

Sinhala Inscription Character Recognition Model Using Deep Learning Technologies

Shashika Ruwanmini, Kapila Dias, Clera Niluckshini, Terrance Nandasara

Abstract— Manual inscription character reading is a time-consuming task. The identification of a character takes nearly one month in the manual process. Characters have evolved into different shapes over the centuries. Archeology experts analyze all these shapes one by one to recognize a character. Reading inscriptions directly using manual procedures would be time-consuming and inefficient due to a lack of consistency. Automating the process of character recognition will be a huge advantage for both archeology experts and the general public. This is the main objective of this research, which focuses on developing a solution using an optical character recognition module to recognize ancient Sinhala inscription characters. The period from 10 A.D. to 12 A.D. was selected to limit the scope of the study. The final output of the research study has two components. The OCR module facilitates proving the recognized characters when the user inputs a scanned image of the inscription. The GIS module is used to present a map for inscription site tracking features that facilitate users' visits to the locations of inscriptions. Mainly, three OCR solutions were developed based on template matching, artificial neural networks (ANN), and convolutional neural networks (CNN). After evaluating each OCR solution, the best-result OCR solution was further implemented.

Keywords—*Optical Character Recognition, Estampages, Sinhala Inscriptions, Convolutional Neural Network, Artificial Neural Network, Geographical Information System*

I. INTRODUCTION

The city called “Anuradhapura” has been recognized as one of the great cultural heritage places in Sri Lanka. There are many inscriptions found in Anuradhapura.

Shashika Ruwanmini is from Institute of Technology University of Moratuwa. (e-mail: ruwanminis@itum.mrt.ac.lk) Clera Niluckshini is from Eastern University of Sri Lanka. (e-mail: saranilu94@gmail.com). Kapila Dias and Terrance Nandasara are from University of Colombo School of Computing. (e-mail: gkad@ucsc.cmb.ac.lk and nst@ucsc.cmb.ac.lk)

DOI: <http://doi.org/10.4038/ict.v16i1.7239>

© 2022 International Journal on Advances in ICT for Emerging Regions

A large number of inscriptions have been found from 10 A.D. to 12 A.D. [1] Stone Book inscription, Mirror Wall, Thonigala inscription are the most popular examples of them.



Fig 1: Thonigala Rock Inscription

The Sinhala inscription provided information about history, and it is important evidence for the language evolution of a nation.

The factors that reveal the evolution of the language can be identified by analyzing the characteristic patterns in the inscriptions. Such as the development of each letter and its structure. The characters in many of the inscriptions found from the 9th to 13th centuries are conjoint consonants. It is believed that these characters evolved through Mahayana Buddhism. Recognizing individual characters in the conjoint consent will be a challenging task since the end points of each character have to be clearly identified. Some of the inscriptions are partially or fully decayed. According to the manual context, inscription reading is a challenge due to several reasons, such as a lack of necessities and specialized knowledge.

In languages like Sinhala, Pali, Tamil, Bengali, and English, there has been a lot of research on optical character recognition in the past few years [11]. According to the information gathered, most of the research has been done on printed and handwritten character recognition. Developing OCR tools to recognize script characters must be a challenging task. According to the Sri Lankan context, a considerable amount of research has been undertaken to recognize Brahmi script characters. Nobody has focused on developing character recognition systems for ancient Sinhala scripts. Therefore, this research gap needs to be filled.

The complexity of ancient Sinhala character structure and the limitations of access to inscription information are major problems that have caused a lack of research in this field. Nowadays, computer software provides greater value in every field.



This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Therefore, developing OCR tools to recognize characters in Sinhala script can provide different and valuable opportunities to researchers in the archaeology field. Such a development will open the door for innovative future research as well.

Therefore, this research will answer the research question, "How to develop a modern technological solution to recognize ancient Sinhala inscription characters?". This study addresses a set of sub-problems, such as: how can we use already developed OCR solutions and algorithms to identify the ancient characters? What is the most accurate technology to develop the solution? How to make a suitable system design that overcomes existing problems in inscription reading. Recognizing each individual character in inscriptions and mapping them into currently used Sinhala is the main purpose of this research. The objectives of this research are mentioned below.

- To recognize the ancient characters.
Identification of the character variations used in the 10th A.D. to 12th A.D. is the main focus of this objective. Estampages of inscriptions are used for data gathering. A constant character set referred to as the estampages which is created with great effort by searching the historical record in books, archival documents, and experts' judgements.
- To map ancient characters with modern Sinhala characters.
Study the current process of mapping ancient Sinhala letters with modern Unicode Sinhala alphabetic letters. Finding a more accurate technological approach to identify the ancient characters from the scanned input image and provide a user-friendly solution to display the modern Sinhala characters for users.
- To track the geographical location of inscriptions.
The geographical location of inscriptions is displayed. It provides guidance on the directions to reach the place where it is located and presents more information on that inscription.

II. RELATED WORK

This section shares knowledge on significant concepts and methodologies related to inscriptions' character recognition domain with the readers. In the recent past, a considerable amount of research has been undertaken to recognize inscription characters in different languages such as Bengali, Sinhala, Tamil, Pali, English, etc.

Currently, many researchers have considered implementing automated character recognition processes to achieve more convenience and efficiency in the inscription reading process. Character recognition is the process of identifying segmenting features in the input image and mapping them into ASCII or another modern Unicode form. Most of the research has been done to recognize characters using optical character recognition techniques.

