A Modified Wellner’s based Binarization on Ancient Malayalam Documents

B.J. Bipin Nair
Department of Computer Science, Amrita School of Arts and Sciences, Mysuru Campus, Amrita Vishwa Vidyapeetham, India. E-mail: bipin.bj.nair@gmail.com

S. Yadhu Krishnan
Department of Computer Science, Amrita School of Arts and Sciences, Mysuru Campus, Amrita Vishwa Vidyapeetham, India. E-mail: yadhukris141@gmail.com

A.M. Akarsh
Department of Computer Science, Amrita School of Arts and Sciences, Mysuru Campus, Amrita Vishwa Vidyapeetham, India. E-mail: akarshnc18@gmail.com

Nakul S Anand T
Department of Computer Science, Amrita School of Arts and Sciences, Mysuru Campus, Amrita Vishwa Vidyapeetham, India. E-mail: nakulsanand@gmail.com

Pravin Sasikumar
Department of Computer Science, Amrita School of Arts and Sciences, Mysuru Campus, Amrita Vishwa Vidyapeetham, India. E-mail: pravinsk98@gmail.com

Received May 15, 2021; Accepted August 18, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI05/WEB18243

Abstract

Preservation of precious knowledge present in Malayalam literature is a tedious task because of the presence of a huge amount of degradations in historical documents. One way to save these documents is to enhance the Malayalam manuscripts and storing them electronically. Here we are using historical Malayalam documents like poems, agreement copies and palm leaves as experimental dataset. In our proposed work is a novel binarization model which is based on modified Wellner’s algorithm. The degraded input image is converted into an integral image and then a modified version of Wellner’s algorithm is applied to it. This would enable us to enhance the document and further proceed with remaining phases of OCR. In binarization mainly we are focusing on the challenges from degraded documents like Non-uniform Background Illumination, Stains, bleed through etc. In our work we are developing a model which effectively binarizes the degraded Malayalam documents, especially the challenges like oil stain, smudge, and uneven illuminations. Finally, we will be able to validate and check the
accuracy of the proposed model effectively. The proposed algorithm yielded an accuracy of 92%.

Keywords

Wellners, OCR, Binarization, Malayalam.

