International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

DEVELOPMENT OF AI BASED ANCIENT LETTER EXTRACTION

BALAMURUGAN V T[1],RATHISH G,[2]MANOJKUMAR P[2] AND KAMALHARSHAN K[2]

[1]*Department of Biomedical Engineering, Bannari Amman Institute of Technology, India*

[2]*Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, India*

Abstract-

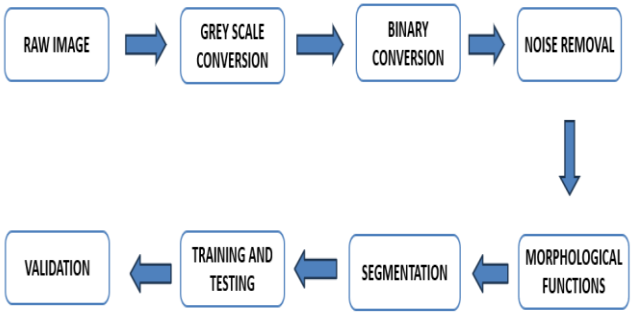
Palm leaf manuscripts are very significant because they provide a wealth of information. As a result, easy access to historical manuscripts must be made available in order to disseminate this knowledge around the world and encourage further research into ancient literature. In this study, an optical character recognition (OCR) system based on convolutional neural networks (CNN) is used to precisely digitize and identify Tamil palm leaf manuscript characters. In this article, the convolutional neural network's classifier, pooling layer, activation layer, fully connected layer, and other layers are used. The character set database was created using the scanned images of palm-leaf manuscripts. The database is split into 67 distinct classes, with about 100 samples in each class. The palm leaf manuscripts' OCR recognition and related issues are illustrated. The CNN model was used to construct a functioning example of the character recognition technique for Tamil palm-leaf text. It was discovered that the CNN model had a higher rate of recognition. Because so many features were extracted for each layer of CNN, the prediction rate and accuracy are very high.

Keywords - Palm leaf manuscript,Digitization,Convolutional NueralNetworks,Classification

1. INTRODUCTION

Information about the grandeur, way of life, economic situation, culture, and administrative practices used by several rulers and dynasties can be found in stone and palm leaf inscriptions found throughout the world. Rocks, pillars, slabs, stones, building walls, and temple bodies all have inscriptions. Additionally, it is present on seals, palm leaves, and copper plates. Palm leaf manuscripts must be preserved because they include knowledge about past religious practices, astrology, astronomy, and medical practices from antiquity. The information in the inscriptions can be cross-referenced to provide insight into the world's dynastic history. The epigraph makes reference to the research and inscriptions' deciphered text. . This information is used by epigraphists to identify graphics, explain their significance, and categorize applications according to historical periods and cultural contexts. Epigraphists employ this knowledge to recognize graphics, describe their meaning, and group applications into different time periods and cultural settings.

Different researchers employ a variety of techniques to get over the difficult issues with character recognition in stone and palm leaf inscriptions In order to overcome the challenging problems with character recognition in stone and palm leaf inscriptions, many researchers use a range of methodologies.. To preserve the information, palm leaf manuscripts are transformed into digital photographs. The process of capturing images using cameras or other optical scanners is known asimage acquisition. Pre-processing a picture is a method for reducing noise in the image and transforming it into a binary
format that can be processed. Grey photos are transformed into binary images using the binarization technique.



Block Diagram of Palm -Leaf Manuscript Character Recognition System

The efficacy of the character detection system may be decreased by the presence of noise. To enhance the frame, the image is first cropped to remove any distracting elements, and then it is skewed. The picture is skewered to the proper synchronization. An orientation angle is controlled by the skew detection function.

The classifier is then used to identify the character and convert it into a current text. An NLP-based tri-gram approach is utilized to anticipate the precise Unicode character. The corresponding Unicode values are used in order for the algorithm's matching pattern to correlate to characters and matched characters to identify the text in the palm leaf manuscripts by numerous researchers

There are numerous language manuscripts in various locations around the world that were used for the study. With the aid of this technology, we are able to decipher long-forgotten languages in addition to old scripts. It gives life to extinct languages and cultures, enabling us to cross over into the present and providing priceless insights into the development of society and the universal human experiences that unite us all.