M. Merline Magrina (2019) [1] has presented a model for recognizing Tamil characters from stone inscriptions using Convolution Neural Network (CNN) architecture and Unicode mapping. From the preprocessed images, characters have been segmented using the bounding box technique and passed to the 14-layer CNN model for feature extraction and classification via the SoftMax layer. Following the classification, characters are recognized by mapping them to a Unicode value, where a numeric value is

assigned for each character. The researcher concluded that this model is a state of art technique that can be applied to other languages as well since the recognition rate was 99.05%.

Yasir Babiker Hamdan and Prof. Sathish (2021) [2] have used OCR-based algorithms for training the model for classification and recognized handwritten text. They proposed a statistical based SVM framework that relies on a kernel-based learning approach that uses principal component analysis for feature extraction. This approach was able to achieve high precision compared to the template matching method, the structural pattern acknowledgement model, and the statistical method. The key factor to notice in the model, it was able to recognize italic and stylish types of text and numeral handwritten characters with higher accuracy than the above stated other approaches.

Saad Bin Ahamed et al. (2019) [3] have published a novel approach to recognize handwritten Urdu characters with a transfer learning method. The MNIST pre-trained network is used for transferring learning experience to the Urdu dataset. This research is highly aimed at the Multi-dimensional Long Short-Term Memory (MDLSTM) network architecture, which has a robust training algorithm accompanied by transfer learning. The most important fact we learned from this model was the transfer learning ability of cursive script, which is highly relatable to our inscription content. This research proved that the enclosure of transfer learning increased the potential of LSTM networks, which leads to state-of-art in data analysis.

Lalitha Giridhar et al. (2019) [4] have developed a model consisting of an OCR scanner to capture the images of the inscriptions. They binarized the images, segmented the images into character blocks, and passed them to the CNN model for training and classification. Tesseract is used for character recognition in order to generate the digitized modern-day Tamil characters from the inscriptions, and it has been converted to audio output format. But the outcome from the model cannot be digitally segmented because of absence of any language parser for ancient Tamil scripts.

Wichai Puarungroj et al. (2019) [5] have discussed a CNN based model for recognizing Thai Noi characters in palm leaf manuscripts. The captured image of a palm leaf manuscript was preprocessed and segmented, and the dataset was created for training with 2600 images, which have 100 images per 26 consonants. The research experiment was carried out with two convolutional models which are Inception-v3 and Inception-v4, using a 10-fold cross-validation design. The accuracy of both was respectively, 76% and 73% due to the lack of training data for each consonant and the quality of the image.

Guoying Liu and Feng Gao (2018) [6] investigated the recognition of Oracle Bone Inscription (OBI) through building a CNN model. OBIs are the ancient Chinese characters written on the cattle bones or shells of turtles using sharp objects. The proposed model consists of five convolution layers with a 3x3 kernel to extract the features and another two fully connected layers to obtain the definitive details of OBI instances. Four pooling functions are used to achieve shift invariance. The model has deployed the stochastic gradient descent technique for training the CNN model and has been presented with a testing accuracy of 91%, but some OBIs cannot be properly classified.

In the research of Kannada character recognition and its period prediction work conducted by Sachin Bhat and Achar (2016) [7], they proposed a model with OCR based algorithmic process. Where they first created a large database with four Kannada periods and related alphabets. The images of stone inscriptions are captured and subjected to binarization and then using a connected component method, a segmentation process is done. To match the captured character with the database image, they calculated the mean and variance of each row or column of the input image. By applying the absolute difference algorithm, the similarity is matched based on the mean and variance of each character block. In the paper presented by Nikhi, Vijayakumar and Kolkure, (2017) [8], with an objective of digitizing hard copies of historical books since they can be damaged in the long term, they have proposed a model with key OCR techniques associated with acquisition of images, preprocessing, extraction of feature, classification, and post processing. With that purpose the researchers have provided an overview of OCR components briefly.

The paper submitted by Shashank (2015) [9], explains an artificial neural network methodology to recognize the visual characters. The concepts of neural networking and pattern recognition have been well executed in this domain. The paper shares knowledge relevant to image digitizing, neural network learning mechanisms, details on the architecture implemented, performance constraints of the model, level of accuracy, and computational complexity. So, the paper highlights that neural networks provide many benefits in pattern recognition and classification.

Hugo Pires and et al. (2015) [10] have submitted a model, where 3D scanned data sets were used to reveal invisible facts from archaeology sites. They have taken advantage of raking light photography techniques, PTM (Polynomial Texture Maps) and 3D scanning techniques. A Latin inscription found in a Roman sanctuary and engravings from a rock art in Monte Faro are taken into consideration in this research, and they were highly damaged due to erosion and other environmental factors. The research has contributed techniques that can be used for 3D data processing and visualization.

Michael Fuchs (2017) [11] has conducted projects with ABBYY Gothic OCR for the automated detection of historical documents. In this paper, he has presented an overview of different elements that play a part in capturing historical documents using OCR technology. As the first step, he has explained details from the image capturing to image optimization. At this stage, he has illustrated the disadvantages of using a low-quality scanner. However, using high-quality scanners may cause difficulties in digitizing analogous materials. As the second step, he has mentioned how to analyze the document. In the third section, he has illustrated individual character recognition using OCR technology. OCR technology uses different patterns saved in software to recognize standard individual characters in printed sources. Based on that, in this section, researchers mentioned different OCR solutions. In the fourth step, they clearly mentioned how the user can do manual post correction. Synthesis and export of document formats become the final step in this research work. Hence, it can generate different output formats with various options.

