Translation of Brahmi Script through OCR Approach

Muskan Choudhary

Student, Department of Computer Science

Graphic Era Hill University, Dehradun

itsmuskan819101@gmail.com

Abstract -- The aim of this paper is to develop a system that involves character recognition of Brahmi Characters from palm manuscripts of Historical Tamil Ancient Documents, anaylsed the text and machine translated the present Tamil digital text format. — 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. Though many researchers have implemented various algorithms and techniques for character recognition in different languages, Ancient characters conversion still poses a big challenge.

Because

Image recognition technology has reached near perfection when it comes to scanning English and other language text. But optical character recognition (OCR) software capable of digitizing printed Tamil text with high levels of accuracy is still elusive. Only a few people are familiar with the ancient characters and make attempts to convert them into written documents manually. The proposed system overcomes such a situation by converting all the ancient historical documents from inscriptions and palm manuscripts into Tamil digital text format. I have collected a huge dataset of Brahmi script models from different verifiable foundations and phonetic circumstances as a feature of my method. The algorithm comprises different stages: i) image preprocessing, ii) feature extraction, iii) character recognition and iv) digital text conversion. The first phase conversion accuracy of the Brahmi script rate of our algorithm is 91.57% using the neural network and image zoning method. My 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

INTRODUCTION

Brahmi script is one of the most important writing systems in the world by virtue of its time depth and influence. It represents the earliest post-Indus corpus of texts, and some of the earliest historical inscriptions found in India. The best-known Brahmi inscriptions are the rockcut edicts of Ashoka in North-central India, dated to 250– 232 BCE. This elegant script appeared in India most certainly by the 5th century BCE, but the fact is that it had many local variants even in the early texts which suggest that its origin lies further back in time. There are several theories on the origin of the Brahmi script. Many researchers have tried to apply many techniques for breaking through the complex problems of character recognition. Most difficult thing is to recognize the sculpting of Brahmi characters in stone compared to the other character recognition from different sources. The Brahmi characters were sculpted in stones, clay pot, copper plate etc. In this paper, a character recognition system based on a statistical approach with a new feature set is proposed.

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.

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

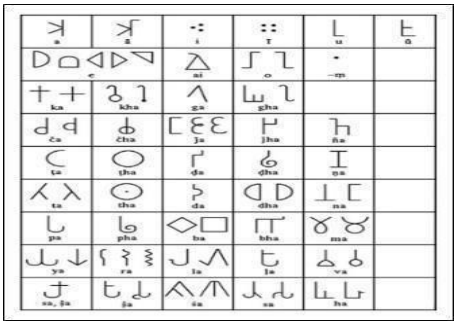


Fig 1. Brahmi Script

LITERATURE REVIEW

Three major approaches used commonly for recognition of handwritten characters are statistical, structural or syntactic and neural network based approaches.

A. Statistical Approach

Statistical or probability functions are used for formulating a recognition algorithm in statistical approach. Features to be given as input are derived from a set of measurements used for pattern recognition. There prevails a difficulty in representing the pattern of classification on the basis of structural information in this statistical approach.

B. Structural or Syntactic Approach

Syntactic approach employs syntactic or structural information about patterns to obtain information relevant to patterns. It extracts similar patterns and formulates pattern syntax or structural rules. The information about the pattern syntax rules helps highlight, classify and identify unknown patterns. The structural approach is favorable for evolving a recognition system for handwritten characters. The reason is that it uses structural rules to build much pattern syntax for handwritten characters, but it is not easy to formulate learning structural rules.

C. Neural Network Based Approach

Neural pattern recognition is a pattern recognition approach that is based on the storing and manipulation of information by a biological neural system. This is an artificial neural system and it is named as "neural networks". This system is expected to solve all the problems with automatic reasoning, including pattern recognition problems. It classifies patterns on the basis of predictable properties of neural networks. However, it represents only a little amount of information about semantic from a network.

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,

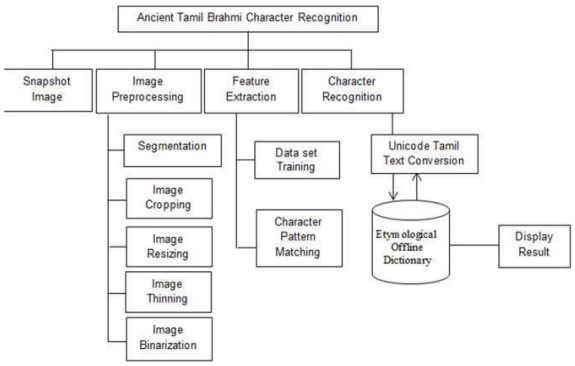


Fig 2: Recognition system for stone inscription

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.

