Camera Captured Kannada Hand-Written Document Binarization & Segmentation in Color Space

Vinod H C, S K Niranjan

Abstract*: Binarization of document images is a major phase in the handwritten text recognition process. Text recognition process gives best result and easy to archive recognition for printed documents, but more accurate and fast Binarization & segmentation methods are required to achieve high accuracy in handwritten character recognition. In this paper we presenting Document Binarization & Segmentation. In Document Binarization initially we convert RGB color space image to CMY color space further calculate the average of C, M & Y. Finally convert the grey scale image to binary by determining threshold value automatically. Segmentation is done by the projection profile method and paragraph skew correction recursively until height of the segmented line image is less than 7% of the input image, Connected Component Analysis is used to segment words. These segmented words can be feed to OCR for recognition; the proposed experimental results are encouraging.

***Index Terms*: CMY Color Space, Connected Component Analysis, Projection Profile, RGB Color Space.**

I. INTRODUCTION

In camera captured and scanned documents, where noise, contrast, and illumination vary, classifying pixels as foreground or background pixels is still a difficult and challenging problem. Binarization of document images is a major phase in the handwritten text recognition process, binarization is the separation of text from the background image. This separation enables algorithms for layout analysis or document transcription to focus only on the written text without the disturbing background noise. Segmentation process is the stage to separate words and characters from image, this process requires better binarization techniques, so binarization makes a very important role in recognition system.

Document Segmentation is process of rending document image into separate regions. Segmentation process usually consists of three process, they are, line segmentation, word segmentation and character segmentation. Automation of document segmentation will improve the time efficiency of recognition system; reduce the human interference and effort. The automation of handwritten document segmentation is a huge challenging task.Text recognition process gives best result and easy to archive recognition for printed documents, but more accurate and fast Binarization & segmentation methods are required to achieve high accuracy in handwritten character recognition when the number of documents increases with different style, font, size, age-group of writer etc., to archive very good efficiency system to recognise handwritten text requires huge number of document images with different possibilities of handwriting.

The text processing problem can be divided into:- Binarization: It refers to the determine the threshold value and suppress the non-text field in document. Text Detection: text detection is the process to determine the occurrence of text in a given document images. Text Localization: It is the process of identifying the location of text in the document image by creating or drawing bounding boxes around the detected text. Text Extraction: It extracts the text from Candidate text region. Text Enhancement: The text region usually has low-resolution and is prone to noise, hence it has to be enhanced. Text Recognition (OCR): The extracted text can be transformed into plain text using OCR.

Yokobayashi and Wakahara [1] presents two key ideas for binarization & recognition of characters in color space, the first key idea is to binarize characters from background of document by estimating threshold value by selecting one band from RGB color space that gives maximum difference thresholding values between the RGB class or axis. Second key idea for distortion- tolerant character recognition. the maximum correlation value between trained and testing images will recognise characters using the global affine transformation correlation.

Westphal et. al., [2] propose a recurrent neural network based algorithm using Grid Long Short-Term Memory cells for image binarization, as well as a pseudo F-Measure based weighted loss function. Vats et. al., [3] presents an automatic document image binarization technique to detect the text from degraded document images. They have removed background noise by using 2-band-pass filtering technique, and obtain the optimal result by automatic hyper parameter selection using Bayesian optimization approach.

Mudhsh et. al., [4] introduces a new hybrid meta-heuristic algorithm to determine the best threshold value for image documents binarization. This hybrid based on the Artificial Bee Colony (ABC) and the Firefly Algorithm (FA), and it is called ABCFA. It is making the two algorithms searching about the best threshold, which maximizes the Otsu function in simultaneous with half of the population for each algorithm. Therefore, this type of hybridization avoids the drawbacks of other hybrid methods.

Şekeroglu and Khashman

Revised Manuscript Received on June 11, 2019

Vinod H C, Department of Information Science and Engineering, SJB Institute of Technology, Bangalore, India.

S K Niranjan, Department of Computer Applications, JSS Science and Technology University, Mysuru, India.

[5] presented an optimal binarization method that efficiently utilized for different types of scanned documents. Initially, described the study of 13 different binarization techniques used for grey-level degraded historical document images, printed documents and handwritten documents. Finally, the 3-new parameters are presented to evaluate the performance and quality of binarized images.