There is another OCR (Optical Character Recognition) software engine originally developed by HP between 1985

and 1995 that is now sponsored by Google Projects (Google Tesseract). Also known as Google's Optical Character Recognition (OCR) software, it now works for over 248 world languages (including all the major South Asian languages). The technology being used to extract the text from the images, scan the printed text, including handwritten text, can extract text images from old books, manuscripts. Google OCR, which is launched as open software with the dependency of Tesseract is the optical character recognition system that is being used in Google Books. The community project of this OCR development was taken over by Google in 1995, and now Tesseract is the most successful OCR engine working across 60 languages. During the processing, it maintains the basic formatting options (italic, bold, strong, line breaks, etc.) without any alterations. Anyhow, some other formations like bullets of numbering, tables, lists, text columns, footnotes, and endnotes sometimes get altered.

III. METHODOLOGY

A. Preparing Training Dataset

The research project creates a responsive web application with a GIS module for tracking locations and an OCR module for reading characters. The GIS module gives information about the inscription sites in Anuradhapura, as well as pictures and more historical information.

When doing this research, thousands of estampage collections were captured with great difficulty. A few experts in this specific industry were consulted to obtain more information. Estampages are obtained from the archaeology department in Sri Lanka.

Between the 10th and 12th centuries, 35 letters in the alphabet were used with different character variations [9]. Approximately 3500 data which follow 100 data for one character are nearly needed to train the OCR engine. The limitation in collecting the data was a challenge to achieving a high accuracy rate for the project. During the data collection phase, only 850 data were collected for some mostly used characters. Such as “ක”, “ග”, “ඵ”, “ම”, “භ”, “ල”, “ය”, “ඝ”, “ඞ”.



Fig 2: An Estampage of an Inscription

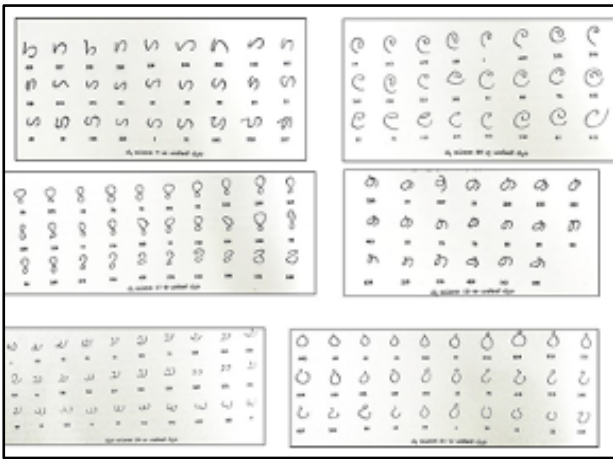


Fig 3: Character Variations [12]

To make it simpler for the system to convert the estampage document into a format that can be read by a computer, the image data must first be digitalized. Here we digitize image data into two formats, namely "bitmap" and "jpeg" because different tested OCR solutions require different formats. For example, an ANN-based OCR solution requires images in bitmap format, whereas a CNN-based OCR solution requires images in jpeg format. The document images are prepared first with a variety of font sizes according to the OCR requirements at 300 DPI quality.

The image processing module was developed using the OpenCV python library for the purpose of preprocessing all digitized images, and it was used to test three OCR techniques.

In the above example, the Intel image processing library was used because it usually works with IplImage data types. The Image Processing Library inherits the actual version of OpenCV [10]. Hence, IplImage has been used from the very beginning. More powerful image processing functions in OpenCV are used to implement the image processing module. Such as the cvLOADImage, cvNamedWindow, and cvShowimage functions were used to capture and show images. The process in the image processing modules has been supported by several functions. Such as cvLoadImage, which loads the image file and returns an IplImage pointer to the file. The description of the window for display is supported by CvNameWindow. cvShowImage displays the image in a specific window. Noise removal should play an important role in image processing. In this research, noise removal is considered a set of tasks for removing black pixels that are not normally supposed to be part of the document but are introduced to the document image under different conditions. In that sense, the dataset contained different types of noise, which required preprocessing to obtain a better recognition rate.

Scanned estampage images are in black and white. Subsequently, images are converted to grayscale. A grayscale digital image is a single sample that brings only acuity information about the image. For this purpose, the estampages used in this research are composed exclusively of shades of gray, varying from black at the weakest acuity to white at the strongest. The black and white scanned image is converted into a grayscale, which gives 256 possible shades of gray from black to white.

Additionally, de-noising the captured images is necessary in the preprocessing process. Noise removal was done by using a process called "Blur," which can be achieved in different ways. The median filter is a kind of nonlinear smoothing technique, and it is considered a basic filter. In this research, the median filtering method is used for the blurring process, and the gray value of every pixel is fixed to a middle value. OpenCV includes a median filter that can be applied to an image by calling the cvSmooth function.

Smoothing is one way to test how the rate of recognition changes for a set of smoothed data. The conversion of a color image to a bi-level one is called binarization, and it can be understood as a classification between the character and background of an image. The binarization process supports separating the image skeleton from the background. This research focuses on completing image binarization processing as much as possible to obtain image information. Additionally, thresholding should play a very important role in setting the correct threshold value, which will determine a pixel as an object or a background, and whose gray level of 0 indicates that the value of the pixel is less than the fixed threshold value, and whose gray level of 1 denotes that the value of the pixel is more than the predefined threshold value. OpenCV's cvThreshold function was used to segment the character from a background.