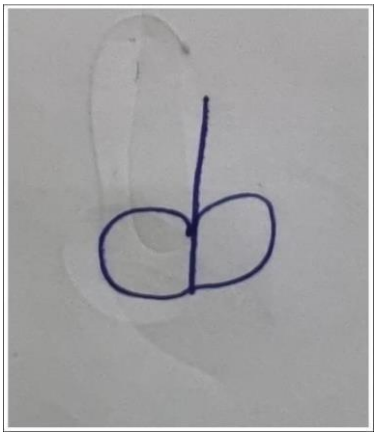


Fig3. Hand drawn Image

C. Data Processing

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

• Binarization: Apply binarization to convert grayscale binary. Choose a global thresholding approach to enhance edge visibility, crucial for character recognition.

Resizing: Normalize the size of characters to a consistent 32x32 pixels. Maintain the aspect ratio to prevent distortion while ensuring uniformity for effective model training.

D. Dropout Technique

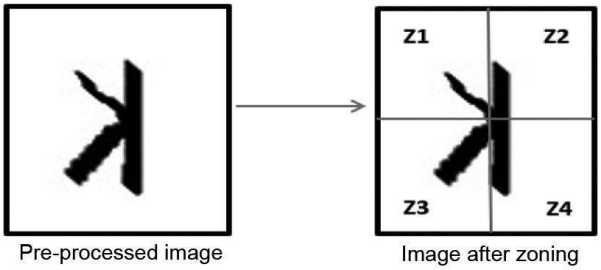
• To counter overfitting, the dropout method randomly deactivates neuron outputs during training, encouraging diverse and robust feature learning within the network Set the outputs of hidden layer neurons to zero randomly during training to encourage robust feature learning.

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

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 preprocessed utilizing strategies including increase, standardization, normalization, and sound decrease.

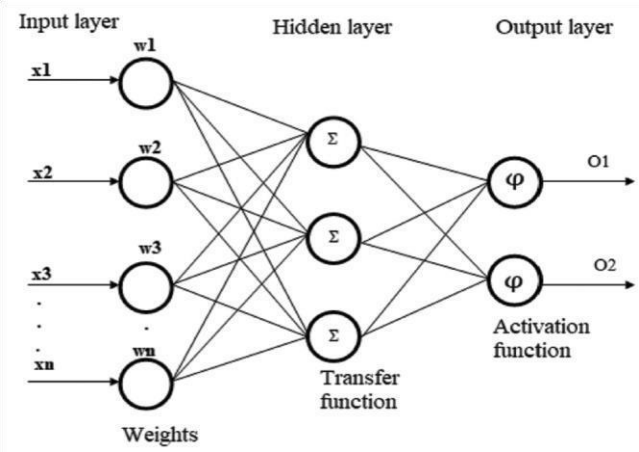


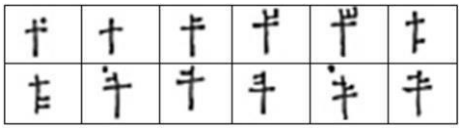
Fig 4: NN layers

VI. RESULTS AND DISCUSSION

Output

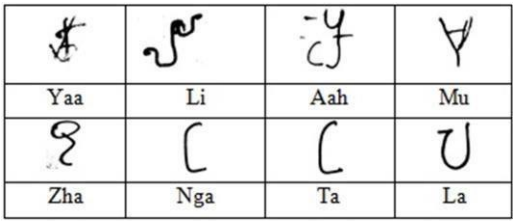
```
In [10]: import random
from PIL import Image
import torch

# Load the trained model
model.load_state_dict(torch.load('classification_model.pth'))
model.eval()

image = Image.open(image_path)


image = transform(image)

image = image.cpu().squeeze(0).unsqueeze(0)


# Make a prediction
with torch.no_grad():
    output = model(image)
    _, predicted = torch.max(output.data, 1)

# Print the prediction
print(f"Predicted class: {dataset.classes[predicted.item()]}")

Predicted class: va
```

The lack of character recognition is due to three reasons. One is that the consonantal vowel characters are similar to consonants.



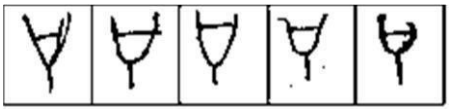
Similarity of consonantal vowels.

The second one, the writers do not write the Brahmi characters properly. Moreover, the characters are stored in different strokes and styles with ambiguity.