In this investigation, we'll delve into the interesting field of AI-based ancient letter extraction and examine its techniques, uses, and significant implications for how we see the past and the human tale. AI is revealing our ancestors' secrets, shedding light on the stories that have created our world, and conserving their legacies for future generations—from the dunes of Egypt to the libraries of Rome.

2.PREVIOUS WORKS

An OCR-based method for identifying ancient Tamil characters in stones was published by Merline et al. [1]. Morphological operations on the input image include pre-processing and segmentation. The salt and pepper noise has been eliminated, the color images have been improved,

and they are now in grayscale. By employing the bounding boxes that are used to divide the dilated image, characteristics such as area attributes and corner points can be extracted. KNN and ensemble learning are used to classify the characters, and subsequently Unicode is used to match them. In addition to introducing the benefit of using the Self Organizing Map (SOM) model to capture and analyze the invariant properties of the Tamil Scripts, Gandhi et al. provided a method for Tamil character recognition [2].

The authors' proposed approach is different from a neural network in that it lacks a character-specific hidden layer. Input and output only require two layers. Gautum et al. [3] demonstrate a Brahmi script recognition system based on OCR. In the work, a geometric technique is used for feature extraction. Images are scanned using optical scanners. Thinning, thresholding, and cropping are all included in the pre-processing stage. During the segmentation process, images are divided into lines, and characters are subsequently extracted from each detected line.

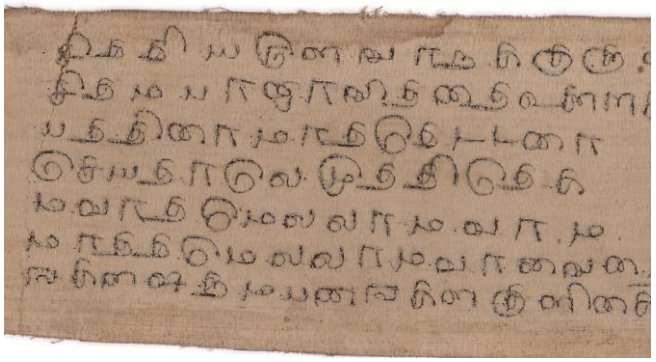
A simple-to-use system for identifying and analyzing Tamil inscriptions was developed by G. Janani et al. [4]. The authors first applied a noise reduction technique to the input image in order to completely remove the disturbance. The following morphological processes to be carried out are dilation and erosion techniques. The letters in the binary image are located using the linked component method. Each character was divided into its component parts and matched to the relevant Tamil language characters using correlation matching.

A project by Ezhilarasi et al. to make the process of interpreting and The following morphological processes to be carried out are dilation and erosion techniques. The letters in the binary image are located using the linked component method. Each character was divided into its component parts and matched to the relevant Tamil language characters using correlation matching. A project by Ezhilarasi et al. to make the process of interpreting and recognizing historic words in writing [5]. The authors proposed a method based on building a neural model using an 11th-century palaeographic stone inscription script for classifying part of speech tags and word prediction. For each new script, a Bi-LSTM model is created with an embedding word-vector layer to serve as the POS Tagging model. The classification of words includes tagging and word prediction.

Based on the technologies of OCR and NLP, Manigandan et al. recommended recognizing Tamil characters in inscription images from the ninth to the twelfth centuries [6]. The writers split the images after collecting, editing, and processing them for the aforementioned work. Prior to segmentation, color images were converted to binary and grayscale using a threshold value. A Scale Invariant Feature Transform

(SIFT) method was used to identify the characteristics of each character image, such as the quantity of lines, curves, loops, and points. Soumya et al. separated the lines and individual characters from the inscription images using the closest neighbor method. The Fourier

and wavelet transforms are used to extract aspects including structural, syntactical, statistical, and loop properties from the relevant data [7]. During the classification phase, genetic algorithms, Support Vector Machines (SVM), and the transductive SVM are employed.



