Applications of Image Processing Techniques on Palm Leaf Manuscripts - A Survey

Nagendra Panini Challa, R. Vasanth Kumar Mehta
1,2 SCSVMV University, Enathur, Kanchipuram, Tamil Nadu
Email: cnpanini@kanchiuniv.ac.in, vasanth.mehta@kanchiuniv.ac.in

Received: 5th June 2017, Accepted: 15th June 2017, Published: 1st September 2017

Abstract
In India there are many organizations/institutions committed for protection of ancient palm leaf manuscripts [8] in order to store our precious knowledge writings. As the time passes by, these palm leaves are getting spoiled by artificial and natural elements. With the growing technology worldwide, we are aimed to digitize all these palm leaf manuscripts which are available in our university library. One such objective is to develop an efficient image processing system for effective retrieval of metadata automatically from these manuscripts. Many image processing techniques have been proposed for the efficient data retrieval. Some of the techniques include image enhancement, segmentation, processing, restoration, compression and acquisition. All these techniques have various algorithms which are implemented successfully with desired results. Since one hundred percent correct recognition rate is impossible with noise and other distorted mediums determining the right technique for image processing is very difficult. So, this paper contains information about some image processing techniques and algorithms applied on palm leaf manuscripts proposed by different authors for successful data retrieval.

Keywords: Palm Leaf Manuscripts; Digitizing; Cataloguing; Metadata; Digital Image Processing; Automatic Extraction

1. Introduction
Manuscript means any information that is handwritten. It is generally related to ancient writings of the past men. Palm leaflets [5] are one of the oldest and cheapest materials used for writing. Palm leaf manuscripts (Figure 1) are produced from two main types of palms namely palmyra and talipot. Although palm leaf manuscripts size differs among other parts of the world, the standard size is 48 and 4 centimeters in length and width.

As the time passed by many palm leaves are under severe extinction from various other external elements such as dampness, fungus, ants and cockroaches and many more. SCSVMV University started preserving its Palm Leaf Manuscripts for future generations and made them available in the university library by digitizing them and storing them accordingly. It has a collection of ~5000 manuscripts collected from various sources.

The collection mostly comprises of manuscripts written in Grantha Script (~80%) which belong to Tamil, Telugu, Hindi, Kannada, and Malayalam languages.

Fig.1. Palm Leaf Manuscript (Source: Manuscripts Section, SCSVMV University)

Each bundle of leaves is usually tied together with cord threads through two holes pierced through the entire manuscript by the insertion of bamboo strips. The resultant bundle is completed by adding the heavy wooden covers at the either sides of the leaves, also tied by the cords or wrapped with a soft textile cloth. The total number of manuscripts available in SCSVMV University manuscripts section can be visualized in Fig 2.

Fig.2. List of Manuscripts available in SCSVMV University Manuscripts Section [10]

1) Image Enhancement
The Authors in [4] mainly focused on image processing techniques which offer an approach selection to counter the degradation quality and make them readable. He classified them to three methods as follows:
(a) Binarization/ Thresholding method
(b) Hybrid Binarization/ Thresholding and other methods
(c) Non-Threshold based methods
Table 5: Challenges and Solutions of Author [4]

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>To counter the palm leaf manuscripts degradations and make them readable</td>
<td>Enhancing the image using Thresholding and non Thresholding based techniques</td>
</tr>
</tbody>
</table>

In IEM I the entropy based method to segment from original image to binary image. This is initiated by threshold values of background and foreground. The adaptive threshold is now applied to every pixel. Firstly, the gray level value for every pixel of image is calculated. Then, local threshold for every pixel is given by

$$f(x,y) = \begin{cases} 
0 & \text{if } g(x,y) < t(x,y), \text{otherwise} \\
255 & 
\end{cases}$$

Fig.3. Image enhancement using Thresholding [4]

Next is Sauvola’s method where threshold \( t(x,y) \) is calculated using the mean and standard deviation (SD). The pixel intensities in a window which is done around the \((x,y)\) pixel. The formula is given as:

\[
t(x,y) = m(x,y) \left[ 1 + k \left( \frac{s(x,y)}{R} - 1 \right) \right]
\]

Here \( R \) is the maximum value for SD (\( R = 128 \)) and \( k \) contains positive values between 0.2 and 0.5. \( m(x,y) \) is local mean and \( s(x,y) = SD \).

