IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol-6, Special Issue-4, May 2018 E-ISSN: 2347-2693

Tamil Palm Leaf Manuscript Character Segmentation using GLCM feature extraction

M. Sornam[1*], Poornima Devi. M[2]

[1*] Department of Computer Science, University of Madras, Chennai, India
[2] Department of Computer Science, University of Madras, Chennai, India

Corresponding Author: madasamy.sornam@gmail.com, Tel.: +91-44-25399628

Available online at: www.ijcseonline.org

Abstract— The main objective of this proposed effort is to advance the system that empowers recognition of Tamil characters from palm leaf and inscription through captured images and stock them for forthcoming use. Some training mechanism has done with several methodologies, but distinguishing Tamil characters stances challengeable mission. Tamil language is considered too complex compared to any other language because of the presences of curved, slope, twist, pits and it will vary writing style of individual to individual. More research needs adapting ancient Tamil characters to modern Tamil characters to extend the aim of creating computerized system for providing improved understanding of human knowledge. This proposed work is applicable for segmenting Tamil characters and store it in an organized system folder for further processing of the image. Gray-Level Co-occurrence Matrix (GLCM) feature extraction is used to quantify the statistical features of segmented characters. At this juncture segmented Tamil Characters are compared with Palm leaf manuscript, Stone Inscription, Handwritten characters and document characters using GLCM feature and the results are promising.

Keywords— Gaussian, Bilateral, GLCM, PSNR, SSIM, MSE, Homogeneity, Angular Second Moment (ASM).

I. Introduction

This exploration work focused on Tamil character recognition from ancient palm leaf because chronological mysteries were concealed in ancient manuscripts. Characters are compared with palm leaf characters, stone inscription characters, handwritten characters and document characters using GLCM feature extraction. Images were preprocessed, segmented as a character and then permitted for feature extraction to quantify the statistical feature.

Captured image differ due to lighting when the image took, eminence of the camera, eminence of the things. So the captured image is permitted to preprocessing. First the image is transformed to grayscale image, then gray scaled image are allowed to normalization. Filters are used to eliminate the noise from an image; here different kinds of filters used are median filter, bilateral filter and Gaussian filter. Then the filtered image is permitted to threshold; simple thresholds, Otsu threshold, adaptive mean threshold, adaptive mean Gaussian threshold are used to examine which thresholding techniques achieve efficient result. Then the output from threshold image is allowed for segmentation. The segmented characters are stored in the system folder as a databank for feature extraction. Finally, GLCM feature extraction is accomplished for the distinct characters stored in the databank to measure the statistical features such as Contrast, Energy, Homogeneity, Correlation, Dissimilarity and Angular Second Moment (ASM). In this paper section II comprehends related work, section III defines methodology, section IV labels feature extraction, section V contains experimental results and discussion, and section VI comprises conclusion and future enhancement.

II. Related Work

Kavitha Subramani and Murugavalli [8] published to recognize Tamil palm leaf characters from historical documents. This system included three processes: preprocessing, binarization and postprocessing. G. Bhuvaneswari et al. [2] proposed a system for ancient character recognition (ACR) for stone inscription using Positional metric which is used to solve the problem occurred in stone inscription. G. Janani et al. [3] introduced to recognize Tamil inscription characters using Image processing techniques. The architecture of this system included image acquisition, noise removal, binary conversion, morphological operation, connected component, feature extraction, segmentation and matching the characters. N Jayanthi and S Indu [4] stated to recover an ancient inscription using Bag of Visual Words (BoVW) technique.

Antony P J and Savitha C K [1] determined to recognize handwritten south Dravidian Tulu script. Preprocessing, feature extraction, learning, classification and recognition, and mapping were the five modules to recognize handwritten

Tulu characters. Saikat Roy et al. [9] presented to identify handwritten Bangla characters using deep neural network. To recognize compound characters, Deep Convolutional Neural Network (DCNN) was compared with supervised Layer wise Deep Convolutional Neural Network (SL-DCNN).

