A Novel Approach in Deep Learning method to Identify the Tamil Manuscripts in Palm Leaves

C.Balaji[1], P.Lalitha[2]

[1]Assistant Professor Department of Computer Science Dr.NGP Arts and Science College Coimbatore, Tamilnadu, India, Email: bala3mca@gmail.com
[2]Professor Department of BCA Hindusthan College of Arts and Science Coimbatore, Tamilnadu, India, Email: lalithavellingiri@gmail.com

Received: 19.04.2024 Revised : 11.05.2024 Accepted: 21.05.2024

ABSTRACT
Manuscripts written on palm leaves are made using specially prepared palm leaves. Usually, the leaves are salted, dried, and then written on. Often, a stylus is used for writing. Religious writings, literary works, medical expertise, astrological information, and other cultural and historical details are among the many types of information found in these manuscripts. Numerous of these manuscripts are written in antiquated scripts like Pallava, Vatteluttu, or Grantha. Manuscripts written on palm leaves have been preserved and digitized because of their historical significance. Careful treatment is necessary during preservation to avoid deterioration, and several digitization efforts have been started to generate digital archives. The cultural legacy of Tamil Nadu and other Tamil-speaking areas is fundamentally based on Tamil palm leaf manuscripts. They offer a sight into the rich past, customs, and wisdom of the Tamil-speaking populace.

Keywords: CDF (Cumulative Distribution Function), Pytesseract, Deep Learning, Gaussian Function

1. INTRODUCTION

There are several palm leaf manuscripts scattered throughout South India. Stalwarts, pillars, slabs, stones, building walls, and temple bodies all have inscriptions on them. Additionally, it is present on seals, palm leaves, and copper plates. Ancient palm leaves must be maintained since they include knowledge about ancient tamil literatures, astrological information, astronomical specifics, land document details, and historical religious traditions. Poem from classical Tamil literature sheds light on ancient Tamil people's way of existence.

Creating palm leaf manuscripts required washing, slicing, and seasoning the palm leaves before soaking them in milk or water. They were then cleaned, oiled, and polished with shells or stone after being buried in wet sand. One or two holes were bored through the center, depending on how wide the leaves were to bind the leaves together to form a book.On palm leaves, texts are inscribed in two different ways. As in birch bark manuscripts, the first method was using a pen or pencil to paint tincture on the written information. This strategy is primarily applied in Nepal and East India. The letters were then incised onto the palm leaf with a metal or bone stylus before being highlighted with a dark pigment produced from lampblack or plant extracts. South India, Southeast Asia, and later North India have seen a rise in popularity for this strategy.

2. LITERATURE SURVEY

[1] In this study, an input image is binarized using Otsu's method. Divide the character into several shapes after binarization, including circles, semicircles, corners, intersections, bifurcations, and termination points. To find the limits, use the method of dilation and erosion. Using the median filter, Gaussian filter, and nonlocal mean denoising filter methods, the noisy pixels are removed. 800 of the 2700 total samples used for training were used for the test set. A sliding window of 3x3 pixels was employed with the CNN Model for 28*28 pixels. ReLU (Rectified Linear Unit) function was utilised to recognise Tamil characters after CNN by vanishing the gradient problem.

[2] In this instance, the input image is changed so that the backdrop is black and the foreground text is white, changing the colour range of 0 to 255 into 0 and 1. so that processing can be done with highly clear characters. Utilizing dilatation and erosion in morphological operations helps to thicken a character's border and fill in any gaps. Using TLS (Text Line Segmentation) as a cutting-edge procedure to segment