The estimation of foreground is done by rough estimation approach after preprocessing task. The binary image is extracted from the foregrounds of image.

Along with this, to separate character from a damaged manuscript background is implemented using Otsu Algorithm.

Fuzzy logic approach [23] is also applied to recognize the script where the equality of the character is given as:

Equality rule: \( S_i (c) = \frac{\sum_{x,y} W(i,j)A_i(x,y)C_i(j(x,y))}{\sum_{x,y} W(i,j)(x,y)} \)

Aspect rule: \( R_j (c) = \min \left( \frac{ar(c)}{ar(j)} - \frac{ar(i)}{ar(c)} \right) \)

Table 2: Challenges and Solutions of Author [2]

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Segmentation of manually written palm leaf manuscripts</td>
<td>The character image is obtained by below 3 steps: Background, Line and Character elimination</td>
</tr>
</tbody>
</table>

Image conversion [2]: Here the image is converted from RGB image to grey scale. When we convert it it depends on the sensitivity response curve detector of light as a function of wavelength. The equation is:

\[ Y = 0.3R + 0.59G + 0.11B \]

Fig.5. Grey Image

Reduction of Noise: The process of eliminating noise from an image is called as Gaussian filtering. The process to apply on each pixel in the image is given as follows:

\[ F(i,j) = 1 / S_{mk} \sum_{m=1}^{k} B_{mn} * F_{gr}(m,n) \]
Fig.6. Grey scale image before and after noise reduction

**Background Elimination:** The main idea is to find an optimal threshold that divides the image objects by finding the two classes from any grey level.

Fig.7. Otsu’s threshold value method

The optimal threshold is calculated by $n = \sigma^2_b$ where $\sigma^2_{\text{Thr}} = \text{total variance}$, is not related to any grey level but only to reduce $\sigma^2_B$ which is within-class variance.

Fig.8. Binary Image after background elimination

**Line Segmentation:** The horizontal PPA is adopted by the author because the text in many images are scaled along the horizontal row. It computes black pixels for each column of the image. When the horizontal projection profile is applied image, a column vector is obtained. Elements of this column vector are the sum of pixel values in each row of the document image.

**Character Segmentation:** In this paper as a final step, the extracted lines are segmented into characters. To find the boundaries between the characters, he applied a threshold value on the length of the space in-between the characters.

Fig.9. Line Segmentation & Character Segmentation

Fig.10. Structure of Text Line Segmentation Technique [5]

The images are analysed with adaptive histogram after normalization. The quality of manuscript characters is more enhanced accurately. Later image line segments are divided by sliding window operation. The skewing operation is performed on the line images for accuracy.

4) **Data Retrieval**

The characters of palm leaf have depth property. The 3D features on Telugu language are extracted using Decision Tree approach. Generally, Telugu has more than 62 million speakers around the globe.

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>To extract 3D features and how to use them in the recognition and classification process</td>
<td>The characters from Telugu palm leaf manuscripts are classified and identified by using SEE5 algorithm of decision tree approach.</td>
</tr>
</tbody>
</table>

[3] The complexity of the characters is reduced by dividing all the characters using printed characters and their coefficient values. The depth of the pixel is directly proportional to the pen/stylus pressure on the palm leaf. The pixel points for every character uses the concept of critical points.

Fig.11. Maximum and Minimum Critical Points

**SEE5 Algorithm:**

a) The attributes of the pixel are identified.
b) Every attribute is separated by a comma and a class label.
c) The training and test data set are made by concatenating each character.

Step: 1- Start
Step: 2-X, Y, Z Coordinates of each pixel are extracted from palm leaf.
Step: 3-Repeat previous step, for every 5 samples.
Step: 4- 3 samples are used for training and other 2 are used as test samples.
Step: 5- SEE5 tool and training set the system is trained.
Step: 6- SEE tool is used to test set database.
Step: 7- Following results are obtained:
   a) Decision Tree
   b) Evaluation on training data
   c) Attribute Usage
   d) Evaluation on test data
Step: 8- End

5) Image Acquisition
The author observed that the mouse started to eat some part of these manuscript folios. The mouse did not destroy the entire manuscript but only ate about 5-10% away from each leaf. After the pre-processing stage the left parts were cropped, and saved as different image files. The boundary coordinates of the leaves are obtained where Fourier descriptors are applied and invariant boundary intersection-based shape descriptor are used to indicate the identity of different leaves.