A.S. Kavitha et al. [7] developed to segment the text from historical document images. Optical Character recognition (OCR) was used to recognize the Indus script images. K. Rajan et al. [11] organized this application to classify Tamil language documents automatically using Vector Space Model (VSM) and Artificial Neural Network (ANN).

III. Methodology

Tamil character recognition is technically a challengeable mission due to century deviation in Tamil language written style and analogy in character. Modern Tamil characters are reoriented from ancient Tamil characters, which lead to the researcher's technically facing difficulty in character classification and recognition. A small number of people are attentive with the ancient Tamil character whose efforts are to translate them into on paper booklets manually. So, digital system has to be developed to preserve the treasure evidence given by our progenitors for future generation. Following steps has been initiated to segment the characters and to measure the features. Figure 1 illustrates the overall flow of the proposed work.

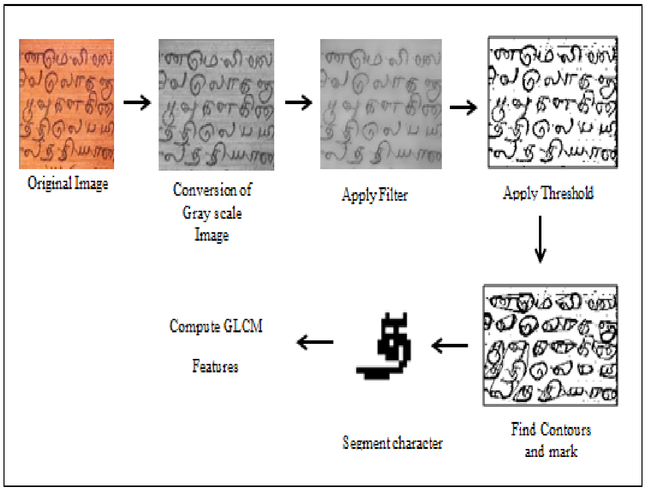


Figure 1. Overall flow of the work

A. Grayscale Conversion

The original images are transformed to grayscale image and the normalization is smeared to adjust the intensity value of pixels. Figure 2 demonstrates the palm leaf input image and Figure 3 demonstrates grayscale transformation of palm leaf manuscript.

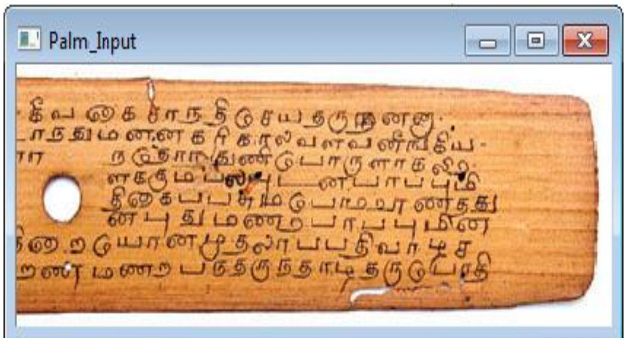


Figure 2. Palm Leaf input image

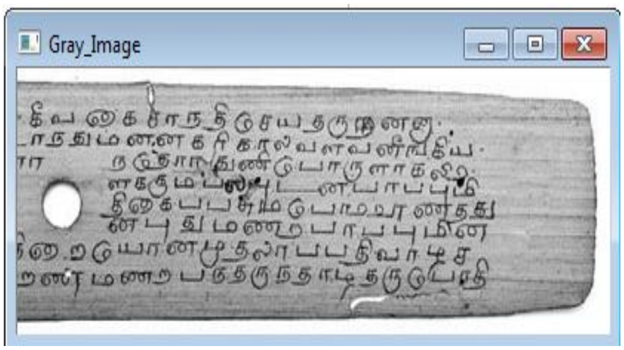


Figure 3. Grayscale image

B. Filter

Filtering is the methodology to adjust or enrich an image, which accomplish neighborhood pixel operations. Here three types of filters are used: Median, Bilateral and Gaussian filter to correlate which filter is appropriate for an image.