Fig 4: Sample Real Image Set

According to our observations, several factors affected the quality of the image. According to this research, obtaining clear and complete estampages must be required. Partially erased or incomplete original sources had to be abandoned to get better OCR results. The scanning resolution should be 300 dpi or above to capture as much image information as possible. Setting the DPI lower than 200 produced unintelligible results and setting it higher than 600 dpi increased the size of the stored file but did not produce much better results.



Fig 5: Sample Preprocessed Image Set

Segmentation is highly challenging, and the impact may differ based on the character structure. According to this research, we are mainly focused on ancient Sinhala characters, which have a more curved structure than lines and strokes. According to similar studies done in this domain, it will be quite easy to build an OCR engine for English, Tamil, and Brahmi characters due to the lack of segmentation errors [5]. According to our observations, there can be so many reasons that cause segmentation problems, such as letter variations, different shapes of the same character, conjoined characters, and complex shapes of individual characters. Based on the complex structure of these characters, there may be some major segmentation problems, such as over segmentation of the basic structure of the character. Ex: “ඔ” recognized as “ඔ”.

B. OCR Evaluation Workflow

For each character, a training data set and a test data set were made by splitting the cropped character dataset into training and test data sets. 90% of the images from the cropped dataset were used as the training data set, and 10% of the images from the cropped data set were used as the test data set for an individual character.

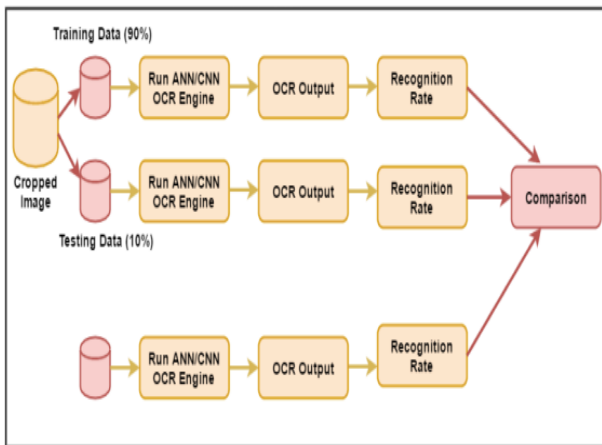


Fig 6: OCR Evaluation Workflow

All the images in the cropped set were preprocessed, and the recognition rates of the three data sets were calculated separately. Subsequently, we obtained the recognition rate for real images with noise. Finally, compare the recognition rates of all three datasets to determine which OCR engine performed well.

C. Tested OCR Techniques

1) *Template Matching*: In the template, the pixel values are compared with the pixel values in the sub-region. If the match score is near the one that enables the object to be present in the image with a close match, then the highest matching is recognized based on the similarity between the region and template with the same size in an image.

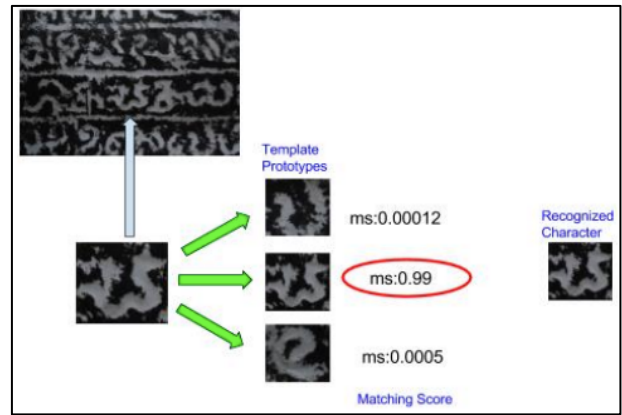


Fig 7: An Illustration of Template Matching

A database that has several images is maintained. The feature detection step is carried out with the assistance of SURF detection. Then extract character features based on the previously designed templates. In this research, deformable templates were used to describe and detect features of the templates. Additionally, the features of interest, like lines in a deformable template and its curve, are described by using parameterized templates. The specialty of the parameterized templates is that allow a priori knowledge about the expected shape of the features. This will be important to guide the detection process.

For the recognition process, template matching means finding either an exact match or the closest match between the template of the input character and the input character. Therefore, the current input character is compared to each template to find the comparison result. When an example is provided for the comparison, if input is denoted by I and $I(x,y)$ is considered the input character, the template is denoted by T , and $T(x,y)$ is the template. The matching function $S(I, T)$ will return a value, which indicates how well the template T matches the I input character.

According to that, character recognition can be done by recognizing when T gives the best value of the matching function of $S(I, T)$. The important point to note here is that this method can only be successful if the input character matches the templates of the same or similar structure.

2) *Artificial Neural Network (ANN)*: ANN-based OCR is a free library that is used to implement ANN-based OCR. The network is trained to recognize 30 characters. The training sample for the conversion is to create a vector of size 100, and this is supported for the converting input part of the training sample. "0" and "1" correspond to the pixel by following "1" in all positions of the letter pixel and "0" in all positions of the background pixels.

In the training process, the network responds with a target output for a specific input, followed by supervised learning. Each training sample is represented as a possible input and target output for the corresponding input. After the training process is done, a new input can be given to the network, which then produces the output data.

Since the CNN implementation used a multi-layer neural network, this OCR implementation only used one layer. 100

inputs have been included in the single layers, corresponding to the size of the input vector.

