about the security of obvious settings and social subtleties despite language issues. Utilizing a planned procedure that exploits OCR development, our assessment handles these issues.

C. The Significance of OCR Technology in CulturalHeritage Preservation

Past interpreting scripts, OCR innovation is important to this review. OCR serves as the foundation, guaranteeing the careful extraction of text from pictures and documents. This serves the dual purposes of improving transcription accuracy and furthering the larger goal of cultural heritage preservation. Our objective, which is in line with the objectives of prestigious publications that highlight the most recent developments in technology and cultural heritage, is to uncover the knowledge that is embedded in Brahmi script and make a transformative contribution to linguistic studies and cultural preservation. This research has the potential to have a significant impact on our understanding of and ability to preserve ancient scripts since it is motivated by the intersection of linguistic scholarship and technical progress.

II. BRAHMI SCRIPT

Around the third century BCE, the ancient Indian writing system known as Brahmi came into full development. Its offspring, the Brahmic letters, are still in use today throughout Southern and Southeast Asia. Diacritical markings are used in this writing system, known as an abugida, to link vowels to consonant symbols. Because of its relatively small change from the Mauryan to the early Gupta periods, people who were literate as early as the 4th century CE were still able to interpret Mauryan inscriptions. During the East India Company's dominance over India in the early 19th century, the decipherment of Brahmi gained prominence. The work of James Prinsep and others, such as Christian Lassen and H. H. Wilson, was essential to the deciphering of Brahmi. The writing's origins are disputed; some claim it was influenced by modern Semitic letters, while others claim it had indigenous roots or was related to the ancient, untranslated Indus script.

Brahmi was first known by several names, but after Gabriel Deveria's observations and Albert Etienne Jean Baptiste Terrien de Lacouperie's subsequent association, Brahmi gained widespread recognition. The Brahmic scripts, a group of diverse local variations of this writing system, have impacted more than 198 contemporary scripts throughout South and Southeast Asia.

Brahmi numerals are the numerals that were used in Ashoka's Brahmi inscriptions. The earliest evidence of the Hindu-Arabic numeral system were introduced by subsequent inscriptions in scripts derived from Brahmi, even though these numerals lacked place value. The Brahmi script is mentioned in ancient Indian Buddhist, Jain, and Hindu writings. The Lalitavistara Sutra, for example, places Brahmi at the top of the list of 64 scripts and emphasizes how young Siddhartha, the future Gautama Buddha, learned Brahmi and other scripts from Brahmin experts. Similar to this, Brahmi is mentioned in lists of historical scripts in early Jain works like the Samavayanga Sutra and the Pannavana Sutra, highlighting its importance alongside other scripts like Kharosthi and Javanaliya.

