International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 7, pp 154-158, July-2024

https://doi.org/10.47001/IRJIET/2024.807016

A Survey on Intelligent Kannada Inscription Character Recognition Using OCR and Machine Learning

[1]*Shivakumar B, [2]Dr. Asha K R, [3]Dr. Kavyashree N

[1]*Research Scholar, SSAHE, Tumkur, India shivakumarkb1995@gmail.com

[2]Associate Professor, Dept. of CSE, SSIT, Tumkur, India ashakr@ssit.edu.in

[3]Assistant Professor, Dept. of CSE, SSIT, Tumkur, India kavyashree1283@gmail.com

***Abstract* - Reading a Kannada inscription character in manual is a time-consuming task. Character recognition system using manual takes over a month to identify the character. Over the decades, the character has evolved into different shapes. The archaeology specialists examine each of this character individually to identify a character. Manual techniques are inconsistent. Reading the Kannada stone inscriptions character directly would be time-consuming and inefficient task. For both the public and archaeologists, automating the character identification procedure will be advantageous. This is the primary objective of this research, which focused on developing a method to identify ancient Kannada inscription characters using optical character recognition method. The time period from 8 AD to 12 AD was chosen to limit the research scope of the study. The final outcome of the research study has two elements. When a scanned image of the inscription is entered by the user, the OCR module makes it easy way to identify the characters and converted into modern character. The GIS module is used to provide a map for features that track inscription sites and make it easier for users to view the locations of inscriptions. The OCR solution with the best results was further applied after each one was evaluated.**

Keywords: Kannada inscription character, Geographical Information System, optical character recognition system.

I. INTRODUCTION

In the field of epigraphy, characters from old stone Kannada inscriptions must be translated into modern Kannada characters. The recognition of the inscription character is tedious task because texture of an image posing several challenges like difficulties such as weather-related deterioration and little background and foreground separation. Three sources give rise to technical difficulties in character recognition from ancient stone inscriptions. The first type of visual flaw is caused by natural weather conditions such as wind, rain, lightning, and thunder. Deformation is the second type of distortion, upon the surface of the rock, cracks and dents are considered a natural and time-dependent feature. The third texture is rock surface, which, as a result of inter reflection, takes on the appearance of self-shadow in an image. An experimental study of characters from an ancient Kannada inscription on the stone has been explored in this paper.

With a history spanning over two millennia, the Kannada language is one of the five ancient Dravidian languages. During this time, numerous dynasties, including the Kadamba, Rashtrakuta, Ganga, Chalukya, Hoysala, and Vijayanagara samrajya, ruled over Karnataka. Manuscripts and historical writings have preserved the rich legacy of the empires for generations. The Brahmagiri edict of Ashoka (3rd century B.C.) is considered to be the earliest written record in Kannada, whereas the Halmidi inscription of Kadambas (4th century A.D.) is considered the first stone inscription in Kannada. These inscriptions are typically discovered on copper plates, pillars, coins, stones, and walls of temples.

To shed light on and comprehend the history, culture, socioeconomic standing, administration, and literature of that age, analysis of these inscriptions is crucial. The skill of reading inscriptions on rocks, plates, palm leaves, and other materials is known as epigraphy. Additionally, paleography is the study of old inscriptions and the art of interpreting old manuscripts. A palaeontologist needs to be familiar with the language, relevant historical writing, handwriting styles, and customs of the time.

The inscriptions of Rashtrakutas and Cholas gave historical information and are significant pieces of evidence for the language evolution of a country. Both palm leaves and stone inscriptions include ancient Kannada characters. Those stone engravings are only readable by epigraphists. We require a strong recognition system that can translate ancient character into modern character in order to increase readability and maintain the historical qualities of the past.

Examining the distinctive patterns in inscriptions can help you identify the elements that show how the language has evolved. like the progression and organisation of every letter. The characters from eighth to the twelfth century are conjoint characters in a large number of inscriptions. It will be difficult to identify certain characters in consensual consent because

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 7, pp 154-158, July-2024
https://doi.org/10.47001/IRJIET/2024.807016

each character's end points must be made clear. Partially or completely deteriorated inscriptions can be found. Inscription reading is difficult for a number of reasons, including a shortage of supplies and specialised expertise, according to manual context.