The important point to note here is that the size of the output vector (30 neurons) has been included in the layer. All samples from the training set are trained through the optical character recognition engine, and error is calculated for each training epoch. The training is done, and the network can be used for recognition when the error becomes less than the defined error limit. 12 iterations were run for recognition, and the output is shown as computation code. For example, "character is represented as "a40", "a41".

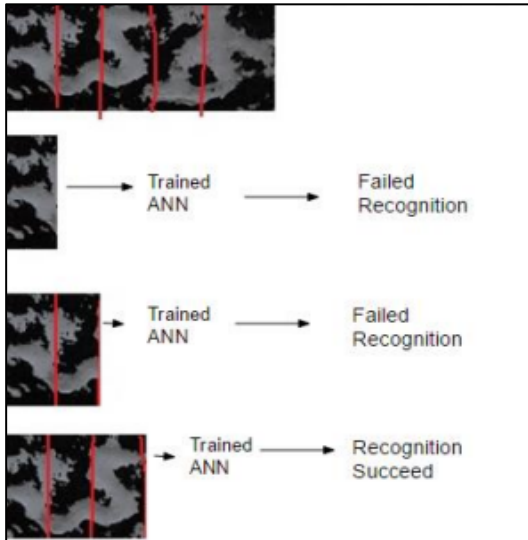


Fig 8: ANN Segmentation Method

3) *Convolutional Neural Network (CNN)*: The convolutional neural network is a perceptible learning model that extracts and learns suitable features. Based on the character lines and curve properties of different scripts, a script can be recognized by CNN based on the shape of its individual characters. The evaluation results are based on these discriminative shape features recognized by CNN, and this is considered an efficient approach. The approach gives good results on different target documents that are not part of the training process. In addition to that, for evaluation purposes, real images with noise were used with the same number of quantities as the cleaned image sample.

The CNN OCR engine requires a lot of time to complete the whole training process. Therefore, someone can say that ANN is better than CNN. This simple version of convolutional NN (that got a 5% error rate) can be implemented in about 2-4 hours, depending on how familiar we are with it. Therefore, the training time of the OCR engine will be quite flexible depending on different factors.

D. System Implementation

The main interfaces of the prototype system are shown below. The interface has been kept as minimal as possible, with only the required items being displayed.

Figure 9 represents the home page of the website, along with the slider bar and two main navigation sections for the OCR module and the GIS module.



Fig 9: Home page of the system

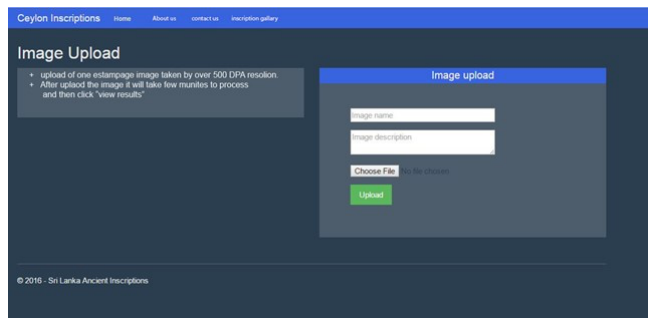


Fig 10: Interface of Image Uploading Function

Figure 10 shows the OCR module interface used to extract characters from estampage by uploading a photograph of the document. It will take a few minutes to process and retrieve the character after uploading. Figure 11 depicts the interface once the process has concluded, and the results are displayed.

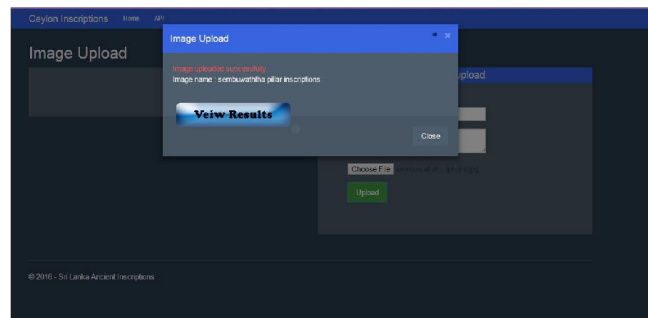


Fig 11: OCR Image Upload & Review the Results

Figure 12 shows the outcomes of a specific inscription image. After viewing the results, both the input images and the identified output are displayed. Unrecognized characters are represented by a dash (-). However, in the future, when new data samples are discovered, additional characters will be added to the training database.



Fig 12: Output Result of OCR Module

1) *Inscription Tracking Module*: A web application was developed to assist users in locating and tracking inscription sites, provide detailed information, and summarize formats. The interpretation of inscriptions at inscription sites can be provided in two ways. The first method involves displaying the history and translated meaning of the inscription, while the second method involves uploading a scanned image of the stamp and displaying the identified character set. The 2D map provides comprehensive information regarding the various inscription sites. This is a type of regional geographic map depicting the location of inscriptions from the time period of our study. Figure 11 displays screenshots of the developed web application, which describe the location details and summarize information for the Galpotha Stonebook inscription.

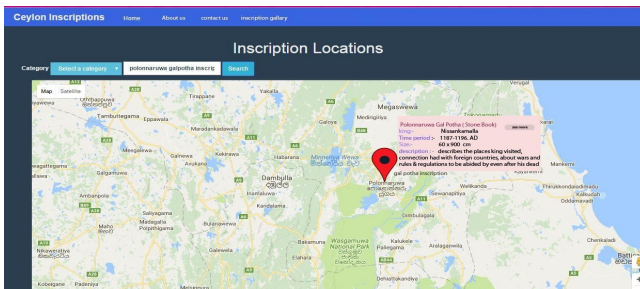


Fig 13: Location Details of “Galpotha” (stone brook) Inscription in Summarized Information

This GIS system also contains a search module to filter out sites by different categories. This will be an advantage for users to search inscription sites by inscription name, Inscription location, nearby inscriptions, district or region, time period, and based on ancient kings’ names.