Introduction

Many precious and valuable Malayalam historical documents were lost because of degradation and other environmental factors. Hence it is quite important that the text present in these documents should be preserved in a non-degradable environment. Enhancing the ancient Malayalam manuscripts is the only way that we can store the documents electronically for a long period of time. Preprocessing can remove the various challenges like ink bleed, illumination, contrast variation etc. Efficient binarization will improve the quality and readability of various Malayalam manuscripts. Clustering based binarization techniques like K-means applied on Korean dataset to remove text stroke which gives an accuracy of 85% (Anh-Nga Lai et al., 2008). A rapid binarization with morphological operators on document images to remove shades and uneven illumination that gives above-average accuracy (Eria Cooksey et al., 2008). An enhanced adaptive binarization on document image to remove ink-bleed and uneven illumination gives an average accuracy (Sridhar Cherala et al., 2008). Modified Adaptive Binarization technique which is applied on document images to remove uneven illumination which gives an accuracy of 94.61% (B. Gatos et al., 2009). A combined approach for local and global binarization method to remove uneven illumination from document images which gives a high accuracy of 99.26% (Mehdy Bohlool et al., 2010). Morphology based binarization method on DIBCO-11 to remove faint text and shade that gives an accuracy of 92.9% (George Carayannis et al., 2012). A local thresholding based binarization method on document image to remove ink bleed, stain gives an average accuracy (V. Rabeux et al., 2013). Combined approach of local and global binarization methods applied on Persian heritage image to remove bleed through, uneven illumination, stain which gives an accuracy of 79.89% and 89.58% (Hossein Ziaei Nafchi et al., 2013). An automatic global binarization method removing ink-bleed, show through from DIBCO-11 that gives an accuracy of 97.9% (Jon Parker et al., 2013). Gabor filters based binarization applied on DIBCO to remove ink-bleed, poor contrast that gives 89% (Abdenour Seha, et al., 2014). Phase-based image binarization of ancient documents and DIBCO to remove uneven illumination and manual correction that provides an accuracy 95.6% (Hossein Ziaei Nafchi, et al., 2014). A novel document image binarization technique is used to correct contrast construction from document image produced an average accuracy (S. Tamilselvan et al., 2014). A combined approach for
Gaussian and Markov binarization method for degraded documents and DIBCO to remove uneven illumination and shade that gives an accuracy of 93% (Shujing Lu et al., 2014). A multi-spectral document image binarization applied on DIBCO to remove uneven illumination which gives an average accuracy of 79% (Nikolaos Mitianoudis et al., 2014). A signal reconstruction binarization method on noisy microscopy images to remove stain which gives an average accuracy (Thomas Lukes et al., 2014). Neutrosophic based binarization technique to remove stroke-like pattern from document image to get high accuracy of 95% (Khalid M Amin et al., 2014). A quad tree based binarization technique to remove stain, ink-bleed from degraded images that gives an outperformed accuracy of 90% (N Sobha Rani et al., 2014). Thresholding based binarization method to remove shade and uneven illumination from ICCICT old dataset that gives an accuracy of 95% (Mrs. Preeti. Kale et al., 2015). A combined approach of local and LSTM binarization method to remove ink bleed, uneven illumination from HDIBCO10 and DIBCO11 that gives an accuracy of 81.73 % and 92% (Muhammad Zeshan Afzal et al., 2015). Global binarization techniques which will help to binarize old halftone documents to remove stain, fadedness that gives an accuracy of 91.98% (Chandranath Adak et al., 2015). An improved global binarization method which removes stain and show through from HDIBCO2012 and DIBCO 2009 gives an average accuracy (Nilima Paul et al., 2016). An adaptive map document binarization method DIBCO which removes uneven illumination and ink-bleed that gives a low accuracy of 60% (B. Vishundharan et al., 2017). Localization and morphological operation binarization technique to remove ink-bleed, shade from document image that gives an accuracy of 90% (Jino P J et al., 2017). Adaptive binarization method which remove uneven illumination from document image which give an accuracy of 92% (Prashant Devidas Ingle et al., 2017). Global filter based binarization method which remove uneven illumination, ink bleed from historic documents gives 92% accuracy (Marcos Almeida et al., 2018). A combined approach of triangle multiple and clustering binarization approach is used to remove uneven illumination and shade from document image that gives the highest accuracy of 97% (Zunaidi Ibrahim et al., 2018). Histogram gradient features binarization on complex document image to remove the marginal noise from UMIACS Tobacco 800 and gives an average accuracy (N Shobha Rani et al., 2018). Noise reduction based binarization on DIBCO to remove ink bleed and show-through from degraded images gives an accuracy of 97% (N. Habibunnisha, et al., 2019). A combined approach of histogram and unique thresholding based approach on document image to remove uneven illumination gives an accuracy of 91.46% (Fitri Arnia et al., 2019). A local thresholding approach to remove ink-bleed and stain from palm-leaf manuscript gives as 80% accuracy (Chris Tensmeyer et al., 2020). A combined approach for CNN and wavelet binarization technique to remove ink stain and uneven illumination from old degraded Documents which has given
an accuracy of 97.02% (Younes Akbari et al., 2020). A combined approach for GMM and Monte Carlo binarization technique applied on DIBCO and Bickley Diary dataset to remove uneven illumination, stain shade gives an below-average accuracy (Robert Krupinski et al., 2020). A method using both global and local binarization technique to remove ink-bleed and show through from Malayalam and Kannada literature document gives 87% accuracy (BJ Bipin Nair et al., 2020). A combined approach for local and global thresholding binarization approach to remove uneven illumination and show through from old poems and books gives 94% accuracy (BJ. Bipin Nair et al., 2020).

The literature summary mainly focused on various document image binarization techniques. Most of the literature introduced a novel binarization method while some tried to improve existing methods like global and local thresholding. DIBCO documents were the most commonly used datasets. Ink bleed, uneven illumination, contrast variation, and smudges were some of the common challenges faced during the binarization stages. 99.26% was the highest accuracy recorded during this survey.

Proposed Method

The Fig.1 describes the proposed work as input image converted into an integral image. The integral image is run through the modified version of Wellner’s method. In the originally proposed method, only the rows of pixels are considered but in our method, both rows and columns of images are considered. The image is then considered as a 2D matrix. In the original method, they are considered as a 1D image since only the rows are considered. After applying the formula the resulting image is then converted into a binary greyscale image using a user-defined threshold value. The binary image obtained is displayed as the result.
**Proposed Algorithm**

1. INPUT image into ‘img’
2. SET rows and columns of img INTO ‘r’, ‘c’
3. CONVERT ‘img’ to Integral Image INTO ‘int’
4. SET ‘Thresh’ BETWEEN [0-100]
5. \[ ResImg[e, f] = int[e + r, f + c] - int[e, f + c] + int[e, f] - int[e + r, f] \]
6. DECLARE ‘grey’ AS GREYSCALE MATRIX WITH SIZE AS ‘r’, ‘c’
7. \[ binImg[grey \leq ResImg \times \frac{100-\text{Thresh}}{100}] = FALSE \]
8. DISPLAY ‘binImg’ AS OUTPUT IMAGE