1) *Median Filter:* It is a non-linear filter to diminish noise from an image as a preprocessing. This filter is broadly used one, because it will safeguard edges sharp while reducing the noise.
2) *Bilateral Filter:* It is also a non-linear filter and safeguards the edges. It smoothens an image by declining noise and it changes the intensity of every single pixel with biased average of neighborhood pixels.
3) *Gaussian filter:* It is linear filter which will blur and decline the contrast of an image while dropping noise.

From these three filters median filter executed well and for this planned work, median filter is nominated to eliminate noise from the image. Figure 4 illustrates evaluation of median, Gaussian and bilateral filter output for the palm leaf input image.

PSNR, MSE and SSIM are used to measure the quality of reconstructed image.

Peak Signal-to-Noise Ratio (PSNR) is generally used for quality measure for recreated image, is the proportion among highest probable of signal and the noise that distresses the image. The signal states to unique image and the error presented by recreated image. Higher the PSNR value designates high quality of an image; it may miscarry for some images. PSNR is computed using (1).

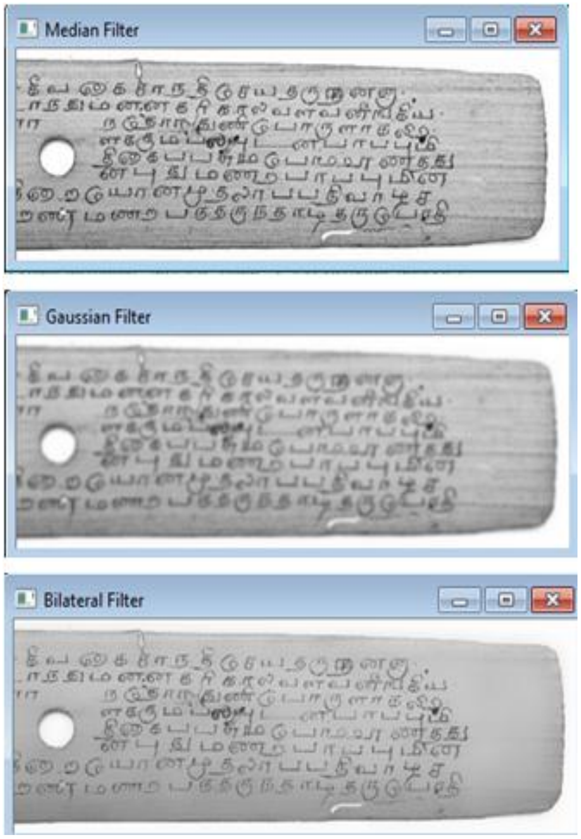


Figure 4. Comparison of Median, Gaussian and Bilateral filter

$$\mathrm{PSNR} = 10.\log_{10}\Big(\tag{1}$$

Here H refers to highest probable of signal and E refers to Mean Square Error (MSE). MSE is computed using (2).

$$MSE = \frac{1}{pq}\sum_{m=0}^{p-1}\sum_{n=0}^{q-1}[O(m,n) - A(m,r \tag{2}$$

where O(m, n) is the original image, A(m, n) is the recreated image, p and q are the dimension of an image. MSE is the cumulative squared error between the original image and recreated image, lower value of MSE refers lesser error in an image.

Structural Similarity Index (SSIM) is a metric used to compute the image quality which includes three properties: luminous, contrast and structure. SSIM is used to measure the similarity between two images using (3) refer Table.1.

$$SSIM(i,j) = [l(i,j)]^{\alpha}.[c(i,j)]^{\beta}.[s(i,j \tag{3}$$

C. *Threshold*

Thresholding is the method of segmenting foreground and background of an image. For executing threshold operation the image must be converted to grayscale image. Here three kinds of threshold performance have been used to elite which executes well. The three threshold approaches are simple threshold, Otsu threshold, Adaptive threshold; adaptive mean threshold and adaptive Gaussian threshold.