Figure 14 shows the searching categories for inscriptions.

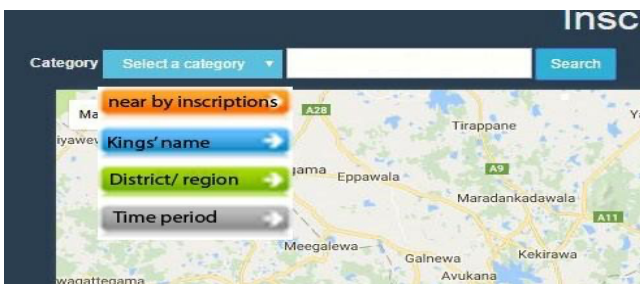


Fig 14: Searching Categories for Inscriptions

2) *Administrative System*: The GIS Admin System is a Windows desktop application that is used for data input for

the Inscription Web Application. The GIS administrator can directly access the inscriptions database, and only admin users have the privilege to access the GIS admin module. Therefore, admin users can create, edit, and remove GIS data using the GIS Admin application. Figure 15 represents one of the interfaces of the GIS administration system.

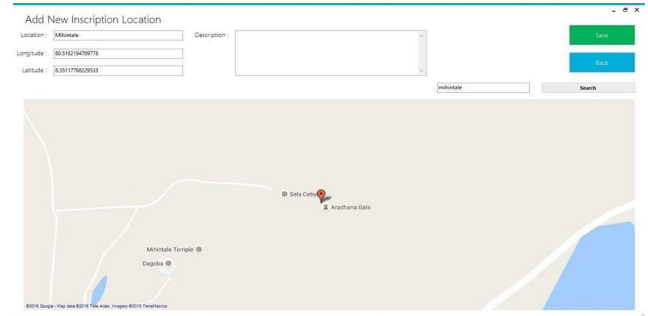


Fig 15: GIS Admin System

The technology used for GIS administration is the.NET stack with C# and SQL Server. Database tables are mapped into objects using the Entity Framework, which gives fast and accurate access to the Inscriptions database. The user can search location data to the 13th decimal point; therefore, it gives the most accurate location data.

IV.RESULTS AND FINDINGS

The evaluation is carried out on two disjoint subsets. The division of the training data set and testing data set is described above in Section 3.2, OCR Evaluation Workflow.

For each character, a training data set and a test data set were made by splitting the cropped character dataset into training and test data sets. 90% of the images from the cropped data set were used as the training data set, and 10% of the images from the cropped data set were used as the test data set for an individual character.

The training set was used for training purposes, and the testing set was used to verify that the performance of the OCR engine was at an acceptable level. The “Ramba Vihara” script is used for further verification of OCR engine performance.

During the testing phase, we utilized two components. Through the image preprocessing module, the pre-processed test data transformed the image into a clean image. In contrast, we obtain authentic test data that has not been OCR-processed and thus contains more noise. The primary objective is to evaluate the impact of noise and character structure/layout on character recognition quality. The evaluation was based on both the performance of each individual character and the overall script. The score is derived from the equation, which serves as the basis for calculating the individual character recognition rate using the following formula.

$$Score = \frac{TNC - TNE}{TNC} \tag{1}$$

Above mentioned accuracy score (1) represents the value of followings,

TNC- Total Number of Characters in dataset

TNE- Total Number of Errors (How many characters were inaccurately predicted)

According to our observations, some of the ancient Sinhala characters have a more basic, similar structure. Similarly, the letter "ඞ" and "ඞ". Both have two curves based on the same basic structure. In such cases, the OCR system gets confused and cannot segment a complete basic

The most important and vital part of making an OCR that works well is segmentation, because classification depends on this step. During the training stage of the system, various possible segments were generated. Most of the possible segments are trained, but test data may have some noise or distortion to generate some new segments. Similarities between these characters' shapes caused recognition failure. The purpose of showing the following summary of comparisons is to give an idea of how small changes in character shapes will impact character recognition success.

A. Experimental Results

In the case of an ANN-based OCR solution in Table 1, the training dataset had 850 training images. For each iteration, 10% of the data (85 images) is used for validation, and 90% (765 images) is used for training. A photograph of a real grid was taken, and the numbers were then manually identified. Additionally, real images (85 images) with noise samples were used in the same quantity as samples of cleaned images for the purpose of evaluation.

The CNN based OCR solution in Table 2, used a training dataset that has 850 training images. For each iteration, 10 percent of the data (85 images) is used for validation and the remaining 90 percent (765 images) for training. A picture was taken of a real grid, and then numbers were manually recognized.

Both OCR have shown less recognition rate for test data set of real images with noise which is presented in Table 3.

character. Sometimes the poor quality of the uploaded image causes over-segmentation of characters. In this research, we addressed some of these problems in the OCR process during segmentation and classification.