An Image of ancient Tamil manuscript

2A .IMAGE PREPROCESSING

In order to analyze and catalog the inscription characters in a database, Bandara et al. created a web-based recognition system [8]. Border detection, segmentation, thinning, and binarization are all parts of the pre-processing stage. Binarization is carried out using Otsu's method, then thinning and feature extraction are carried out. The inscriptions were first scanned by Gautum and Chai using optical scanners, after which the photos were cropped. The pre-processing stage involved the use of techniques for thresholding and thinning. Characters are then extracted from each line that is recognized once images have been divided into lines. Six geometric entities—circle, semicircle, corner, intersect, bifurcation, and termination points—are used to extract the properties of the characters in the inscription [3].

Using the Otsu approach during pre-processing, Dias et al. binarized the pictures [9]. Then, boundaries are detected and dilation and erosion are applied to the binarized image. In the pre-processing step of old

document photographs, Soumya et al. applied noise reduction and image enhancement utilizing a variety of spatial filtering methods, including Median, Mean, Bilateral, and Gaussian Blur Filters [7].

. Using a gentle low-pass filter from the Fast Fourier Transform (FFT), noise in the image is reduced and removed in step [10] of the pre-processing process. Rajkumar et al. used c fuzzy median filters to minimize the noise in inscription photographs. [11]. The ZS Thinning algorithm, Enhanced Parallel Thinning algorithm, WHF Thinning algorithm, Quadratic Integral Ratio (QIR) algorithm, LW Thinning algorithm, Arabic Parallel Thinning algorithm, Matching algorithm, and Otsu's binarization algorithm [12], [13] were used to enhance the results on the input image. It is possible to evaluate hybrid pre-processing algorithms and create new, efficient thinning and thresholding techniques.

In Tomar et al. [14], various filtering methods for the pre-processing stage were analyzed. A non-linear bilateral filter is excellent for quality increases because, although it smoothens, it sharpens the edges, contrary to what median filtering was once thought to be best for: salt and pepper noises. At the time, the authors looked into hybrid methods such active contour, background and forefront independent binarization, shading estimations, and binarization. The simplest thresholding technique is global thresholding, to sum up.

Using a median filter to remove noise from the pictures and Otsu's approach to binarize the filtered image, Subashini et al. [15] suggested a period prediction system for ancient Tamil inscriptions. The binarized image is then further processed using a standard thinning method. Vertical and horizontal projection profile techniques are employed in the segmentation process. In the preprocessing step, an effort was made to reduce background noise and improve the image quality for better individual sharpness and computer recognition [16]. Using the Laplacian filter, Gaussian blur, and Unsharp mask, the input image is either smoothed or sharpened.

By scaling the character bitmap, Holambae et al. [17] were able to derive three separate attributes. The character image is initially segmented and thinned before intersecting characteristics are removed to create a single-pixel skeleton. 16 shadow qualities are then extracted from the character's 8 octagonal images. The final step is to extract the histogram characteristics from the original scaled character image by detecting and segmenting contour points. For each segment, the histogram chain code features are obtained. Guruviah et al.'s [18] focus was on enhancing CNN's OCR capabilities for reading Tamil inscriptions. Prior to binarization, the raw color image is first converted to grayscale. Binarization is the process used to extract text from a noisy backdrop using the Otsu algorithm threshold. The binarized image will subsequently be sliced into corresponding letter blocks, each corresponding to the Tamil characters.

A segmentation strategy based on machine learning was used by Mohamed et al. [19]. The Ti-MBL toolkit, which has been shown to be helpful for many NLP applications, was employed by the authors. Within-word features and sentence-context characteristics have both been considered as primary feature sets for segmentation. The authors came to the conclusion that more data must be present in the training dataset in order to achieve improved accuracy. Liu and Gao conducted research on the identification of oracle bone inscriptions (OBIs) [20]. By utilizing CNN and the current era's deep learning boom, the writers were able to recognize OBIs. Eight layers make up the neural network. In a series of convolutional layers, which were manually scanned OBI images, only filters of size 3x3 were utilized.The number of convolutional strides is fixed at one.

After each max-pooling process, the number of feature maps increases gradually. The method proposed by Panagopoulos et al. [21] that appears to work better, which uses the pixel intensity histogram for each letter and its lower turning point. The pixels have less intensity than this value is assumed to be from the letter. This method may produce a variety of artefacts, which can be removed using morphological filters. Each contour pixel must have an exact pair of neighboring pixels, and isolated pixels are not permitted, according to the authors' assurance.