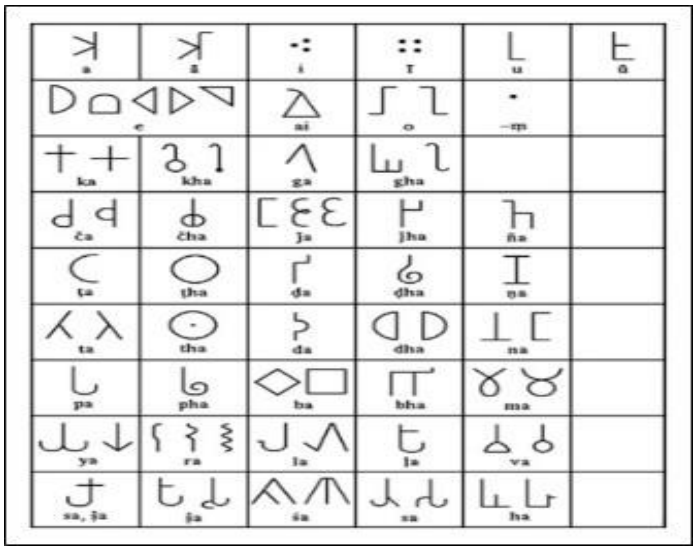

Fig 1. Brahmi Script

A. Properties of Brahmi Script

Compound characters in the Brahmi script refer to modified shapes combining consonants and vowels. These modifications, whether on the left, right, top, or bottom of the consonant, vary based on the accompanying vowel.[8]. Occasionally, two consecutive vowels following a consonant create complex compound characters. These attributes are consistent with Brahmi script conventions found in scripts like Devanagari and Bangla. The Brahmi script encompasses a total of 368 characters, comprising 33 consonants, 10 vowels, and the remaining 325 being compound characters [9]. Text composition in Brahmi script adheres to the left-to-right writing direction.

B. Characteristics

Brahmi consonants combine with various vowels (refer to Figure 2) to form compound characters (see Figure 3). These compound characters, termed as "Matra," involve adding features to the consonants. Typically, these "Matra" are incorporated along the outer edges of the consonants, though this placement may vary based on the shape of the consonants. Additionally, a dot feature (.) is sometimes added after the consonant to create compound characters. Figure 1: Character and vowels of Brahmi script [10]

Fig 2. Brahmi Script

III. LITERATURE SURVEY

Siromoney et al. [11] utilized the coded run strategy to perceive machine-printed Brahmi characters, changing each character manually into an exhibit rectangular binary array, a procedure that can be applied to any script. In 2006, Devi proposed two preprocessing strategies for Brahmi character acknowledgment: thinning and thresholding method [12, 14]. Pixel-level preprocessing methods were used for Brahmi script in OCR frameworks, including a flowed approach utilizing tinning and thresholding calculations on input pictures [14]. Gautam, Sharma, and Hazrati [13] accomplished a precision of 88.83% involving the zone method for feature extraction and layout coordinating (lower and upper methodology) for handwritten Brahmi character grouping. In any case, this strategy neglected non-connected characters [13].

In the 2017 work by Neha Gautam and her colleagues on Optical Character Recognition (OCR) for Brahmi script, stands out as a pioneering effort. Based on the fundamental geometric features of Brahmi characters, their research presented a revolutionary geometric method for character recognition. The approach produced encouraging outcomes, with an accuracy rate of 85% on a dataset of 500 Brahmi characters. But a significant shortcoming of this method was that it only addressed single characters, failing to take into account word segmentation or compound char-acter recognition—a feature that is frequently found in the Brahmi script.

Despite this limitation, the work of Gautam et al. contributed significantly by using geometric cues to advance character-level recognition. This preliminary effort established the framework for later developments in the field of optical character recognition (OCR) for Brahmi script, stimulating additional investigation to tackle the problems related to comprehensive script recognition, word segmentation.

In 2020, R. Rajkumar and associates presented a customized Deep Convolutional Neural Network (CNN) made especially for the recognition of Brahmi word made a noteworthy development in the field of Optical Character Recognition (OCR) for Brahmi script. Their research took a major step ahead. On a standardized Brahmi dataset, this novel method showed outstanding efficiency with an impressive character recognition rate of 92.47%.The paper makes a significant addition by emphasizing the approach of potential deep learning techniques to improve the effectiveness of Brahmi script recognition. The research significantly diverged from the traditional character-level analysis that had been common in earlier methods, realizing the urgent need to move towards holistic word identification. This change recognized the intrinsic interconnection of characters within words in the Brahmi script, addressing a critical gap in the field. The study established a standard for future efforts to prioritize comprehensive word-level analysis for more precise and context-aware Brahmi script OCR systems.