TABLE 2
CNN BASED OCR PREDICTION RESULTS (INDIVIDUAL CHARACTER WISE)



















Character	Samples for Training	Samples for testing (Cleaned Images)	Sample for testing (Noisy images)	Accuracy (%)
	120	12	12	93%
	100	10	10	88%
	70	7	7	80%
	80	8	8	88%
	110	11	11	82%
	75	7	7	80%
	70	9	9	82%
	70	7	7	81%
	70	7	7	82%

TABLE 3
RECOGNITION RATE FOR THREE DATA SET

	Training Data set		Test Data (Preprocessed)		Test Data (Real Images)	
ANN	765	85%	85	80%	85	45%
CNN	765	95.16%	85	92%	85	65%

TABLE 1

ANN BASED OCR PREDICTION RESULTS (INDIVIDUAL CHARACTER WISE)

Character	Samples for Training	Samples for testing (Cleaned Images)	Sample for testing (Noisy images)	Accuracy (%)
	120	12	12	82%
	100	10	10	72%
	70	7	7	55%
	80	8	8	60%
	110	11	11	62%
	75	7	7	65%
	70	9	9	40%
	70	7	7	71%
	70	7	7	42%

B. Summary of Comparison

Tables 4 and 5 represent the OCR prediction results for the input image from the "Ramba Vihara" script. Both OCR engines have failed to recognize some characters in the input image, which were not given to the OCR engine due to a lack of data.

Several ancient Sinhala characters share a more fundamental structure. Similarly, the letters "ඞ" and "ඞ". Both have two curves based on the same basic structure. In such cases, the OCR system becomes confused and is unable to segment an entire basic character. Occasionally, the poor quality of the uploaded image causes excessive character segmentation. In this study, we addressed some of these issues during segmentation and classification in the OCR process. Segmentation is the most vital and important aspect of designing an effective OCR, as classification success is contingent on this step.

TABLE 4
RESULTS OF SAMPLE TEST IMAGE 1


Input Image		
Expected Output	කොව	
	ANN	CNN
OCR engine output	-ක-	කෙප
Recognition Rate	20%	46%
Post evaluation score	$((3 - 3) / 3) * 100 = 0\%$	$((3 - 2) / 3) * 100 = 33.33 \%$

TABLE 5
RESULTS OF SAMPLE TEST IMAGE 2


Input Image		
Expected Output	වහලස	
	ANN	CNN
OCR engine output	-හ-ස	පහලස
Recognition Rate	52%	82%
Post evaluation Score	$((4 - 2) / 4) * 100 = 50\%$	$((4 - 1) / 4) * 100 = 75\%$

TABLE 6
EXPERIMENT RESULTS OF CNN OCR SOLUTION

Trial	Epoches	Accuracy	Output
First 2 Letters	589	98%	ක-ක ප-ප
First 4 Letters	589	96.16%	ක-ක ප- ප ස- ස ග- ග
First 6 Letters	589	93%	ක-ක ප - ප ග- ග ස- ස හ- ග ය- ය
First 9 Letters	589	90%	ක-ක ප - ප ග - ග ස - ස හ - ග ය - ය ම-ම ල & ක Not Recog -nized

TABLE 7
OVERALL ACCURACY ON EXPERIMENTS

	ANN Accuracy	CNN Accuracy
Cleaned images	85%	90.25%
Real Images	34%	45.12%
Smooth Images	93%	95%

V. DISCUSSION

This research is focused on finding a suitable approach to recognizing ancient Sinhala characters from 10 A.D. to 12 A.D. For that purpose, three OCR techniques were tested: template matching, artificial neural network (ANN) and convolutional neural network (CNN) for preprocessing, feature extraction, and character recognition. From the results, the CNN-based OCR solution is identified as the suitable solution to recognize ancient characters more accurately by considering its recognition rates. Additionally, the geographical tracking module of inscriptions, with the inscription details, is integrated with the final system.

From the results obtained during the series of approaches, it reveals that the CNN-based OCR model gives a better recognition rate than the ANN-based OCR and template matching for all the types of feature extraction.

The CNN recognition rate has shown the highest recognition rate of the three data sets used. As described in the above sections, the training data set and test data set for each character were prepared by dividing the cropped character dataset using estampages. 90% of the images from the cropped dataset were used as the training data set, and 10% of the images from the cropped data set were used as the test data set for an individual character. The highest recognition accuracy is from CNN, which is 95.16% from the training dataset, 95% from the validation set, and 95% from the test data. The calculation formula is described in the above sections.

So, the proposed OCR module for an ancient character recognition system is developed using the CNN approach due to its high accuracy rate. Another problem we mentioned was a lack of information about inscriptions that is available to the public. So, the people must visit inscription sites or museums to look at that inscription information. To address that problem, we have implemented a web application that guides us to the places by tracking the inscription sites and providing inscription information in both detailed and summarized formats. So, archeology researchers and the public can get detailed information [12] about archeological places while they are traveling.

Ancient character recognition systems are still a burning research area in optical character recognition. Each and every step contributes directly to the accuracy of the system, like preprocessing, segmentation, feature extraction, training methods, etc. So, all these areas are open to independent research. In each step, a lot can be improved. There are limited systems available in the world, such as in India, Greece, and China. But there is no current system available for ancient character recognition in Sri Lanka.

VI. CONCLUSION AND FUTURE WORKS

Currently, this system identifies each word or small segment including less than four characters. The major limitation was the different character structures in each character and the scarcity of resources to prepare the dataset. Obtaining the meaning of each recognized word or sentence will be a hugely challenging task, and the meaning of a word can be different from time to time or from region to region. Therefore, it will take time to do the research and implement a system at that level. Hence, implementing a solution that provides meaning to the script can be identified as a future work.