the lines and find the overlapping characters Thimunet architecture is used in the CNN Model to recognise the characters at various levels.
[3] This study used the Otsu method to transform the colour image into grayscale, which is subsequently converted into binarization with a threshold value of 0.6. According to the outcomes of each phase, the character is identified using 3 separate steps. Phase I of the 2DFFT algorithm will employ an image that is 32*32 in size. After the Phase I image was converted using the 2D-DWT method and utilised as an input for Phase II, the desired result was not obtained. Using Phase II's image as input, Phase III employs the 2D IDCT and 2D-DCT algorithms for character identification in Tamil. Phase IV use the 2D-DWT+IB method to identify the character.
[4] The character strokes in this paper's example go beyond the text area and extend the line space known as the obstacle following the conversion of binarization of an image. This Text Line Slicing method uses four different sorts of variables to segment characters with or without obstacles, including horizontal space, horizontal track, vertical space, and vertical track. According to a few downward strokes, the length of the appeal in Tamil overlaps to another character.
[5] In this case, picking a character after binarization, locating the pixel position, and measuring the x, y, and depth of the pixel using measurescope with the aid of the dial gauge indicator to identify the character.
[6] Data pretreatment, feature extraction, and clustering are the three stages of image processing that are covered in this study. Binarization comes first in the data pretreatment procedure, then clean data is reduced as noise. Scripts can be distinguished from their backgrounds with the use of border cropping. The procedure of thinnin helps to extract the image skeleton. Reduce the amount of data needed for feature extraction. A bottom-up technique for clustering that is applicable to average-link, complete-link, and single-link systems is agglomerative hierarchical clustering. Apply the K-mean clustering algorithm to separate the data observations into k-clusters. Utilizing the k-centroids c1, c2,... ck, one may determine the nearest distance to the cluster centroid.
[7] The identification of telugu composite characters occurs in this context at various stages, including data collecting, data preprocessing, character segmentation, pattern development, and character recognition. During the data collecting phase, the text line's distance, length, and size are scanned using the coordinates x, y, and z. Eliminate the background noisy pixel during the data preprocessing stage. Using the vertical projection algorithm, character points are segmented according to their size and shape in character segmentation. The XY, YZ, and XZ projection planes are used to normalize the fixed value M (M=50) based on the character's height and width during the pattern generating phase. Four variables—DTB (Distance from Top to Bottom), DLR (Distance from Left to Right), and DRL (Distance from Right to Left)—are utilized in character recognition to extract characters using distance profiles and histogram profiles (Distance from Bottom to Top). It was possible to ascertain the limits and form of the character by utilizing these variables.

3. The Suggested Method

3.1Data Collection:

The collection of Tamil-language palm leaf manuscripts covering a variety of subject areas was the focus of this investigation. The study considered a variety of Tamil literary leaf manuscripts in order to determine the target characters. Table 2 below gives a summary of the number of palm leaf manuscripts used in this study.

Table 1. Current Research Approaches for Computerizing Texts on Palm Leaves.

Paper	Methodology	Language
CNN-powered Character Identification and Categorization	CNN Model using ReLU	Tamil
Using Deep Learning to Identify Tamil Characters	Thimunet Architecture using Text Slicing Algorithm	Tamil
Using the IBL method, a transform-supported extraction approach for PLHCR	PLHCR using IBL algorithm	Malayalam
Text Line Breaking in Tamil Handwritten Manuscripts	Text Line Slicing algorithm	Tamil
Telugu Palm Leaf Character Recognition Isolated	Radon Transform Method	Telugu

Balinese Script Clustering in Palm Leaf	Agglomerative Clustering and K-means Clustering	Balinese

Table 2. Total Count of Manuscripts on Palm Leaf

Title	Count of Palm Leaf Typescripts
Agananoor	159
AgathiyarVaithyar	147
Iyengurunoor	158
IynthinaiImbathu	104
KalavaliNaarputhu	117
Kalingathu Barani	67
OovaiyarVaakundam	57
Thiruvasagam	81
Kanthar Alangaram	83
Mookudarpalu	113
Total	1086