Many studies on optical character recognition in languages like Bengali, English, Tamil, and Kannada have been conducted in the past few years. The majority of study has focused on printed and handwritten character recognition, based on the data acquired. It must be challenging to develop OCR software that recognises script characters. The recognition of Brahmi script characters has been the subject of extensive investigation, as per the Kannada context. There has been no focus on character identification algorithms for Kannada scripts from the past. Therefore, this research gap has to be filled.

The main issues that have prevented more research in this area are the complexity of the ancient Kannada character structure and the access to information found in inscriptions. These days, computer software is more valuable in every industry. For this reason, creating OCR technologies to identify characters in kannada inscription can provide archaeologists new and useful avenues for investigation. This development will also pave the way for creative future study.

Thus, the research topic, "How to develop a modern technological solution to recognise ancient Kannada inscription characters?" would be addressed by this study. This study tackles lots of related issues, including how to recognise the ancient characters using OCR algorithms and solutions that have previously been developed. Which technology is most accurate for creating the solution? How to create an appropriate system architecture that addresses current issues in reading inscriptions. The primary goal of this research is to identify each distinct character in inscriptions and map them into Kannada as it is now utilised. The following list includes the goals of this investigation.

- To recognise the ancient Kannada character in inscription.
  The primary goal of this research is to identify the character variations employed from the 8AD to 12AD. Inscription stamps are employed in data collection. A consistent character set known as the estampages that are meticulously developed by looking through books, archive records, and expert opinions to find the historical record.
- Map ancient Kannada characters onto modern kannada character
  Examine how old Kannada letters are currently being mapped to current Unicode Kannada characters. figuring out a more exact technical method for recognising the old characters from the scanned input image and offer a user-friendly way to show the users the present Kannada characters.
- To keep track of inscriptions exact location.
  Inscriptions exact location is shown. It gives further details about the inscription and instructions on how to get to the location where it is.

In the future, Machine learning will probably become increasingly more crucial for understanding Kannada historical inscriptions. This is a result of machine learning approaches increasingly in powerful and effectiveness. This will enable the development of systems capable of accurately and automatically transcribing and translating ancient inscriptions. This would bring new insights into the history and culture of ancient Karnataka and enable the study of ancient Kannada inscriptions on a far larger scale.

Speed: The process of decoding these inscriptions can be greatly accelerated by using machine learning to automate the character recognition and information retrival process.

Accuracy: A vast range of character forms may be identified by machine learning algorithms, which enhances transcribing accuracy.

Scalability: Deciphering a large number of inscriptions is made possible by the ability of machine learning algorithms to scale to big datasets.

II. LITERATURE SURVEY

Readers will learn important ideas and techniques in the field of inscription character recognition from this part. A great deal of research has been done recently to identify inscription characters in several languages, including Tamil, English, and Kannada. At the moment, a lot of researchers are thinking about using automatic character recognition systems to improve the ease and effectiveness of reading inscriptions. The process of locating and segmenting features in an input image and translating them into contemporary Unicode character form is known as character recognition.

The majority of research has focused on optical character recognition methods for character recognition.

Sachin Bhat *et.al* [1] suggested an algorithmic technique based on optical character recognition. Where a sizable database containing four Kannada eras and associated alphabets was initially assembled. After taking pictures of the stone inscriptions and binarizing them, a segmentation procedure is carried out using the linked component approach. They computed the mean and variance of each row and

ISSN (online): 2581-3048
Volume 8, Issue 7, pp 154-158, July-2024
https://doi.org/10.47001/IRJIET/2024.807016

column of the input image in order to match the captured character with the database image. The similarity is matched based on the mean and variance of each character block by using the absolute difference algorithm.