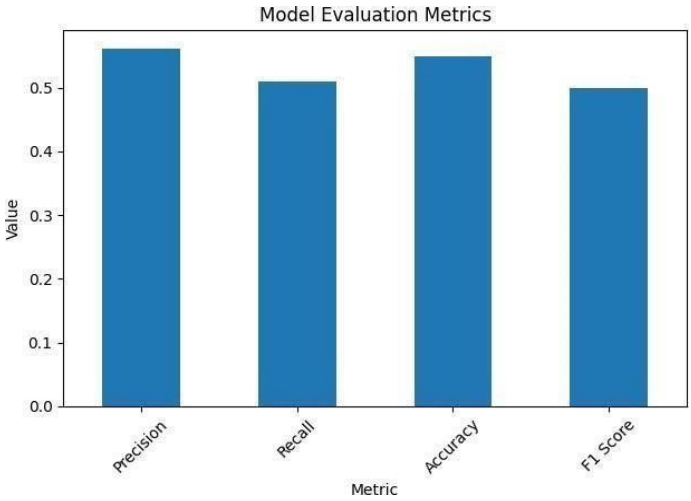


. Unknown characters in the data set

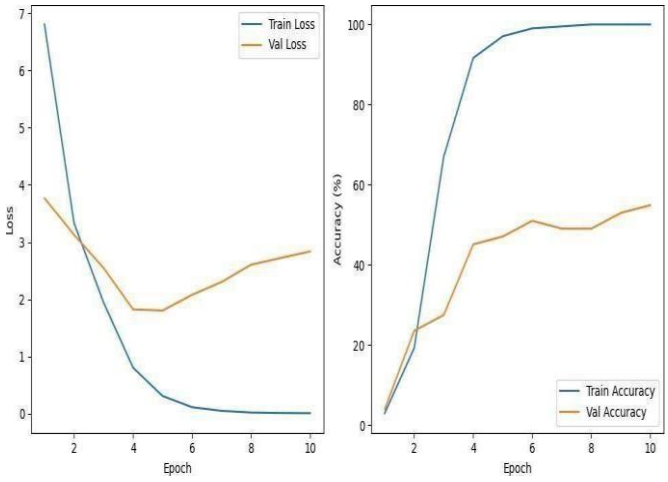
The same character with different style and stroke from different writers.



Maa character from different writers.



Evaluating the performance of the Brahmi script recognition system is crucial to understanding its accuracy, efficiency, and robustness. To achieve a comprehensive assessment, we employ a set of well-established performance metrics. These metrics provide quantitative insights into how well the system performs in recognizing and interpreting Brahmi script characters. The key metrics we use include precision, recall, F1-score, and accuracy, each offering a unique perspective on the system's capabilities.



CONCLUSION AND FUTURE WORK

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. In future, the system can be modified with the incorporation of better feature extraction methods to improve the accuracy rate. The writers shall get trained in writing the Brahmi letters to create the data set. By that, the error rate of ambiguous letters will be reduced. In addition, the system in future will recognize the Brahmi letters from the stone inscriptions for improving the result in using some hybrid algorithms.

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.

N. Gautam, S. S. Chai, and M. Gautam, “Translation into Pali Language from Brahmi Script,” in MicroElectronics and Springer, 2020, pp. 117-124.

[10] 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

[11] 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.

[12] 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

[13] H. K. Anasuya Devi Thresholding: A Pixel-Level Image Processing Methodology Preprocessing Technique for an OCR System for the Brahmi Script Ancient Asia 2042-5937 Ubiquity Press,Ltd.2006-12-011161
https://cir.nii.ac.jp/crid/1360013172726828672
https://doi.org/10.5334/aa.06113

[14] Vimal, Vrince, et al. "Artificial intelligence-based novel scheme for location area planning in cellular networks." Computational Intelligence 37.3 (2021): 1338-1354.

[15] H. Narang, R. Saklani, K. Purohit, P. Das, B. B. Sagar and M. Manjul, "Wheat Disease Severity Assessment Using Federated Learning CNNs for Agriculture Transformation," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 938-943, doi: 10.1109/ICTACS59847.2023.10389932.

[16] Durgapal, Ayushman, and Vrince Vimal. "Prediction of stock price using statistical and ensemble learning models: a comparative study." 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical,
Electronics and Computer Engineering (UPCON). IEEE, 2021

[18]. Plamondon, R. and Srihari, S. N., Online and off-line handwriting recognition: a comprehensive survey. IEEE Trans. Pattern Anal. Mach. Intell., 2006, 22(1), 63–84.

[19]. Sumathi, C. P. and Karpagavalli, S., Techniques and methodologies for recognition of Tamil typewritten and handwritten characters: a survey. Int. J. Computer Sci. Eng. Surv., 2012, 3(6), 23–35. 21. Indra Gandhi, R. and Iyakutti, K., An attempt to recognize handwritten Tamil character using Kohonen SOM. Int. J. Adv. Net. Appl., 2009, 1(3), 188–192. 22. http://en.wikipedia.org/wiki/brami_script

[20]. Suganya, T. S. and Murugavalli, S., A survey on character recognitions in Tamil scripts using OCR, 2012.

[21]. JeyaGreeba, M. V. and Bhuvaneswari, G., Recognition of ancient Tamil characters in stone inscription using improved feature extraction. Int. J. Recent Develop. Eng. Technol., 2014, 2(3), 38–41. 25. Zernike, F., Physica, 1934, 1, 689.

[22]. Wang, Qiu-Feng Yin, Fei and Liu, Cheng-Lin, Handwritten Chinese text recognition by integrating multiple contexts. IEEE Trans. Pattern Anal. Mach. Intell., 2012, 34(8), 1469–1481.

[23]. Tharwat, A., Classification assessment methods. Applied Computing and Informatics,2018.