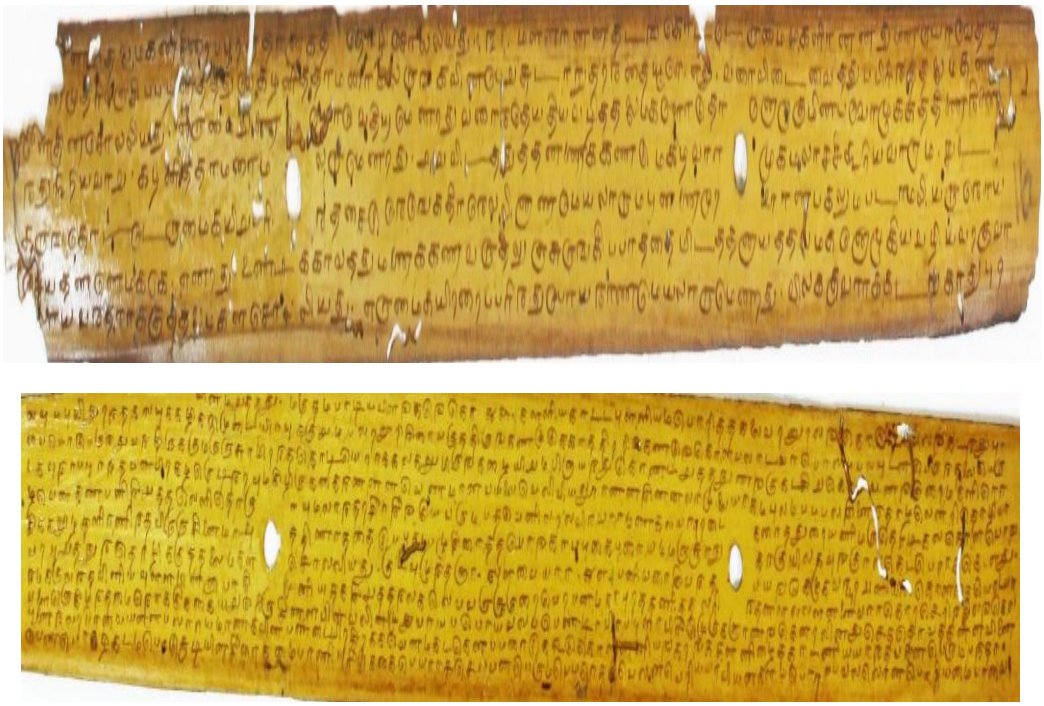


Figure 1. An Example of Palm Leaf Manuscripts in Tamil Literature

There were 1,086 palm leaf manuscripts gathered in all. A few manuscripts of Tamil literature that were gathered for the data samples are displayed in Figure 1.

3.2 Flowchart of the Processes

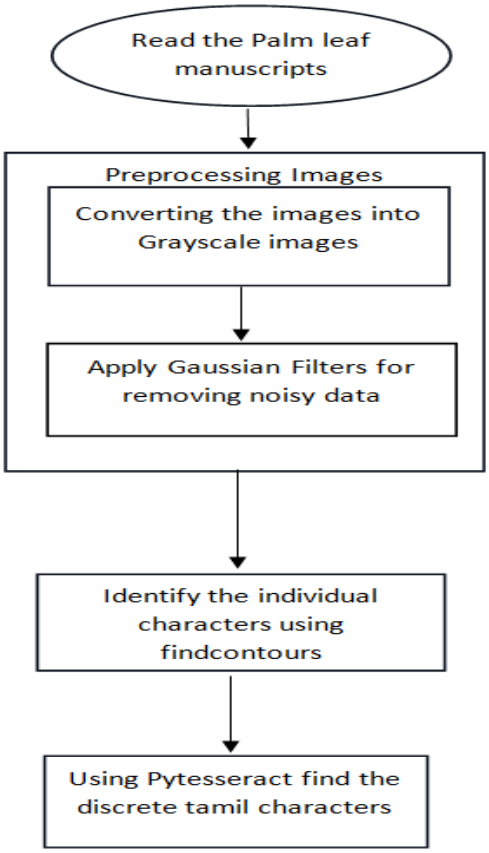


Figure 2. Sequence of Operations

3.3 Proposed Approach

The handwritten method is used to build the dataset. One of the fiddliest and most fascinating areas of pattern recognition and image processing study has been handwriting recognition. Specifically, it primarily advances automation techniques and creates a human-machine interface for many applications.

Step 1: Read and resize all the Tamil palm leaf manuscript in the size of 301 pixels of height and 3548 pixels of width approximately with 72 dpi quality.

Step 2: Convert the colour images into grayscale images or black and white images to identify the character without difficulty. Binarization is the process of applying a predetermined threshold value to assign 0s and 1s. In palm leaf manuscripts, 0 denotes the white background and 1 denotes the black foreground text. T displays the global threshold of 50. The pre-processed images are shows in (Fig. 3).