Using an advanced recognition algorithm, H S Mohana *et.al* [2] has presented a method for identifying the era of Kannada stone inscriptions from the Ganga and Hoysala phases. Typically, this method is applied by first selecting a template, calling the search picture, and then computing the sum of products between the coefficients by simply comparing the template over each point in the search image. It detects the character based on this computed product value. We expanded the approach for many Kannada dynasties spanning different eras in the current paper. Better findings were obtained. The experimental results demonstrated good accuracy in recognising the stone inscription characters of the Chalukya, Hoysala, Kadamba, and Vijayanagara time frames.

SVM classifiers have been used by S. Rajkumar *et al.* [3] to recognise Tamil sounds from stone inscriptions dating back to the seventh century. He has suggested using neural networks to recognise Tamil characters. Character segmentation and recognition are done using global texture analysis as a basis.

Michael Fuchs *et.al* [4] has worked on initiatives that Automated historical document detection using ABBYY Gothic OCR. He has offered an in this paper. Summary of the many components that go into utilising OCR technology to capture old documents. He has first gone over the specifics, covering everything from picture capture to image optimisation. He has now demonstrated the drawbacks of utilising a subpar scanner. Nevertheless, digitising similar things digitally could be challenging when employing top-notch scanners. He has discussed how to analyse the document as the second phase. He has used OCR technology to demonstrate individual character recognition in the third section. OCR technology recognises standard individual characters in printed media by using various software-stored patterns. Considering that, in this segment, Several OCR solutions were mentioned by researchers. They made it apparent in the fourth stage how the user might perform manual after revision. In this research work, the synthesis and export of document formats become the last stage. As a result, it can produce a variety of output formats and options.

Guoying Liu *et.al* [5] Using a CNN model, looked into Oracle Bone Inscription (OBI) recognition. OBIs are the ancient Chinese characters that were inscribed with sharp implements on cow bones or turtle shells. The suggested model has two fully connected layers to get the precise details of OBI instances, and five convolution layers with a 3x3 kernel to extract the features. To attain shift invariance, four pooling functions are used. A 91% testing accuracy has been reported for the model, which used the stochastic gradient descent technique to train the CNN model. However, certain OBIs cannot be correctly categorised.

Chaki and *et.al* (2014) [5] examined the fundamental principles behind binarization techniques for determining the threshold value. They outlined five key methods for binarizing images, which include entropy-based approach, clustering technique, image variance calculation, error measure derived from the optimal threshold, and analysis of contrast variations in images.

Das *et al.* [6] utilized Fast, Independent Component Analysis (FICA) to convert characters from Kannada stone inscription images into binary format, despite the presence of strong correlated noises. To normalize the image, they employed the linear regression method. By determining the cumulative residual entropy and applying a global threshold, they successfully binarized the image, achieving superior outcomes compared to the Otsu method.

Jayanthi [7] has introduced a technique for extracting inscription characters from the background using a blind source extraction method. In this method, each independent component of the source signal is analyzed, and the contrast is enhanced through higher order cumulants. The proposed approach significantly improved the legibility of the text in comparison to both ICA and Fast ICA methods.

Chandrakala *et al.* [8] proposed a retinex technique to enhance digitized estampages of Kannada inscriptions. The length of inscriptions protected on this work is the 11th century. The usage of the retinex approach, the great of the image is stronger by using highlighting foreground characters even in deteriorated estampage snap shots. Ultimately, the digitized great of the estampages is progressed earlier than they're saved inside the database.

Soumya *et al.* [9] made an effort to binarize deteriorated Kannada epigraphic pictures. There are four stages to the work. First, several filters are used to improve deteriorated photos. Second, the Gaussian blur algorithm is used to smooth out the pictures. Thirdly, unsharp masking is used to build a mask of the original images. Finally, spatial operations are applied to the pixels using a Laplacian filter. The Otsu thresholding approach is used to determine the ideal thresholding in order to binarize the images. While the above technique yielded good results, only few datasets were processed.