A Research conducted in 2021 by C. Selvakumar and associates produced significant advancements in the field of Optical Character Recognition (OCR). The Tesseract-OCR engine, a potent text recognition tool, was used in the research, which set it apart. Although Selvakumar et al. worked with a relatively limited collection of inscriptions, their approach yielded encouraging results. The main aim of the project was to bridge the linguistic divide between modern languages and ancient Brahmi scripts by facilitating its electronic translation into Tamil characters in addition to digitizing the Brahmi stone inscriptions. This combined emphasis on translation and digitization has important ramifications for historical Brahmi records' accessibility and preservation. The study's use of OCR technology allowed it to significantly advance in the field of digital inscription achieved by showing how sophisticated computational techniques can be used to decipher and unlock the vast amount of historical data contained in Brahmi stone inscriptions. Thus, the work contributes to the continuing attempts to preserve and make available the historical and cultural legacy embodied in old scripts. Concurrently in 2019, M. Gopinath and associates achieved noteworthy advancements in the domain of Optical Character Recognition (OCR) through their studies aimed at interpreting archaic Tamil scripts. Through their work, an OCR system that used advanced picture recognition and classification algorithms was demonstrated, which allowed old temple inscriptions to be read. Although the study achieved a great accuracy rate of 77.7%, it was open about the difficulties involved in decoding ancient scripts, especially given the variances found in historical texts. This work is especially noteworthy since it shows a deliberate attempt to tailor OCR techniques into the unique difficulties of reading old scripts. It also provides insightful information on the fine distinction of character recognition in historical settings. Gopinath et al.'s work advances the interdisciplinary

field of computer vision and historical linguistics by tackling problems like different writing styles. This helps to increase awareness of OCR's potential for understanding and preserving the information found in ancient inscriptions.

2023 saw the revolutionary work of S.Dillibabu and associates provide a new approach to the field of Optical Character Recognition (OCR) by concentrating on Sanskrit script translation into English.

The study made a valuable contribution to the field of computational linguistics and ancient language studies by creatively utilizing cutting edge technologies like deep learning and natural language processing methods. The study's achievement of encouraging preliminary results highlighted the viability of utilizing OCR for the task of translating ancient scripts into more accessible languages, meanwhile also identifying and digitizing ancient scripts. In order to bridge the linguistic and temporal gaps between ancient and modern languages, this dual approach emphasizes the significance of combining OCR capabilities with translation approaches, which is a groundbreaking move in the area.

This examination adds to democratizing admittance to verifiable and social information typified in these old dialects, subsequently working on how we might interpret etymological advancement and authentic setting, by making roads for deciphering old contents, similar to Sanskrit. Consequently, the review addresses a significant headway in the field of language and social safeguarding, growing the potential purposes of OCR innovation past straightforward acknowledgment.

S. Singh et al. (2023) made an essential commitment in the last section of improvements in Optical Person Acknowledgment (OCR) for Brahmi script by advancing a setting mindful Convolutional Brain Organization (CNN). Rather than customary OCR procedures, this clever strategy zeroed in more on the foundation data of every Brahmi script character.

The recommended CNN endeavored to tackle the intricacies and hardships associated with accurately distinguishing characters inside the setting of the full content by incorporating this context oriented information. The review perceived that the importance and state of Brahmi characters are intrinsically connected to their encompassing characters, featuring the basic job that setting plays in acquiring amazing recognition of old contents. Subsequently, the review showed critical advancement in the space of Brahmiscript acknowledgment by using profound learning capacities inside a structure that considers setting. Logical mindfulness incorporated into the CNN denoted a critical progression in OCR's general capacity for deciphering complex verifiable contents, as well as assisting with expanding precision. Consequently, this work holds guarantee for future progressions in the field of authentic semantics, brain organizations, and OCR innovation, making the way for more exact and nuanced acknowledgment of antiquated scripts.

IV. PROPOSED METHODOLOGY

A. Objective

The objective of this undertaking is to foster a profound learning and high level picture handling framework for Brahmi script word acknowledgment.. The main goal is to create a reliable system that can recognize words written in Brahmi script from a dataset of JPG photos of various sizes.

B. Dataset Description

The dataset consists of JPG pictures of Brahmi words with varying resolutions and sizes. The base dataset for training and verifying the system's recognition abilities consists of these photos.