2B. FEATURE EXTRACTION

Characters are described by their H-center, V-center, height, width, HP-Skewness, VP-Skewness, and horizontal, vertical, right-to-left, and left-to-right diagonal stroke densities. Features are then retrieved using these properties [15].
Kannada language characters in epigraphs were identified using optical character recognition and fuzzy logic by Soumya et al. [7]. In order to extract statistical features, the authors used three different techniques. A one-dimensional statistical analyzer was used to compute Mean, Standard deviation, and Variance. A histogram analyzer was used to determine kurtosis, entropy values, and skewness, and a GLCM analyzer was used to determine the smoothness and coarseness of a picture. When attempting to identify Brahmi characters, Vellingiriraj et al. [22] focussed on zonal density to extract characteristics.

The recognition of Bangla compound letters and words was a focus of some investigations. Using zones and row-wise longest runs, SVM and RBF were utilized to extract features from Bangla compound characters [23]. Using structural decomposition, Bangla compound characters were also detected [24]. Using Gabor features and zone-based techniques for extracting the characteristics, Soumya et al.'s main goal was to recognize the ancient script (Brahmi script) used during King Ashoka's reign [25]. Mousavi et al. created the Tesseract engine for segmentation, learning, and classification [26]. Images with salt and pepper and Gaussian noise are first sharpened and filtered with a median filter; after that, small objects are removed using a morphological technique. The characters in the image are then recovered using a different dilation technique.

2C.CLASSIFICATION METHODS

For grouping the characters in old inscriptions, Dias et al. employed the K-means algorithm [9]. By using a Transform Feature Scale-Invariant (SIFT), Rajakumar et al. [11] acknowledged old Tamil characters, and a new strategy based on keypoint bags was developed. Subashini et al. [15] employed a Support Vector Machine (SVM) to categorize data based on the concept of decision planes that establish choice limitations. The authors choose the radial function kernel type to improve the performance of SVMs in this system. Multiclass SVM turned out to be a very effective method for predicting the Tamil inscription centuries. Neural networks were used to train images and match them to modern Tamil letters [11]. The SVM classifier and means algorithm were employed in [11] to increase classification accuracy overall. The SVM classifier uses a feature vector with key points to assign a character to the image. S. Li et al. [27] proposed a CNN-based character extraction method that featured character embedding and convolutional layers that included k-max-pooling as well as a Bi-LSTM layer and a tag inference layer for the extraction of the character's hidden local feature.

The major issues presented by Merlin et al. [28] that need little computing expense are simple convolution neural networks and Unicode mappings for image categorization. The 18 layers that make up the CNN design characteristics can be retrieved and categorized by the Softmax operator. The created architecture is trained, test data is applied to the trained networks, and the precise categorized character is transferred to Unicode values. A convolution neural network was suggested by Prashanth et al. [29] to be used in the development of a Tamil character handwriting recognition system. Utilizing stochastic pooling, weighted probabilistic pooling, and standardization of local contrast, the ConvNetJS package was expanded to learn functions.

Using a method for identifying characteristics from the Temple walls, Rajakumar and Subbiah Bharathi [11] proposed an identification methodology for the characters from the twelfth century. The suggested approach uses a contour-let to precisely manage the curved images. Clustering algorithms will be used to identify input characters, and a fuzzy median filter will be used to reduce noise. Finally, a neural network evaluates a system for contrasting historical characters with contemporary characters.

According to Clanuwat et al. [30], there are parallels between early and modern Japanese character classifications. The experiment made use of two fully connected convolutional autoencoders. Data Recurrent Neural Networks (RNN) and mixed density networks (MDN) are employed for training. Bhowmik, et al. [31] examine the classifier's performance using the elliptical technique and a variety of classifiers, such as Naive Bayes, Dagging, Bagging, SVM, and MLP. For the purpose of recognizing Bangla words, the collected features were categorised using an ANN and the Histogram Oriented Gradient (HOG). A cropped image dataset is provided as input to CNN for image classification and detection. CNN is trained using data augmentation and transfer learning in the work of Guruviah et al. [18] to categorize Tamil letters based on Keras and Tensorflow

3.MATERIALS

3A. Image capturization

The palm leaf manuscript must be identified and scanned at a resolution of at least 300 dpi on a high-quality scanner. Since the palm leaf is at least 150 years old and has a significant likelihood of leaf degradation, high-quality scanning is necessary to retrieve the data without losing any of it. In the preprocessing stage, high quality scanned images were helpful in locating and eliminating noise in very old manuscripts.