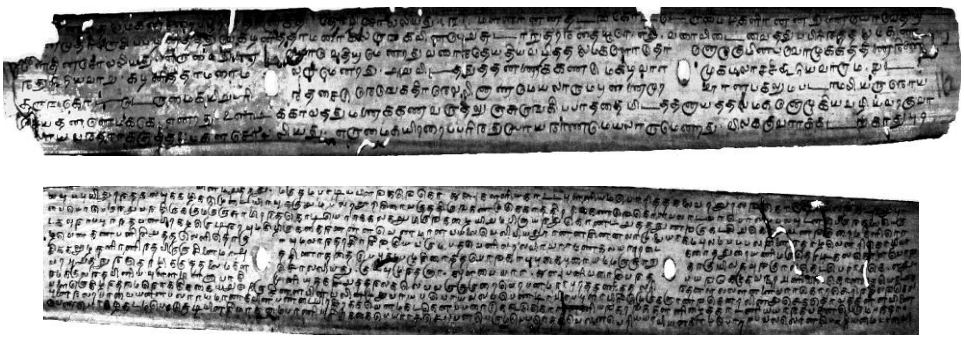
$g(x,y)=I(f(x,y)\geq T)$

Where $I(\cdot)$is the indicator function, defined as:

$I(P)=1$if P is true

0if P is false

In this case, Pis the condition $f(x,y)\geq T$. The function $g(x,y)$ will output 1 if $f(x,y)\geq T$ and 0 otherwise, which is equivalent to the original piecewise function.



Step 3: The Gaussian kernel forms the basis of the Gaussian filter, which is a two-dimensional matrix of values obtained from the Gaussian function. The kernel is usually square and has a centre with higher weights and lower weights towards the edges.

$$G(x, y) = \frac{1}{2\Pi^{\sigma}} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$$

where σ is the Gaussian distribution's standard deviation and (x, y) are the kernel's coordinates. It is the standard deviation that governs the Gaussian curve's spread. The input image and the Gaussian kernel are then convolved. The process of convolution entails moving the kernel across the image and calculating the weighted sum of the pixel values it covers at each location. A weighted average of the surrounding pixels, with higher weights given to pixels closer to the kernel's center, is the final pixel value in the output image. By minimizing abrupt changes in pixel values and high-frequency noise, the convolution process successfully blurs the image. The Gaussian distribution's standard deviation (σ) determines how much smoothing or blurring occurs. The blur becomes smoother and wider with a greater σ. The Gaussian distribution determines the weights of a convolution operation using a Gaussian kernel, which is how a Gaussian filter smoothes a picture. Control over the degree of image smoothing is possible thanks to the standard deviation parameter.

Step 4: After this process, by applying histogram equalization to improve the character enhancement in all images. By distributing the intensity levels, it improves the contrast of the image. It works by altering pixel intensity values in an image to improve the uniformity of the image's cumulative distribution function. As a result, the distribution of intensity levels becomes more balanced, which might enhance the contrast overall. The distribution of the image's intensity values is shown by the histogram. Using the histogram as a guide, compute the cumulative distribution function. The CDF provides the cumulative likelihood that a pixel's intensity will be less than or equal to a specific amount. Adjust the CDF to fall between [0, 1]. To guarantee that the transformation is done consistently to all intensity levels, this step is crucial. Take the matching normalized CDF value and multiply it by the maximum intensity level for each pixel in the image to replace its original intensity value. This mapping effectively redistributes the intensity values.

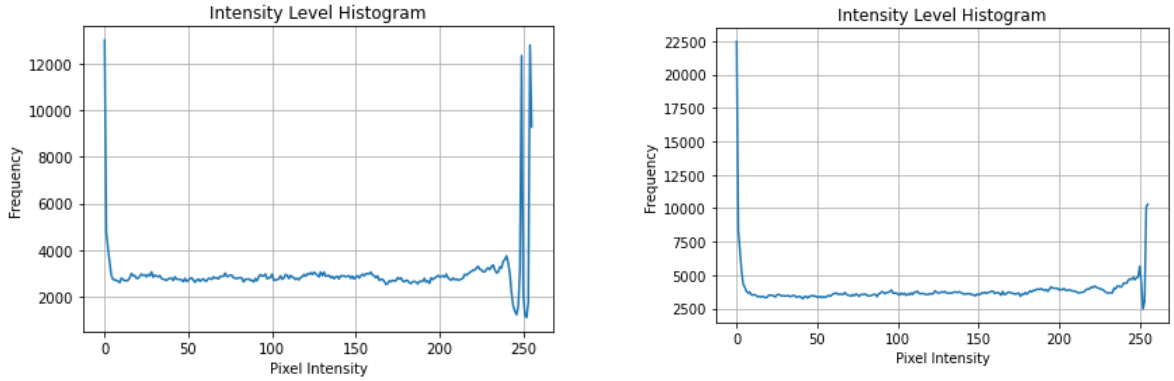
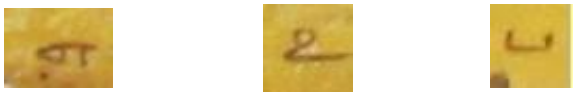


Figure 4. CDF Distribution Intensity Level of Sample Images

Step 5: Identify the individual character edging through findcontours. Contours are simply the boundaries of objects in an image. The ability to recognize and extract a character's outlines can be helpful in identification. Green color boxing squares off each character.