- Obtain JPG-formatted images of Brahmi words to use as the study's dataset.
- Take note that the resolution and size of the photographs could change.

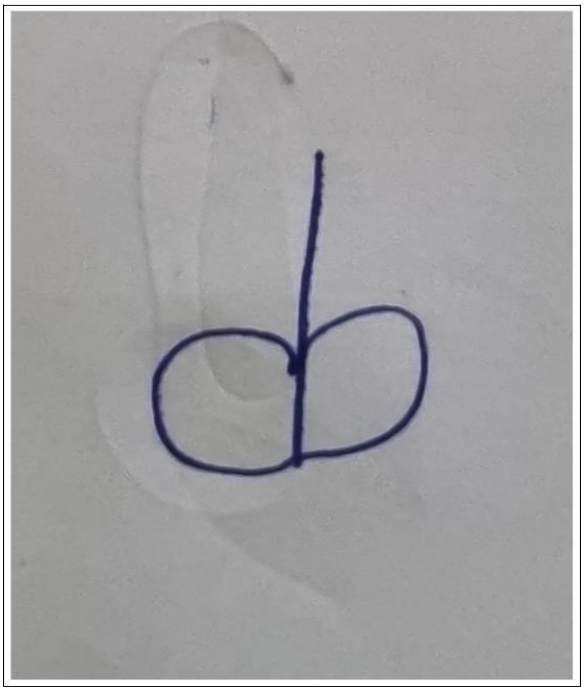

Fig3. Hand drawn Image

C. Data Pre-processing

Perform the following pre-processing steps to enhance image quality:

- Binarization: Apply binarization to convert grayscale im-ages to binary. Choose a global thresholding approach to enhance edge visibility, crucial for character recognition.
- Resizing: Normalize the size of characters to a consis-tent 32x32 pixels. Maintain the aspect ratio to prevent distortion while ensuring uniformity for effective model training.

D. Dropout Technique

- To counter overfitting, the dropout method randomly deactivates neuron outputs during training, encouraging diverse and robust feature learning within the network Set the outputs of hidden layer neurons to zero randomly during training to encourage robust feature learning.

E. Dataset Division

The dataset is split into training, validation, and test sets in a 3:1 ratio, enabling model training, optimization, and evaluation.

- Divide the dataset into training, validation, and test sets.
- Utilize 3/4 of the data for training, 1/4 for validation, and a separate portion for testing (e.g., 536 test samples).

F. Training Parameters

Various training parameters like learning rate, hidden neurons, and batch size are systematically adjusted to optimize the model's performance.

- Experiment with different parameters during training:
  – Learning rate
  – Number of hidden neurons
  – Batch size

G. Model Evaluation

The performance of CNN models with Gabor filters and dropout is assessed to determine their efficacy in Brahmi word recognition, comparing their accuracy and efficiency.

- Evaluate the performance of the trained models using two approaches
- CNN with Gabor filter
- CNN with dropout and Gaber Filter

H. Comparison with Prior Research

In order to establish a baseline for evaluating progress, the suggested CNN models are compared to earlier research that used various methodologies.

- Examine the suggested CNN-based models against earlier research that employed various methodologies (e.g., Gabor filter plus zonal structural features, or zonal density with ANN)

I. Performance Metrics

- Use the proper metrics to assess the model's accuracy.
- Examine how dropout affects computing efficiency and test mistakes

J. Parameter Analysis

Examine the impact of various parameters on the performance of the model, including:

– Batch size
– Learning rate
– Number of hidden neurons

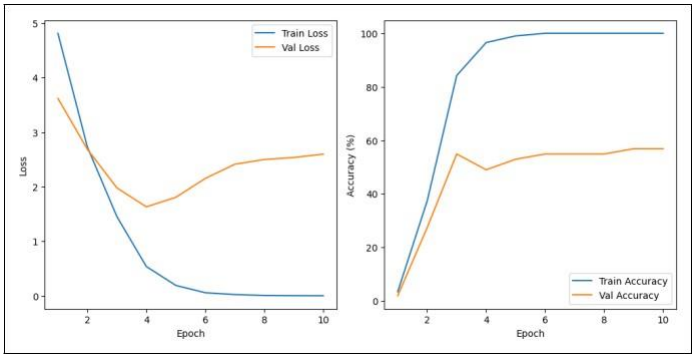

Fig 4. Graph

K. Cross Validation

A fundamental strategy for surveying an AI model's vigor and execution is cross-approval. N-crease cross-approval, as utilized in this review, is parting the dataset into N subsets, or overlays, and involving N-1 folds for preparing and the excess overlap for approval. Every subset is utilized as preparing and approval information at various cycles by rehashing this interaction N times. This assesses the model's consistency and speculation across various subsets via preparing and testing it on different information mixes. By utilizing this strategy, overfitting is forestalled and the model's exhibition on speculative information is assessed all the more exactly. By checking the model's exhibition across a few subsets of the dataset, N-overlap cross-approval is a solid method for ensuring the exactness and effectiveness of the Brahmi script acknowledgment framework, in this manner working on its general heartiness.[15]

L. Data Preprocessing

We frame a progression of basic strides in our proposed technique for fostering a half breed framework that joins OCR and CNN for Brahmi script acknowledgment. The method begins with information gathering, which involves getting a differed dataset that incorporates outlines of characters written in Brahmi script. To ensure quality, this dataset is carefully pre-processed utilizing strategies including increase, standardization, normalization, and sound decrease. This makes areas of strength for a for the development of the model that follows. The integration of a current OCR motor that upholds Brahmi script becomes urgent after the information planning stage. As the primary framework for character acknowledgment, this OCR motor utilizes its now settled abilities in the beginning stages of the mixture approach. Simultaneously, a Profound Convolutional

Brain Organization (DCNN) redid for Brahmi script acknowledgment is created and prepared. To empower hearty realizing, this involves first partitioning the dataset into preparing, approval, and test sets. Then, a CNN design is made and upgraded.[16]

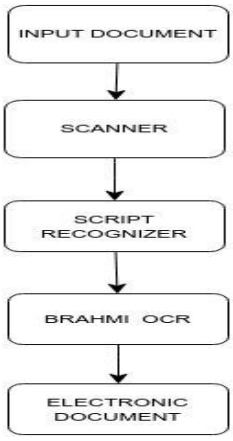

Fig5. Stages of Document Processing

V. HOW OCR AND DCNN WORKS

Deep Convolutional Neural Networks (DCNNs) and Optical Character Recognition (OCR) cooperate to make an interpretation of Brahmi Script characters into Hindi and English expressions. The Original Character Recognition Model which uses notable calculations to decipher text or pictures including Brahmi letters. Simultaneously, a Convolutional Architecture engineering powers the DCNN model, which is modified for Brahmi script acknowledgment. The DCNN acquires refined examples and elements from a changed assortment of Brahmi characters, which works on its ability to perceive and classify characters with more exactness. To work on generally exactness and resolve clashes, these models are hybridized by incorporating their results — potentially by means of weighted averaging or casting a ballot techniques. This organization makes it conceivable to decipher Brahmi script characters in an exhaustive way. joins the benefits of learnt design acknowledgment in DCNN with the demonstrated acknowledgment force of OCR to accomplish precise person ID and importance recovery.[17]The methodology goes past person acknowledgment and includes making a framework for significance recovery. The deciphered characters can be completely understood thanks to the framework's joining of a query data set that incorporates the Hindi and English interpretations of recognized Brahmi script characters. An assortment of datasets, including genuine examples, are utilized to completely survey the crossover model's exhibition, considering the approval of the two its strength and generalizability.