1) *Simple threshold:* If the pixel value is superior to the threshold value, then it will be allotted to one value (white) otherwise another value (black).
2) *Otsu threshold:* It robotically calculates a threshold rate from image histogram.
3) *Adaptive Threshold:* The global threshold may not be appropriate for all instance, so adaptive threshold were used. It figures out the threshold for actual region of the image, so distinctive region will acquire different threshold values.
   a. *Adaptive mean threshold:* Threshold value is computed by using the mean of the adjacent pixel region.
   b. *Adaptive Gaussian threshold:* The threshold value is computed by weighted aggregate of neighbor pixel values, where the weight represent Gaussian kernel.

From this three threshold methods, adaptive threshold achieved improved result and adaptive mean threshold is nominated for this proposed work. Figure 5 displays the evaluation of threshold methods.

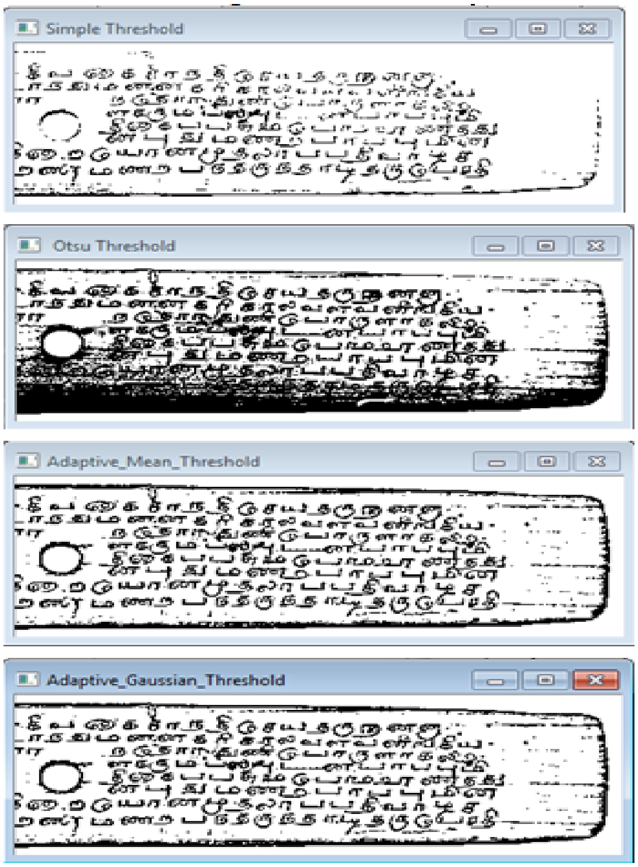


Figure 5. Comparison of threshold methods

D. Segmentation

Segmentation is the procedure of segregating image into dissimilar segments. The concluding outcome of the segmentation will be the gathering of segments that comprises the set of complete image. Threshold action should be done earlier to segmentation. The segmentation can be categorized as line segmentation, word segmentation and character segmentation. In this paper the foremost aim of the effort is to segment the characters from palm leaf manuscript. Palm leaf manuscript was inscribed with merged components and cannot isolate the word by nature. So here the characters were segmented from an image using contours. After thresholding, contours are computed and noticed the contour to segment. All the segmented characters are stored in the organized system folder. Figure 6 demonstrates the contours marked for segmentation.

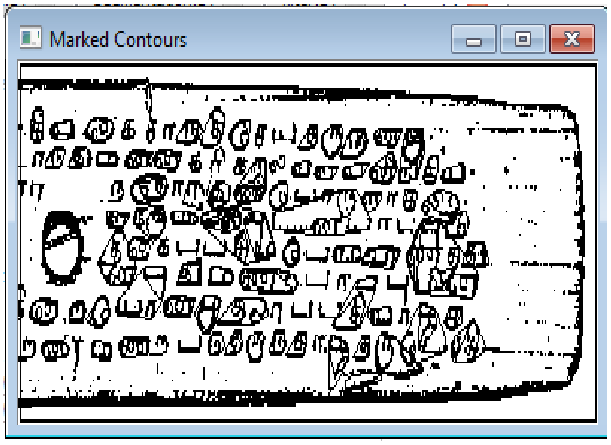


Figure 6. Contours marked for segmentation

The drawback of this segmentation is some of the connected components have to be segment further, the blank space and pits between the characters are also segmented which lead to unwanted space storage. This has to be rectified in imminent work. Figure 7 displays segmented character.