Figure 5. Identification of Individual Character in sample image

Step 6: After determining which individual characters, using Pytesseract Optical Character Recognition technology, correlate with one another. With support for over 100 languages andbeing open source and licensed under the Apache license, Tesseract stands out above competing OCR engines. It is compatible



with all popular computer operating systems and platforms. Aside from this, it uses its own or pre-existing GUIs or APIs (Application Programming Interface) to generate its output in many formats, including text, PDF, and other portable document formats.

Tesseract is one of the most widely used and reliable open-source optical character recognition (OCR) engines. It began as a doctoral research project at the Bristol Hewlett-Packard (HP) labs. Moreover, Google took the lead in 2006 with Tesseract's 2005 Open Source release. Tesseract is available from the official repositories maintained by Google Inc. and Ray Smith.It is possible to integrate the repositories with any programming language because they offer extensive support for a wide range of languages. Since Tesseract lacks integrated Graphical User Interface (GUI) support, one can test the essential functions of Tesseract by utilizing these Tesseract OCR engine repositories. It is possible to integrate an existing open source GUI repository or create a new one using their GUI.

Figure 6 shows how the Tesseract OCR engine actually operates. An picture is input in the first stage, where OCR is required. It then moves on to the "Adaptive Thresholding" step, where it is transformed into a binary image. A later processing step for the binary image was "Connected Component Analysis," which divided the text and words using character outlines. Further delivered to a two-way pass for word recognition are the outlined characters. Pass-1 recognizes words or text and then processes them further into an adaptive classifier that uses the information as training data.

Discussion, Conclusion and Future Enhancement

Tamil, one of the world's oldest languages, is gaining importance among academics worldwide due to its historical significance and ability to endure for generations. Millions of people speak it as their mother tongue all over the world, including those living in the modern day, which accounts for its popularity. The personal interests and philosophies of the scholars have influenced the way that study on medical manuscripts in Tamil literature has evolved throughout time. Character recognition of ancient Tamil features, which focuses on Tamil character identification to maximize data collection, is the most popular study field in the real world. It was possible to observe old medicinal systems because of the palm leaves that were included in ancient Tamil inscriptions. Even if several people held palm leaf manuscripts by default, regrettably, poor handling has led to their annihilation. For example, most of the participants owned the texts as family property and showed little interest in traditional medicine. Another explanation for the manuscript's destruction is inadequate comprehension of either digitization or conservation. During the field research, it was found that the distinctive palm leaf writings had been obliterated by moths, mice, and harsh weather. Most of those that had been kept on hut roofs had been destroyed by rain. Thus, a thorough analysis of Tamil palm leaf manuscripts, their translation, and a listing of their publications were required in order to thoroughly research the subject of vital medicine. Numerous organizations and individual healers still hold hundreds of palm leaf manuscripts however, they need to be digitized and appropriate catalogues created for future reference. Remember that most of the palm leaf manuscripts that are currently in print largely include the therapeutic formulas that practitioners use in addition to their unique clinical experiences. Parallel to this, the survey includes a large number of studies on the recognition of handwritten Tamil text. Thus, a thorough analysis of Tamil palm leaf manuscripts, their translation, and a listing of their publications were required in order to thoroughly research the subject of vital medicine. Numerous organizations and individual healers still hold hundreds of palm leaf manuscripts; however, they need to be digitized and appropriate catalogues created for future reference. Remember that most of the palm leaf manuscripts that are currently in print largely include the therapeutic formulas that practitioners use in addition to their unique clinical experiences. Parallel to this, the survey includes a large number of studies on the recognition of handwritten Tamil text. To be sure, there has been very little research on medical manuscripts unearthed in Tamil literature, but there is still much to learn about the language given its rich heritage and history. Because there aren't many research on the classification of the language or the subject of handwriting identification, it is also vital to ascertain the origin of the Tamil language and the extent to which its characters have evolved over time. It's interesting to note that very few previous researchers tried to generate a significant number of real, handwritten Tamil character datasets for training and classification.