Fig.12. Front face of a leaf with hole for bundling up. Left side is damaged by the mouse

### Table 4: Challenges and Solutions of Author [1]

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>The manuscript bitten by mouse should be re sequenced</td>
<td>Fourier Descriptor Algorithm, Rotation-Translation Invariant Boundary Algorithm.</td>
</tr>
</tbody>
</table>

Table 4: Challenges and Solutions of Author [1]

Fig.13. Proposed Orthographic view of Author [1]

Connectivity mode inner boundary tracing method is used as a starting point to segment the leaves from image background. The Fourier descriptors, based on fourier analysis of the function derived from the boundaries. Therefore the boundary points are represented as \((x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots \ldots \ldots, (x_n, y_n)\). The author represents the coordinate pair as complex numbers. Thus reducing a 2D problem to 1D we get the below equation

\[
S(k) = x(k) + jy(k), k=0,1, \ldots \ldots N-1 \quad \text{---------(1)}
\]

The euclidean distances among the leaves were calculated and an overall symmetric matrix was generated.

To evaluate the information, the author used evaluating of shape data using Tree search algorithm. In the final step, a Tree Search scheme that starts from the 18th (fixed) leaf and ends at the 66th leaf gives the most probable sequence as shown in Fig 13. Every node in the tree was defined as a leaf and linked to the neighbour leaves.

### Author | Title | Methodology Used
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Akca , A. Gruen</td>
<td>Re-Sequencing Palm Leaf Manuscript using Boundary-Based Shape Descriptors</td>
<td>Fourier Descriptors, rotation-translation invariant boundary, intersection based shape descriptor algorithm, inner boundary tracing algorithm</td>
</tr>
<tr>
<td>Olarik Surinta and Rapeepon Chamchong</td>
<td>Image Segmentation of writings from palm leaf manuscripts.</td>
<td>Background elimination by Otsu’s algorithm, Line and Character segmentation.</td>
</tr>
<tr>
<td>P Narahari Sastry and Ramakrishna</td>
<td>Classification and Identification of written Telugu Characters using Decision Tree Approach</td>
<td>The characters from Telugu palm leaf manuscripts are classified and identified by using SEE5 algorithm of decision tree approach</td>
</tr>
<tr>
<td>Htwe Pa Pa Win, Phyo Thu Thu Khine, Khim Nwe Ni Tun</td>
<td>Bilingual OCR System for Myanmar and English Scripts with Simultaneous Recognition</td>
<td>Image Enhancement Methods using Binarization, Thresholding and non thresholding techniques</td>
</tr>
</tbody>
</table>

### Conclusion
Although a lot of research works exist in the field of image processing in context with palm leaf manuscripts, but there is a need to prepare an efficient database (DB) for the automation and digitalisation of these manuscripts which can be accessed by all the researchers worldwide (such as a manuscripts DB). All the techniques discussed above gave good results but more work has to be done on Indian languages due to large character set and linguistic features. From this study, few data sets of manuscripts are used for obtaining the test results but there is a need to train and test huge data sets.

Therefore, it is necessary to prepare an efficient manuscript DB which can automatically generate the metadata from palm leaf manuscripts.
References

[1] D. Akca, A. Gruen, Re-sequencing a historical palm leaf manuscript with boundary based shape descriptors. Institute of Geodesy and Photogrammetry, Swiss Federal Institute of Technology (ETH) Zuerich, ETH Hoenggerberg, CH 8093 Zuerich, Switzerland.


[11] Rapeeporn Chamchong, Chun Che Fung, Optimal Selection of Binarization Techniques for the processing of Ancient Palm Leaf Manuscripts, School of Information Technology, Murdoch University 90, South Street, Murdoch, Western Australia 6150, 2010.