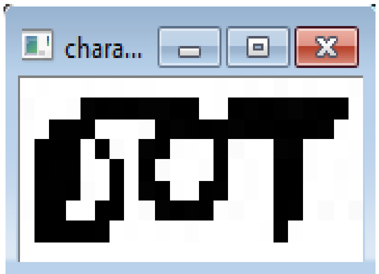


Figure 7. Segmented character

IV. FEATURE EXTRACTION

1) GLCM Feature Extraction:

In the proposed method Gray Level Co-occurrence Matrix (GLCM) has been used for feature extraction, which is also called as Gray Tone Spatial Dependency Matrix (GTSDM) [10]. Justification is needed here GLCM is the second order statistical degree of texture computation. Second order means which dealing with relation between two classes of pixels, where the first order will also compute the statistical degree but do not contemplate the relation between two classes. Up to here GLCM has mainly two steps: compute co-occurrence matrix and find the texture features. The parameters used here are distance and orientation. Distance is taken as d which can be d=1,2,3,..etc. and orientation can be represented in eight directions θ= 0, 45, 90,135,180,225, 270,315. The frequently used directions are 0 (horizontal), 45 (front diagonal), 90 (vertical) and 135 (back diagonal).

Steps to compute GLCM matrix:

a) For computing co-occurrence matrix:

Step 1: Get the pixel values from an image.

Step 2: Fix the neighbor pixel and reference pixel to perform matrix manipulation.

Step 3: Fix the distance and orientation. Here d=1 and θ= 0.

Step 4: Find the co-occurrence matrix for image pixel

Example:

Step 5: Calculate GLCM matrix, then calculate sum of elements of resultant matrix to perform normalization, which is 16.Divide all the elements of resultant matrix with 16 to obtain horizontal GLCM.