Document picture binarization became completed with the aid of Lu *et al.* [10] by way of via locating stroke edges.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 7, pp 154-158, July-2024

https://doi.org/10.47001/IRJIET/2024.807016

Bataineh *et al.* is proposed adaptive binarization for degraded document pictures that includes surface comparison versions. Here, the variance cutting-edge the pixels is measured to find comparison variations between pixels. Geometric function extraction is executed by producing geometric and pointer pics. The experimental assessment changed into accomplished on statistical measures to supply better outcomes, as compared to other thresholding techniques

Pannirselvam *et al.* [11] proposed a segmentation set of rules for handwritten Tamil document photos, primarily based on the horizontal and vertical projection projection profile method technique. This method, attempted on various file snap shots, produced green results with a excessive segmentation fee.

Sridevi *et al.* (2012) [12] applied a segmentation set of rules on historic Tamil script file pictures. Segmentation is vital task whilst characters overlap. Hence, the Particle Swarm Optimization (PSO) is applied to segment lines from the script. A mixture trendy related additives and the closest neighbourhood helped segment the characters and achieve exact consequences.

Nabil Aouadi *et al.* [13] proposed a novel technique for textual content line segmentation, chiefly concerning additives from Arabic manuscripts. Segmentation is executed in steps: (1) the localized textual content thing is found, based totally on the interpolation and shape brand new the character, and (2) primarily based on a similar version saved inside the dictionary, the relevant point modern day the text thing is located. This approach retains the skew modern-day the individual and the output is creator-impartial.

Dave *et al.* [14] explained the degrees state-of-the-art segmentation and baseline man or woman extraction, accompanied by word segmentation thru locating the shortest path through the pixel-counting technique. The pixel-counting method outperformed different techniques.

Roy *et al.* [15] proposed a singular method called zone segmentation for Indic scripts (Bangla and Devanagari). in this approach, the script is subdivided into three unique zones - higher, lower and centre - to reduce the variety trendy distinct element instructions. The region segmentation approach better the overall performance modern day the recognition fee using the HMM classifier.

Ryu *et al.* [16] proposed a binary quadratic project approach to find gaps among man or woman words in handwritten file pictures. The irregularity in handwritten characters is trained using a aid Vector gadget (SVM) through structural mastering. The proposed approach changed into tried on Latin and numerous Indian languages and produced state of the artwork performances.

III. RESEARCH GAP

The system primarily carries out the steps of image enhancement, Binarization, and noise reduction in the pre-processing of the old Kannada inscription images. A component of the system that samples the characters from the input document pictures of the Kannada inscription is segmentation. The segmentation of the current ancient Kannada inscription is done using the pre-processing stage, which is primarily created with consideration for the images of ancient inscriptions. The following is a list of the gaps between the current and proposed work that were found after the thorough literature review mentioned above:

- Handwritten and printed document images in multiple languages have been the exclusive focus of character recognition research to date. On the other hand, little work has been put into the inscription images.
- The unrestricted nature of stone inscriptions contributes to their intricacy. It is suggested to use digital acquisition to recognise characters on stone. This cuts down on the amount of time needed and shields the stone from harm caused by etching.
- In stone inscription images, the contrast between the foreground and backdrop is much less than in document images. Therefore, to recover foreground characters, an effective image preprocessing technique needs to be devised.
- A range of segmentation algorithms have been studied in the literature review. Nevertheless, owing to character breakage and very little space between characters, segmenting characters in an inscription is challenging.
- A large number of character sets must be used to train the system in order for it to recognise and transliterate old characters. It is planned to create a machine learning technique that can identify ancient letters, streamline manual transliteration, and outperform the current system in terms of speed.
- The kannada inscription character are recognised and converted into modern language.it will helpful of epigraphers to read the inscription easily.

IV. CONCLUSION

This technology recognises every word in an inscription. The primary constraints were the distinct character structures found in every character and the lack of resources available to prepare the dataset. It will be very difficult to determine the meaning of every word or sentence that has been recognised, as word meanings can vary often or between geographical areas. Thus, conducting the necessary study and putting the

ISSN (online): 2581-3048
Volume 8, Issue 7, pp 154-158, July-2024
https://doi.org/10.47001/IRJIET/2024.807016

system in place at that level will take time. Therefore, putting into practice a solution that gives the script significance might be noted as future effort.