The input image is stored in the variable ‘img’ and its rows and columns are stored in ‘r’ and ‘c’ respectively. The image is then converted into an integral image and stored in ‘int’. The image is then applied to the Modified version of Wellner’s algorithm where ‘e’ and ‘f’ stand for the difference between the rows and columns of the integral image and the rows and columns of the original image respectively. The resulting image stored in ‘ResImg’ is then applied to an equation that would implement the threshold value and convert it into a binary image stored in ‘binImg’.

This is a modified version of Wellner’s algorithm. The image is considered as two-dimensional matrix. Here, r and c stand for the number of pixels in rows and columns of the image. E and f stand for the difference between the rows and columns of integral image and the rows and columns of original image respectively. ‘int’ stands for integral image.

\[ ResImg[e, f] = int[e + r, f + c] - int[e, f + c] + int[e, f] - int[e + r, f] \] (1)

The resultant image from Equation(1) is converted into a binary image using Equation(2) Here, ‘grey’ is a greyscale image declared in the form of a matrix with height and width same as the input image. ‘ResImg’ is the output image from Equation.1. ‘Thresh’ is a user-defined value which should be between 0 and 100.

\[ binImg[grey \leq ResImg \times \frac{100-\text{Thresh}}{100}] = FALSE \] (2)
Table 1 Datasets Collection

<table>
<thead>
<tr>
<th>Source</th>
<th>Place/District</th>
<th>Type of Script</th>
<th>Number of Samples</th>
<th>Degradation Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nedumpally Mana</td>
<td>Palakkad, Kerala</td>
<td>Palm Leaf</td>
<td>200</td>
<td>Non illumination, Faded characters</td>
</tr>
<tr>
<td>Public Library</td>
<td>Wayanad, Kerala</td>
<td>Notebook</td>
<td>15</td>
<td>Ink bleed, Smudges</td>
</tr>
<tr>
<td>Puzhakkara mana</td>
<td>Thrissur, Kerala</td>
<td>Palm leaf</td>
<td>150</td>
<td>Non illumination, Brittleness</td>
</tr>
<tr>
<td>Public Library</td>
<td>Malappuram, Kerala</td>
<td>Notebook</td>
<td>20</td>
<td>Show Through, Non illumination</td>
</tr>
</tbody>
</table>

The above table shows us the datasets used in the proposed method. This dataset consists of over 300+ palm leaves and 35 notebook samples. The sources of these datasets are taken from 4 different districts of Kerala, India. The major challenges faced in these datasets are Ink Bleed, Non-Illumination, Smudges and Show Through.

Experimental Results

For the experiment around 300 datasets are used as input image which includes Palm leaf scripts, literature scripts, Astrology readings and Agreement copies. The experiment was also done on DIBCO 2017 datasets. The collected datasets are run through the proposed method along with three already existing binarization methods, namely Otsu, Niblack and Sauvola. The efficiency of the algorithms is measured and compared using MSE, PSNR, SSIM, Precision, Recall and F-measure.
Fig. 3 Describes experimentation done on Malayalam Literature using existing vs proposed algorithm
Fig. 4 Describes experimentation done on Agreement Copies using existing vs proposed algorithm
Proposed Method

Fig. 5 Describes experimentation done on DIBCO 2017 using existing vs. proposed algorithm