3B. Image preparation

Skew correction, binarization, border detection, segmentation, and thinning are all parts of the pre-processing stage. The binarization process employs adaptive thresholding, which is followed by thinning and deep learning feature extraction. The image from the acquisition stage might not be perfectly aligned; it could be placed at any angle. Skew correction is therefore necessary to guarantee that the image conveyed to the succeeding phases is correctly oriented. The scanned color image is then changed to a grayscale version.

For every AI-based recognition model, clean data is the final need to attain maximum accuracy. Noise removal is the next significant step in preprocessing. Noise removal is therefore a crucial step in the proposed approach to provide cleaner data to the AI model. Different filtering methods, such as the median filter, Gaussian filter, and non-local means de-noising filter, are used to remove noise.

A non-linear digital filtering method called the median filter is used to eliminate noise from photos and data. This form of noise reduction is widely applied as a pre-processing technique for palm leaf manuscript processing in order to enhance the outcomes. Because median filtering can sometimes preserve edges while lowering noise, it was chosen for the proposed investigation

A 2D convolution-based smoothing filter called the Gaussian filter is used to lower noise in visual data. The kind of kernel used is the only distinction between a Gaussian filter and a mean filter.

$$G(x)= 1/\sqrt{2\pi\sigma}\, e^{(-x^2/(2\sigma^2))}$$

the standard deviation is where

Given is the 2D Gaussian function.

G(x,y)=1/(2**2) e(-(x**2+y**2/(2**2))

Image processing employs a non-local technique to image denoising. Every pixel in the image is averaged by non-local mean filters, which weight each pixel according to how similar it is to the target pixel. Consequently, after filtering, the image's clarity is greatly improved and its details are preserved.

u(p)=1/c(p) f(d(B(p),B(q))) u(q)dq

Where the normalizing factor is identified as c(p), the Euclidean distance between image patches with centers at p and q is given as d(B(p),B(q), and f is a decreasing function. Because the manuscripts for the proposed work are so ancient, it's possible that certain letters or characters have degenerated. The non-local mean de-noising function is therefore the most appropriate.

A binary image is created after the filtered image. A number of methods, including Otsu's binarization, adaptive thresholding, and the local maxima and minima method, can be used to binarize the image. The suggested approach uses adaptive thresholding since it dynamically generates a threshold value based on the location in the image in order to binarize the image.

Palm leaf writings require the segmentation of several characters according to continuity. Therefore, morphological processes are necessary for character segmentation. Text extraction uses a variety of core

morphological techniques, such as dilatation, erosion, opening, and closing, as well as the OCR method for character identification from photographs. The image is then modified morphologically after edge detection. First the image is enlarged, then it gets eroded.

The three types of image segmentation that can be done are at the line, word, and character levels. It is suggested to employ character level segmentation because the proposed

task comprises handwritten script, where the words or lines cannot be separated. At this point in the segmentation process, a character-based image is used.

All segmented and annotated patch images will be combined into a dataset of isolated characters. It has 8,500 character samples and 67 character classes, with 6,700 character samples being used as a train set and chosen based on character classes. As a test set, the remaining samples will be used. There are various numbers of sample photos in each class.

In our collection of palm leaf manuscripts, some classes are commonly used while others are not.Convolutional neural networks are used in this study instead of feature extraction. The first layer of the network receives as input data all of the pixel values in a particular image. In this study, the TensorFlow library is used.

A multilayer convolutional neural network is applied in this case. The network's architecture is depicted in figure 3. The input grayscale image of 28x28 pixels is used to compute the first convolutional layer (C1) using a sliding window of 3x3 pixels. The convolution output is followed by the ReLU function.