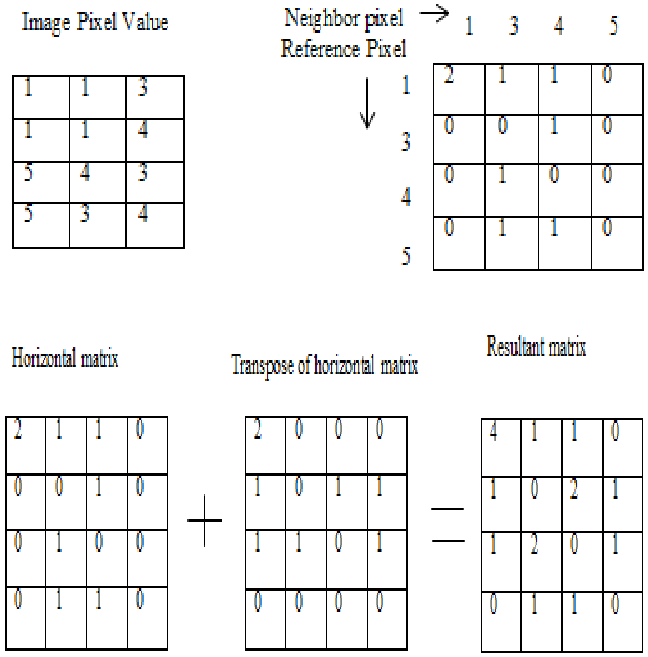


Image Pixel Value

1	1	3
1	1	4
5	4	3
5	3	4

Neighbor pixel →
Reference Pixel ↓

	1	3	4	5
1	2	1	1	0
3	0	0	1	0
4	0	1	0	0
5	0	1	1	0

Horizontal matrix

2	1	1	0
0	0	1	0
0	1	0	0
0	1	1	0

+

Transpose of horizontal matrix

2	0	0	0
1	0	1	1
1	1	0	1
0	0	0	0

=

Resultant matrix

4	1	1	0
1	0	2	1
1	2	0	1
0	1	1	0

Horizontal GLCM

0.25	0.06	0.06	0
0.06	0	0.12	0.06
0.06	0.12	0	0.06
0	0.06	0.06	0

b) For finding texture feature for corresponding GLCM:

Step 6: Calculating Dissimilarity,

$$D = \sum_{m,n=0}^{k-1} P_{(m,n)} |m - \qquad (4)$$

$$\begin{array}{ccccc} (m\text{-}n) & & \text{Horizontal GLCM matrix} & & \\ \begin{bmatrix} 0 & 2 & 3 & 4 \\ 2 & 0 & 1 & 2 \\ 3 & 1 & 0 & 1 \\ 4 & 2 & 1 & 0 \end{bmatrix} & * & \begin{bmatrix} 0.25 & 0.06 & 0.06 & 0 \\ 0.06 & 0 & 0.12 & 0.06 \\ 0.06 & 0.12 & 0 & 0.06 \\ 0 & 0.06 & 0.06 & 0 \end{bmatrix} & = & \begin{bmatrix} 0 & 0.12 & 0.18 & 0 \\ 0.12 & 0 & 0.12 & 0.12 \\ 0.18 & 0.12 & 0 & 0.06 \\ 0 & 0.12 & 0.06 & 0 \end{bmatrix} \end{array}$$

In conclusion, sum all the elements to attain dissimilarity,
D= 1.2

Step 7: Similarly find texture feature for other GLCM statistical feature measure.

In this paper GLCM features were computed using the distance, d=1 and angle θ= 0. Contrast, Energy, Homogeneity, Correlation, Dissimilarity and Angular Second Moment (ASM) are the most generally used features. Table.2 clarifies the above mentioned feature details for characters of palm leaf manuscript, stone inscription, handwritten document and for printed document.

Contrast: It categorized underneath contrast group measure and is also known as sum of squares variance. If m is equal to n, then there is no contrast. If m is not equal to n, then it is on diagonal and increase progressively according to values.

$$C = \sum_{m,n=0}^{K-1} P_{(m,n)} (m - \qquad (5)$$

Energy: Energy categorized under orderliness measure, which is used to compute the regularity of an image using.

$$E = \sqrt{\sum_{m,n=0}^{k-1} (P_{} } \qquad (6)$$

Homogeneity: Homogeneity categorized under contrast group measure and also known as Inverse Difference Moment (IDM). It produces high rate for uniformity image and produces small rate for unrelated image.

$$H = \sum_{m,n=0}^{k-1} \frac{P_{(m,}}{1+(m\text{-}} \qquad (7)$$

Dissimilarity: It is also categorized under contrast group measure and weights of the dissimilarity upsurges linearity.

$$D = \sum_{i,j=0}^{k-1} P_{(m,n)} |m - \qquad (8)$$

Angular Second Moment (ASM): ASM is categorized under orderliness and produce high value when it is highly well-ordered. The maximum value is 1 for undistinguishable image.

$$A = \sum_{m,n=0}^{K-1} (P_{m,} \qquad (9)$$

V. Experimental Results and Discussion

Dataset for palm leaf were taken from Tamil Nadu Archaeological Department, containing mesa lagnam leaf 40 parts, virsabha lagnam leaf 26 parts, mithuna lagnam leaf 22 parts, kataka lagnam leaf 22 parts, simha lagnam leaf 18 parts, thula lagnam leaf 23 parts, virchika lagnam leaf 28 parts, thanoor lagnam leaf 15 parts, magara lagnam leaf 13 parts, gumpa lagnam leaf 25 parts and meena lagnam leaf 14 parts. Each part contains 10 leaves, totally 2460 palm leaf images with dimension 198 x 2063.