This technique improves the Kannada translation even further. Travellers would find it simpler to discover historical details about Karnataka by reading the Kannada version. This is a clever way to help the tourism sector in Karnataka.

The research should focus on improving the dataset. Only a few numbers of letters were gathered because there were small chances for data access. This situation may change in the future, as archaeologists will be able to quickly upload an increasing amount of recently found material to the database.

REFERENCES

[1] S. Bhat and B. Achar, H.V, "Character recognition and Period prediction of ancient Kannada Epigraphical scripts," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 3, no. 1, pp. 114-118, 2016.

[2] Dr.H.S.Mohana, "Era identification and recognition of Ganga and Hoysala phase Kannada stone inscriptions using advance recognition algorithm", 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT).

[3] S. Rajakumar, V. Subbaih Bharati, "Eighth century Tamil consonants recognition from stone inscriptions", IEEE, 2002.

[4] G. Liu and F. Gao, "Oracle-Bone Inscription Recognition Based on Deep Convolutional Neural Network," Journal of Computers, vol. 13, no. 12, pp. 1442-1450, 2018.

[5] ChakiN, Shaikh, SH & Saeed, K 2014, "A comprehensive survey on image binarization techniques", In: Chaki N (ed) Exploring image binarization techniques. Springer, New Delhi, pp. 5-15.

[6] Das, S & Banerjee, S 2015, "An algorithm for Japanese character recognition". International Journal of Image, Graphics and Signal Processing (IJIGSP), vol. 7, no. 1, pp. 9-15.

[7] Jayanthi, N, Tomar, A, Raj, A, Indu, S & Chaudhury, S 2014, "Digitization of historic inscription images using cumulants based simultaneous blind source extraction". Proceedings of the Indian Conference on Computer Vision Graphics and Image Processing, ACM, p. 51.

[8] Chandrakala, HT & Thippeswamy, G 2017, "Epigraphic Document Image Enhancement Using Retinex Method". In International Symposium on Signal Processing and Intelligent Recognition Systems (pp. 178-184). Springer, Cham.

[9] Soumya, A, & Kumar, GH, 2014, "Pre-processing of camera captured inscriptions and segmentation of handwritten Kannada text". Int J Adv Res Comput Communication Engineering vol. 3 no.5 pp. 6794-6803.

[10] Lu, S, Su, B & Tan Chew Lim 2010, "Document image binarization using background estimation and stroke edges". Int J Doc Anal Recognition 13.4, pp. 303-314.

[11] Pannirselvam, S & Ponmani, S, 2014, "A Novel Hybrid Model for Tamil Handwritten Character Segmentation". International Journal of Scientific & Engineering Research, vol. 5 no. 11, pp. 271-275.

[12] Sridevi, N & Subashini, P 2012, "Segmentation of Text Lines and Characters in Ancient Tamil Script Documents using Computational Intelligence Techniques". International Journal of Computer Applications, vol. 52, no. 14.

[13] Aouadi, N & Kacem, A 2017, "A proposal for touching component segmentation in Arabic manuscripts". Pattern Analysis and Applications, vol. 20, no. 4, pp. 1005-1027.

[14] Dave, N 2015, "Segmentation methods for hand written character recognition". International journal of signal processing, image processing and pattern recognition, vol. 8, no. 4, pp. 155-164.

[15] Roy, PP, Bhunia, AK, Das, A, Dey, P & Pal, U, 2016, "HMM-based Indic handwritten word recognition using zone segmentation". Pattern Recognition, vol. 60, pp. 1057-1075. 75.

[16] Ryu, J, Koo, HI & Cho, NI, 2015, "Word segmentation method for handwritten documents based on structured learning". IEEE Signal Processing Letters, vol. 22 no. 8, pp. 1161-1165.

Citation of this Article:

Shivakumar B, Dr. Asha K R, & Dr. Kavyashree N. (2024). A Survey on Intelligent Kannada Inscription Character Recognition Using OCR and Machine Learning. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 8(7), 154-158, Article DOI https://doi.org/10.47001/IRJIET/2024.807016

****