Thus, in order to provide training datasets, this study constitutes one of the first attempts to manually and automatically separate Tamil palm leaf manuscripts. Using the dataset, future researchers could build expert systems for a range of applications, including the classification of characters based on the century in which they may have evolved, the identification of extinct characters, and the identification of characters whose forms may have changed. It is important to keep in mind that the study's scope was limited to Tamil palm leaf literary manuscripts. The project's objective was to create a vast number of Tamil character databases by sorting ancient Tamil palm leaf manuscripts. The identification of characters from Tamil palm leaf manuscripts in literature has not been the subject of any prior research. Consequently, the study's findings ought to be useful in handwriting recognition of a number of unique Tamil characters, especially in the literary field. Future machine learning and neural network algorithms (as well as expert systems) can use the produced training data as input to aid in categorization and unearth fresh details regarding the evolution of the characters included in Tamil-language medical manuscripts.

REFERENCES

[1] Pravin Savaridass M, Haritha J, Balamurugan V T, Vairavel K S, Ikram N, “CNN Based Character Recognition and Classification in Tamil Palm Leaf Manuscripts”, IEEE, 2021.
[2] Gayathri Devi S, SubramaniyaswamyVairavasundaram, Yuvaraja Teekaraman, Ramya Kuppusamy and Arun Radhakrishnan, “A Deep Learning Approach for Recognizing the Cursive Tamil Characters in Palm Leaf Manuscripts”, Hindawi, 2022.
[3] S. Gopinathan, I. Jailingeswari, “Transform Supported Extraction Approach for PLHCR using IBL Algorithm in Malayalam Palm Leaf”, International Journal of Mechanical Engineering, 2022
[4] Dr. M. Mohamed Sathik , R. SpurgenRatheash, “Text Line Segmentation in Tamil Language Palm Leaf Manuscripts – A Novel Approach”, Journal of Tianjin University Science and Technology, 2021.
[5] Panyam Narahari Sastry, Ramakrishnan, “Isolated Telugu Palm Leaf Character Recognition Using Radon Transform – A Novel Approach”, IEEE, 2012
[6] Anastasia Rita Widiarti, C. Kuntoro Adi, “Clustering Balinese Script Image in Palm Leaf Using Hierarchical K-Means Algorithm”, ICIST, 2020
[7] Ayush Pradhan, Sidharth Behera, and Pushpalata Pujari, "Comparative Study on Recent Text Line Segmentation Methods of Unconstrained Handwritten Scripts", ICECDS, IEEE, 2017.
[8] Dona Valy, Michel Verleysen, and Kimheng Sok, “Line Segmentation for Grayscale Text-images of Khmer Palm Leaf Manuscripts”, IEEE, 2017.
[9] Himanshu Jain, Archana Praveen Kumar, "A Bottom up Procedure for Text Line Segmentation of Latin Script", IEEE, 2017.
[10] Kathirvalavakumar, Thangairulappan, and Karthigaiselvi Mohan, “Efficient Segmentation of printed Tamil Script into characters Using Projection and Structure”, ICIIP, IEEE, 2017.
[11] Dona Valy, Michel Verleysen, and Kimheng Sok, “Line Segmentation for Grayscale Text-images of Khmer Palm Leaf Manuscripts”, IEEE, 2017.
[12] T.M. Kodinariya and R.R. Makwana, “Review on determining number of cluster in K-Means Clustering,” International Journal of Advance Research in Computer Science and Management Studies, vol. I, 2013.
[13] A. Triayudi and I. Fitri, “A new agglomerative hierarchical clustering to model student activity in online learning,” Telkomnika vol. 17(3), 2019.
[14] V.N.Manjunath Aradhya, G.Hemantha Kumar, S.Noushath, “Robust Unconstrained Handwritten Digit Recognition Using Radon Transform”, IEEE-ICSCN, 2007.
[15] Panyam Narahari Sastry, Ramakrishnan Krishnan and Bhagavatula Venkata Sanker Ram, Classification and Identification of Telugu hand written characters extracted from palm leaves using decision tree approach, ARPN Journal of Engineering and Applied Sciences, Vol. 5, No. 3, March 2010.
[16] Vijaya Kumar Koppula, AtulNegi, "Fringe Map Based Text Line Segmentation of Printed Telugu Document Images", IEEE, 2011.