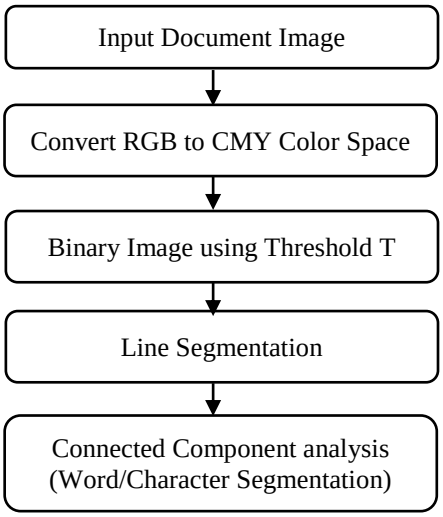

Figure 1: The proposed System Architecture.

Lin et. al., [6] proposed a hybrid technique to segment & classify the contents like graphics, text & space in document images. In proposed system initially divide the document into sub-images called blocks, later, extract the Grey Level Co-occurrence Matrix (GLCM) features for each blocks. By using these features, group the blocks into 3-clusters by applying K-Means clustering algorithm. These groups are classified to graphics, text & space by using heuristic rule classifier. Wick and Puppe [7] presented a high-performance fully convolutional neural network (FCN) for historical handwritten document segmentation for single page by processing in one step.

Barna et.al., [8] presents a document image segmentation system for multiple languages, in this proposed system, segments the heterogeneous printed document contents into homogeneous contents like texts, graphics, tables and halftones in three steps, they are: extraction of tables, halftones image extraction and table extraction. These modules use morphological operators to extract contents, these extracted homogenous segmented contents are indicated with color bounding boxes.

Zahour et.al., [9] described text line extraction method. Segmentation of printed document is done by typical horizontal projection profile technique and then regrouping using connected component analysis, these techniques are not best suited for handwritten unconstrained documents because of the base line skew, shift in base line, intensity of text and different inter line distance.

In this paper we proposed a system that contains binarization & Segmentation. Initially we convert the RGB color space to CMY color space and binarize the document image by determining threshold automatically on averaged CMY grey image, and secondly uses the projection profile with document skew correction recursively to segment lines and connected component analysis to segment the characters. The overall proposed system architecture is showed in Figure 1. The proposed system is divided into two sections as i.e., Document Binarization and Segmentation.

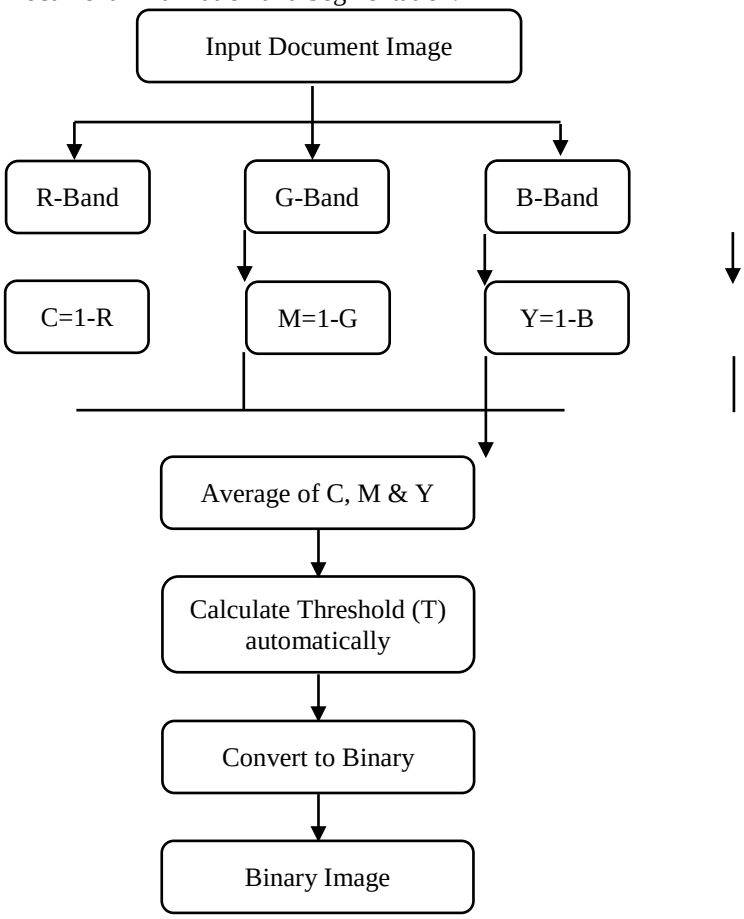

Figure 2: Proposed Binarization Technique.

II. PROPOSED SYSTEM

The proposed system is divided into two sections as i.e., Document Binarization and Segmentation. The input document image will be pre-processed by dividing input image into R, G & B bands. After pre-processing we convert the image to binary format by applying few more techniques that highlights the character present in the pre-processed image. This binary image will be feed to segmentation, in segmentation first segment the document image into lines and then connected component analysis has been applied to segment words. The overall architecture of proposed system is showed in Figure 1.