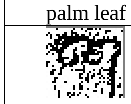
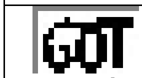
Table.1 shows the comparison of different filters with quality measure (PSNR, MSE and SSIM). SSIM=1 refers that both the images has no difference and SSIM ≥ 0.8 refers somewhat similar between two images, SSIM < 0.8 refers noisy image and SSIM <0.7 refers blurred image. From Table.1 PSNR value is highest for median filter and SSIM produced maximum similarity (0.9) between images. MSE for median filter is lesser compared to other filters. So, median filter is selected to decrease the noise from an image.

Table.1 Comparison of quality measures with filters.

	Palm Leaf Image			**Stone Inscription Image**			**Handwritten Image**			**Document Image**		
	PSNR	***MSE***	***SSIM***	***PSNR***	***MSE***	***SSIM***	***PSNR***	***MSE***	***SSIM***	***PSNR***	***MSE***	***SSIM***
Median	27.81	107.65	0.99	21.48	461.50	0.94	36.02	16.27	0.99	30.09	63.75	0.99
Gaussian	21.14	500.01	0.70	17.10	126.56	0.68	25.45	185.40	0.72	16.22	154.28	065
Bilateral	21.51	459.50	0.73	19.23	775.31	0.78	25.14	199.04	0.68	22.87	335.28	0.64

In forthcoming work, from these statistical features, it needs to be validated that the characters segmented from different Tamil manuscript and inscription images are similar or not and to classify and recognize. If it is comparably same character then the statistical measure of correlation will be exactly 1, if it is somewhat related then the correlation value will lie between 0.5 to 0.9, if there is no similarity then there is correlation relationship between characters.

Table 2. Comparison of GLCM feature for different images for Tamil character.

Measures	C	E	H	D	ASM
 Sample original character	1403.445	0.590	0.741	9.762	0.348
 palm leaf	16233.075	0.267	0.439	64.853	0.071
 stone inscription	18996.281	0.243	0.395	75.715	0.059
 handwritten	15443.393	0.279	0.429	61.849	0.077
 printed document	19678.273	0.246	0.410	78.301	0.061

VI. CONCLUSION AND FUTURE ENHANCEMENT

The foremost intention of segmenting Tamil characters is for recognition purpose. It is meritoriously satisfied using GLCM feature extraction which measures statistical data of segmented characters. The core difficulty in segmenting and recognizing is a challengeable task in palm leaf manuscript due to more connected components and recognizing ancient Tamil font styles. In this paper, the evaluation is studied whether the segmentation with GLCM feature works for all categories of Tamil manuscript or not. Various filters and threshold mechanisms are implemented and the characters are segmented as individual character from the images and stored it in system folder. The result for this proposed work was implanted with the system with Windows 7, Intel (R) Pentium(R) CPU 2127U @ 1.90GHz processor and 4 GB of RAM. Images were processed using Spyder, Anaconda version 3 (Python 3.7).

The main problem is some connected components are present even after segmentation. This has to be resolved in future and then the segmented characters obligate to be transformed into text document so that the continuation of the characters will be more expressive for enhanced understanding. The work will extend with PCA (Principal Component Analysis) to reduce the dimensionality and to extract the enhanced feature for classification, which is noteworthy to highlight similarity and dissimilarity between the images [12]. So, in future the work will be implemented in Deep learning Neural Network using Convolutional Neural Network (CNN) with PCA to recognize the Tamil characters in a superior and efficient way.

REFERENCES

[1] Dr. Antony P J and Savitha C K, "*A framework for recognition of handwritten south Dravidian Tulu script*", Conference on Advances in Signal Processing(CASP), IEEE, pp. 7 - 12, 2011.