Table 2 Average F-measure value of existing Vs proposed algorithm

<table>
<thead>
<tr>
<th>INPUT IMAGE</th>
<th>OTSU</th>
<th>NIBLACK</th>
<th>SAUVOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Palm leaf</td>
<td>68.5</td>
<td>74.9</td>
<td>86</td>
</tr>
<tr>
<td>2-Palm leaf</td>
<td>65.9</td>
<td>76</td>
<td>72.9</td>
</tr>
<tr>
<td>3-Literature Text</td>
<td>69.6</td>
<td>67.2</td>
<td>85.7</td>
</tr>
<tr>
<td>4-Literature text</td>
<td>78.7</td>
<td>63.4</td>
<td>78.9</td>
</tr>
<tr>
<td>5-Agreement Copy</td>
<td>80.7</td>
<td>62.6</td>
<td>80.7</td>
</tr>
<tr>
<td>6-Agreement Copy</td>
<td>79.8</td>
<td>63.8</td>
<td>81.1</td>
</tr>
<tr>
<td>7-Astrology Readings</td>
<td>82.2</td>
<td>63.9</td>
<td>82.9</td>
</tr>
<tr>
<td>8-Astrology Readings</td>
<td>80.6</td>
<td>60.9</td>
<td>81.8</td>
</tr>
<tr>
<td>9-DIBCO 2017</td>
<td>70.6</td>
<td>57.6</td>
<td>74.54</td>
</tr>
<tr>
<td>10-DIBCO 2017</td>
<td>66</td>
<td>60.6</td>
<td>85.1</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>74.26</td>
<td>65.09</td>
<td>80.964</td>
</tr>
</tbody>
</table>
The above Table.3 depicts the average F-measures value of various existing algorithms like OTSU, NIBLACK and SAUVOLA as well as the proposed method using the various datasets. The table clearly indicates that the proposed method shows highest accuracy.

![Graph](image.png)

**Fig. 6 Graph depicting Average F-measure of existing vs proposed**

The above Fig.6 and table.2 indicate the graphs depicting average F-measure metric of 10 datasets for existing and proposed algorithm. The graphs interpreting the evaluation metric F-measure shows highest accuracy in the case of proposed algorithm for all 10 datasets. It clearly shows that the proposed algorithm is efficient in removing the noise in documents compared to the three well-known existing Binarization methods.

**Table 3 Performance evaluation metric for proposed method**

<table>
<thead>
<tr>
<th>INPUT IMAGE</th>
<th>MSE</th>
<th>SSIM</th>
<th>PSNR</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Palm leaf</td>
<td>912.32</td>
<td>0.95</td>
<td>52.85</td>
<td>92.8</td>
<td>92.8</td>
<td>92.8</td>
</tr>
<tr>
<td>2-Palm leaf</td>
<td>387.93</td>
<td>0.98</td>
<td>51.97</td>
<td>91</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>3-Literature</td>
<td>1300.29</td>
<td>0.92</td>
<td>48.98</td>
<td>86</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>4-Literature</td>
<td>2259.98</td>
<td>0.86</td>
<td>47.38</td>
<td>91.3</td>
<td>91.3</td>
<td>91.3</td>
</tr>
<tr>
<td>5-Agreement Copy</td>
<td>1056.69</td>
<td>0.92</td>
<td>40.64</td>
<td>92.7</td>
<td>92.7</td>
<td>92.6</td>
</tr>
<tr>
<td>6-Agreement Copy</td>
<td>489.67</td>
<td>0.96</td>
<td>53.89</td>
<td>94.5</td>
<td>94.5</td>
<td>94.6</td>
</tr>
<tr>
<td>7-Astrology Readings</td>
<td>2975.85</td>
<td>0.83</td>
<td>47.48</td>
<td>92.4</td>
<td>92.4</td>
<td>94</td>
</tr>
<tr>
<td>8-Astrology Readings</td>
<td>1826.78</td>
<td>0.92</td>
<td>51.35</td>
<td>92.8</td>
<td>92.8</td>
<td>92.8</td>
</tr>
<tr>
<td>9-DIBCO 2017</td>
<td>2720.57</td>
<td>0.88</td>
<td>49.93</td>
<td>89</td>
<td>89</td>
<td>90.2</td>
</tr>
<tr>
<td>10-DIBCO 2017</td>
<td>5817.46</td>
<td>0.85</td>
<td>50.5</td>
<td>91.3</td>
<td>91.3</td>
<td>91.3</td>
</tr>
</tbody>
</table>
The above Table 3 and Fig. 7 depicting the proposed method performance evaluation metric as well as graph which is obtained through various measures like MSE, SSIM, PSNR and F-Measure. The proposed algorithm is validated with the help of 10 datasets of different formats. The table 3 indicates that F-measure gives higher value compared to the other performance evaluation measures.

Conclusion

The proposed work devised a novel binarization method which uses a modified version of Wellner’s algorithm. The algorithm is tested on various datasets including palm leaves and document copies. The result of the experiment was compared with three existing and well-known binarization methods namely, Otsu, Niblack and Sauvola. It is observed that the proposed method was able to achieve an average accuracy of 92% which is observed to be greater than the existing methods. This would allow Optical Character recognition Softwares to easily enhance historical documents. The proposed algorithm’s efficiency drops while dealing with datasets with large bleed through, brittle nature as well as folding marks. Methods like deep learning and a combination of various filters can be used to further improve the algorithm and ancient documents become noise free.

References