ReLU functions do not saturate during convnet training, assisting in avoiding the vanishing gradient issue. In order to reduce the processing intensity of the architecture, a maxpooling operation may be utilized after each convolution operation, depending on the architectural design.

4.RESULTS

The project uses Python altogether. To segment images and characters, open-source libraries like OpenCV and PIL are employed. Convnets are implemented using Keras, and TensorFlow serves as the backend. The Google Co-lab GPU environment serves as the training environment. The character recognition model is developed using the training data from the palm leaf manuscript that was used to train the CNN. After preprocessing, the model is evaluated on a fresh palm leaf manuscript with various character data sets. Then, in the manuscript output image, the detected characters are shown on top of each character.

Using the CNN model, the proposed palm leaf manuscript characters are identified and categorized. Using examples, the CNN algorithm instructs the computer to learn. The generated data set was used to develop and train the CNN model. For the palm leaf text, a unique data collection was developed. Each character in the preprocessed image is segmented, and character features are extracted from the segmented image to form the data set. To train each recognized character, 100 input photos with various writing styles are used. The segmented characters are accurately detected by the CNN model with 81.62% accuracy using the newly preprocessed palm leaf manuscript image.

5.REFERENCES

[1] Language (Baltim), vol. 192, 2009, pp. 188–192. R. I. Gandhi,An Attempt to Recognize Handwritten Tamil Character Using Kohonen SOM”

[2] “Tamil Character Recognition from Ancient Epigraphical Inscriptionm Using OCR and NLP” 2017 Int. Conf. Energy, Commun. Data Anal.Soft Comput., pp. 1008–1011; T. Manigandan, V. Vidhya, V.Dhanalakshmi, and B. Nirmala.

[3].”Recognition of ancient Kannada epigraphs using fuzzy-based Approach”2014 Int. Conf. Contemp. Comput. Informatics, pp. 657–662, by A. Soumya and G. H. Kumar.

[4] “GPU Based Optical Character Transcription for Ancient Inscription Recognition” 15th Int. Conf. Virtual Syst. Multimed., 2009, pp. 154–159.

[5] “12th Century Ancient Tamil Character Recognition From Temple Wall Inscriptions” S. Rajakumar and V. S. Bharathi, i-manager's J. Embed.Syst., vol. 1, no. 2, pp. 27–31, 2012.

[6]"Thresholding: A Pixel-Level Image Processing Methodology Preprocessing Technique for an OCR System for the Brahmi Script" by H.M. Devi, published in 2006.

[7] "Ancient Indian Scripts Image Pre-Processing and Dimensionality Reduction for Feature Extraction and Classification: A Survey" A. Tomar, M. Choudhary, and A. Yerpude, 2015.

[8] "Preprocessing of Camera Captured Inscriptions and Segmentation of Handwritten Kannada text," International Journal of Advanced Research in Computer and Communications Engineering, vol. 3, no.5, pp. 6794-6803, 2014.

[9]"Combining Multiple Feature Extraction Technique and Classifiers for Increasing Accuracy for Devanagari OCR," no. 4, 2013, pp. 38–41. A.N. Holambe and R. Thool.

[10] "Preprocessing of Camera Captured Inscriptions and Segmentation of Handwritten Kannada text," International Journal of Advanced Research in Computer and Communications Engineering, vol. 3, no.5, pp. 6794-6803, 2014.

[11]"A Novel Approach to OCR Using Image Recognition based"Classification for Ancient Tamil Inscriptions in Temples" ArXiv, vol.abs1907.04917,

[12] "6 Inscriptions from Ethiopia" by A. Bausi and P. M. Liuzzo.Crossing Experimental Digitized Epigraphy, Encoding Inscriptions in Beta, 2018.

[13] "Automatic Writer Identification of Ancient Greek Inscriptions"; IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, pp. 1404–1414, 2009. M.Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S.Tracy.

[14] "An enhanced harmony search method for Bangla handwritten character recognition using region sampling" 2015 IEEE 2nd International Conference on Recent Trends in Information Systems, pp. 325–330.

[15] "Recognition of Bangla compound characters using structural decomposition" Pattern Recognit., vol. 47, no. 3, pp. 1187–1201,2014. S. Bag, G. Harit, and P. Bhowmick.