[2] Mrs. G. Bhuvaneswari and Dr. V. Subbiah Bharathi, "*An Efficient Positional algorithm for recognition of Ancient Stone Inscription Characters*", International Conference on Advanced Computing(ICoAC), IEEE, pp. 1 - 5, 2015.

[3] G. Janani, V. Vishalini and Dr. P. Mohan Kumar, "*Recognition and Analysis of Tamil Inscriptions and Mapping using Image Processing Techniques*", International Conference on Science Technology Engineering and Management(ICONSTEM), IEEE, pp. 181 - 184, 2016.

[4] N Jayanthi and S Indu, "*Inscription Image Retrieval using Bag of Visual words*", ICMAEM, pp. 1 - 7, 2017.

[5] Er. Kanchan Sharma, Er. Priyanka, Er. Aditi Kalsh and Er. Kulbeer Saini, "*GLCM and its feature*", International Jorunal of Advanced Research in Electronics and Communication Engineering(IJARECE), Vol. 4, Issue. 8, pp. 2180-2182, 2015.

[6] Karthigaiselvi. M and T. Kathirvalavakumar, "*Recognition of words in Tamil script using Neural Network*",International Journal of Computer Research and Apllication, Vol.7, Issue.3, pp. 62-70, 2017.

[7] A.S. Kavitha, P. Shivakumara, G.H. Kumar and Tong Lu, "*Text Segmentation in degraded historical document images*", Egyptian Informatics Journal, Vol.17, Issue.2, pp. 189 - 197, 2016.

[8] Kavitha Subramani and Dr. S. Murugavalli, "*A Novel Binarization method for degraded Tamil palm Leaf image*", International Conference on Advanced Computing(ICoAC), IEEE, pp. 176 - 181, 2016.

[9] P. Rajan and S. Sridhar, "*Identification of Ancient Tamil Letters and Its characters: Automatic date fixation based on Contour-Let technique*", Association for Computing Machinery (ACM), pp. 40 – 43, 2017.

[10] Saikat Roy, Nibaran Das, Mahantapas Kundu and Mita Nasipuri, "*Handwritten Isolated Bangla compound character recognition: A new benchmark using a novel deep learning approach*", Pattern Recognition Letters 90, Vol. 90, pp. 15 - 21, 2017.

[11] Shijin Kumar. P.S and Dharun V.S, " *Extraction of Texture features using GLCM and shape Features using connected regions*", International Journal of Engineering and Technology(IJET), Vol. 8, No. 6, pp. 2926-2930, 2017.

[12] M. Sornam and C. Vishnu Priya, " *Deep Convolutional Neural Network for Handwritten Tamil Character Recognition using Principal Component Analysis*", Intl. Conf. on Next Generation Computing Technologies (NGCT 17), University of Petroleum and Energy Studies, Dehradun,India, pp. 102-112, 2017.

Authors Profile

Dr M. Sornam received her MSc in Mathematics from the University of Madras in the year 1987, Master's Degree in Computer Applications from the University of Madras in the year 1991 and received her Ph.D from the University of Madras in the year 2013. Since 1991–1996, she worked as a Lecturer in Computer Science at Anna Adarsh College, Chennai. Later, from 1996 to 2000 she worked as a Lecturer in Computer Science at T.S. Narayanasami College of Arts and Science, Chennai. Since 2001, she has been working in the Department of Computer Science, University of Madras. At present, she is working as an Associate Professor in Computer Science at the University of Madras. Her area of interest includes artificial intelligence and artificial neural networks , image processing, data mining, pattern recognition and applications.

Miss. Poornima Devi. M pursed Bachelor of Science in Soka Ikeda College for Women from University of Madras, India in the year 2013, Master of Science in Queen Marys College for Women from University of Madras, India in the year 2015 and Master of Philosophy in University of Madras, India in the year 2016. She is currently pursuing Ph.D. in Department of Computer Science, University of Madras, India since 2017. Her main research work focuses on Artificial Intelligence, Artificial Neural Network, Image Processing, Deep Learning Neural Network.