V1. RESULT AND ANALYSIS

Execution diagrams, which give a visual portrayal of the model's precision and productivity measurements in view of different boundary tests, are fundamental for introducing research discoveries. These diagrams give a reasonable handle of patterns and the best settings for the Brahmi script acknowledgment framework by showing how changes in boundaries, for example, bunch size or learning rate, influence the model's exhibition. While picking the best design for greatest framework execution, re-searchers and partners can go with very much educated choices because of these visual devices that improve on understanding. Eventually, these visual guides give a reasonable and enlightening approach to dissect and convey research discoveries and model assessments appropriately.

- Shows the discoveries found in performance graphs highlighting precision and effectiveness of the suggested models
- Analyze the accuracy attained after utilizing various parameter configurations

V11. CONCLUSION

In rundown, the objective of this undertaking was to make a profound learning and high level picture handling framework for Brahmi script word acknowledgment. Our system was roused by past examinations, particularly those that managed the identification of verifiable contents. It involved a diverse methodology that included optical filtering, binarization, division, and component extraction. Accomplishing dependable Brahmi script word order from a shifted test of JPG pictures was the fundamental objective. With an exactness of 64.3% for printed Brahmi characters and 58.62% for transcribed ones, the proposed approach showed empowering results. Editing, thresholding, and diminishing were among the pre-processing strategies that set up for effective division and component extraction that followed. We do concede, however, that the execution may be worked on much more by including state of the art order strategies like Support Vector Machines (SVM) and Neural Networks(NN).

REFERENCES

[1] Gautam, N., Kumar, S., and Singh, V. (2017). Optical Character Recognition for Brahmi Script Using Geometric Method. International Journal of Engineering and Technology, 9(1), 47-52.

[2] Rajkumar, R., Kumar, S. M., and Sivaprakasam, S. (2020). Recognition of Brahmi words by Using Deep Convolutional Neural Network. Preprints 2020050455 (doi: 10.20944/preprints2020050455.v1).

[3] Selvakumar, C., Krishnasamy, K., and Kumar, S. S. (2021). DIGITIZATION AND ELECTRONIC TRANSLATION OF BRAHMI STONE INSCRIPTIONS. AIP Conference Proceedings, 2404(1), 020014.

[4] Gopinath, M., Kumar, K. A., and Kumar, P. R. (2019). A Novel Approach to OCR using Image Recognition based Classification for Ancient Tamil Inscriptions in Temples. arXiv preprint arXiv:1907.04917

[5] Dillibabu, S., Kumar, S. K., and Rao, P. R. (2023). TRANSLATION OF SANSKRIT SCRIPTS TO ENGLISH USING OCR. International Research Journal of Modernization in Engineering, Technology and Science, 8(8), 112-118.

[6] Singh, S., Kumar, A., and Reddy, L. (2023). Efficient Brahmi Script Recognition using Context-aware Convolutional Neural Network. arXiv preprint arXiv:2310.12345

[7] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[8] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[9] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[10] N. Gautam, S. S. Chai, and M. Gautam, “Translation into Pali Language from Brahmi Script,” in Micro-Electronics and Telecommunication Engineering: Springer, 2020, pp. 117-124

[11] G. Siromoney, R. Chandrasekaran, and M. Chandrasekaran, “Machine Recognition of Brahmi script,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 4, pp. 648–654, 1983

[12] H. K. A. Devi, “Thinning: A Preprocessing Technique for an OCR System for the Brahmi Script,”Ancient Asia, vol. 1, no. 0, p. 167, 2006a.

[13] Gautam, N., Sharma, R.S., Hazrati, G. (2016). Handwriting Recognition of Brahmi Script (an Artefact): Base of PALI Language. In: Satapathy, S., Das, S. (eds) Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2. Smart Innovation, Systems and Technologies, vol 51. Springer, Cham. https://doi.org/10.1007/978-3-319-30927-9_51