The methods mentioned above raise the quality of the English translation even further. Reading the English version will make it easier for tourists to learn historical facts about Sri Lanka. This is a creative method to support Sri Lanka's tourism industry.

Enhancing the dataset should be addressed for future studies. Due to the limited opportunities for access to data, only a few letters were collected. In the future, this circumstance can be altered, and archaeological specialists will be capable of rapidly expanding the database by adding more and more newly discovered data.

REFERENCES

- [1] M. M. Merline and M. Santhi, "Ancient Tamil character Recognition from epigraphical inscriptions using image processing techniques," *Journal of Telecommunication Study*, vol. 4, no. 2, pp. 40–48, 2019.
- [2] B. Hamdan and A. Sathesh, "Construction of Statistical SVM based Recognition Model for Handwritten Character Recognition," *Journal of Information Technology and Digital World*, vol. 3, no. 2, pp. 92–107, Jun. 2021, doi: 10.36548/jitdw.2021.2.003.
- [3] M. Husnain et al., "Recognition of Urdu Handwritten Characters Using Convolutional Neural Network," *Applied Sciences*, vol. 9, no. 13, p. 2758, Jul. 2019, doi: 10.3390/app9132758.
- [4] L. Giridhar, A. D. And, and V. Guruviah, "A Novel Approach to OCR using Image Recognition based Classification for Ancient Tamil Inscriptions in Temples," *arXiv (Cornell University)*, Jul. 2019, doi: 10.48550/arxiv.1907.04917.
- [5] W. Puarungroj, P. Kulna, T. Soontarawirat, and N. Boonsirisumpun, "Recognition of Thai Noi Characters in Palm Leaf Manuscripts using Convolutional Neural Network," *ResearchGate*, Nov. 2019, [Online].
- [6] 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.
- [7] 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.
- [8] P. Agvale and V. Kolkure, "Comparative Study of Different Image Feature Extraction Algorithm and Representation Techniques", *AJCT*, vol. 1, no. 1, Dec. 2017.
- [9] S. Araokar, "Visual Character Recognition using Artificial Neural Networks," *MGM's College of Engineering and Technology, University of Mumbai, India*, 2005.
- [10] H. Pires, J. A. Rubio, and A. E. Arana, "Techniques For Revealing 3d Hidden Archeological Features: Morphological Residual Models As Virtual-Polynomial Texture Maps," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-5/W4, pp. 415–421, Feb. 2015, doi: 10.5194/isprsarchives-xl-5-w4-415-2015.
- [11] K. Yu, M. Kim, and J. R. Choi, "Memory-Tree Based Design of Optical Character Recognition in FPGA," *Electronics*, vol. 12, no. 3, p. 754, Feb. 2023, doi: 10.3390/electronics12030754.
- [12] R. Samaveda, "The Development of Sri Lankan Epigraphy" *Postgraduate institute of Archaeology University of Kelaniya*, vol. 81, no. 1, pp. 91–100, 2016.
- [13] P. K. Charles, V. Harish, M. Swathi, and C. H. Deepthi, "A review on the various techniques used for optical character recognition," *Ijera.com*. [Online]. Available: https://www.ijera.com/papers/Vol2_issue1/DB21659662.pdf.
- [14] M. Auer, "3D-Sutras: A web-based atlas of laser scanned Buddhist stone inscriptions in China," 2010.
- [15] S. G. Dedgaonkar, A. A. Chandavale, and A. M. Sapkal, "Survey of methods for character recognition," *Ijeit.com*. [Online]. Available: https://ijeit.com/vol%201/Issue%205/IJEIT1412201205_36.pdf.
- [16] A. M. Sabu and A. S. Das, *A Survey on various Optical Character Recognition Techniques*. 2018. doi: 10.1109/icedss.2018.8544323.
- [17] R. B. Rustamov, *Geographic Information Systems in Geospatial Intelligence*. IntechOpen, 2020. doi: 10.5772/intechopen.84925.
- [18] "Epigraphy | Sri Lanka Archaeology," Oct. 13, 2020. <https://www.archaeology.lk/category/epigraphy/>
- [19] S. Budha, N. Pant, and B. K. Bal, "Nepali OCR Project Research Report 2.0," *www.academia.edu*, Apr. 2018, [Online]. Available: https://www.academia.edu/35122109/Nepali_OCR_Project_Research_Report_2_0
- [20] Y. M. Alginahi, "Preprocessing Techniques in Character Recognition," in *Sciyo eBooks*, Sciyo, 2010. doi: 10.5772/9776.

- [21] H. Mara, J. Hering, and S. Kromker, "GPU based optical character transcription for ancient inscription recognition," in 2009 15th International Conference on Virtual Systems and Multimedia, 2009.
- [22] S. Rajakumar and V. S. Bharathi, "Century Identification and Recognition of Ancient Tamil Character Recognition," *International Journal of Computer Applications*, vol. 26, no. 4, pp. 32–35, Jul. 2011, doi: [10.5120/3090-4237](https://doi.org/10.5120/3090-4237).
- [23] A. Tomar, M. Choudhary, and A. N. Yerpude, "Ancient Indian Scripts Image Pre-Processing and Dimensionality Reduction for Feature Extraction and Classification: A Survey," *International Journal of Computer Trends and Technology*, vol. 21, no. 2, pp. 85–93, Mar. 2015, doi: [10.14445/22312803/ijctt-v21p1116](https://doi.org/10.14445/22312803/ijctt-v21p1116).