A. Document Binarization

The proposed binarization flow is shown in Figure 2. Initially the input document image is divided into R-band, G-band and G-band, these bands are shown in Figure 3(b-d) for input document showed in Figure 3(a). Further convert RGB color space image to cyan, magenta and yellow (CMY) color space image, they are computed as C = 1-R, f or M = 1-G and Y = 1-B, CMY results are showed in Figure 3(e-g). The average of C, M & Y images has been computed as below:

$$AVG_{CMY} = (C + M + Y)/3$$

The AVG_{CMY} is a grey scale image as showed in Figure 3(h), this grey scale image further converted to binary format [bin(i,j)] by using

automatic calculated Threshold (T) showed in Figure 3(i). The Threshold (T) has been determined as follows.

$$T = \frac{SC}{(m * n) + NHF}$$

Where, m & n is the number of rows and columns respectively, SC is the sum of co-efficient value in grey image and computed as below:

$$SC = \sum_{i=1}^{m} \sum_{j=1}^{n} AVG_{CMY}(i, j)$$

NHF is the number of high frequency coefficients that satisfy $AVG_{CMY}(i, j) > MC$. Where MC is the mean of co-efficient values and its computed as follows:

$$MC = \frac{1}{m * n} \sum_{i=1}^{m} \sum_{j=1}^{n} AVG_{CMY}(i, j)$$

B. Segmentation

Binary image contains the multiple lines of text, further to segment the lines in document. For segmentation we apply horizontal projection profile technique, horizontal projection gives the line by line segmented images. For all these images we will check the height of the image, if height of the segmented image is greater than 7% of input document image, it indicates that segmented image contains two or more lines in segmented image. This occurs due to the skew in handwritten document, i.e., paragraph skew in document. If the value is greater, then apply skew correction and apply horizontal projection profile to segment further. If the height is less than 7% then continue to next step. Finally apply connected component analysis to segment words. This segmented words or characters are feed to OCR for recognition purpose. The step by step procedure for the segmentation is showed in Algorithm-1.

Algorithm 1: Segmentation

//Input: Binary Document Image that contains 'n' lines of text.

//Output: n-segmented line images and segmented word images.

Step-1:Read Binary document image.

Step-2: Segment the document into multiple lines by computing horizontal projection profile.

Step-3: if (H_{SL}>7% of H_{ID}) *then*

goto *Step-4*

else

goto *Step-5*

Where H_{SL}-> Height of the Segmented Line Image.

H_{ID}-> Height of the Input Document Image.

Step-4:Apply Skew correction, rotate segmented line by desired angle, segment line by applying horizontal projection, goto step 3.

Step-5:Apply connected component analysis to segment words & draw bounding box for segmented words.

Step-6:Stop.

III. EXPERIMENTAL RESULTS

The proposed system has been implemented in MATLAB R2017a. We have collected our own dataset; our own dataset consists of 110-handwritten Kannada document images. These documents are collect by various people with different age group, written in single column, 10-20 lines and it contains only text objects. We have created document images by capturing handwritten document using camera phones, while capturing it created different noises, shadow, without shadow, indoor, outdoor. For this proposed system we have considered skew corrected document images. While capturing documents camera skew will be created, this created skew has been removed with multi-level Run-Length-Smoothed Image (RLSA) skew detection and correction technique [10]. The skew corrected document images are considered for the proposed system experiments.

Figure 6 shows the Binary images of our proposed system, the Feng thresholding method [12], the Sauvola method [13] and binarization using Haar wavelet decomposition method [11]. The experimental result of proposed binarization technique is encouraging. In Figure 6 we can notice that, the proposed system works better even the document contains shadow in the image when compared to Feng, Sauvola and our previous proposed Haar wavelet decomposition methods.

The proposed system best suited in camera captured document images with different color foreground and background as shown in Figure 6(e) & Figure 7(c). Figure 4 & 5 shows the line segmentation and word or character segmentation using connected component analysis. If we observe the result of proposed system & Haar wavelet decomposition technique [11], in case of Haar wavelet decomposition technique the edge thickness of each character is thicker due to the thickness, characters in word are connected. The edge thickness of characters in proposed system are thinner, due to this each characters in a word is separated and its easy to segment characters using connected component analysis as shown Figure 4 & 5. The proposed system is not suited if the document contains shadows as shown in Figure 7(f).

IV. CONCLUSION

Binarization of document images is a major phase in the handwritten text recognition process. Text recognition process gives best result and easy to archive recognition for printed documents, but more accurate and fast Binarization & segmentation methods are required to achieve high accuracy in handwritten character recognition. In this paper we presenting two modules, they are Document Binarization & Segmentation. In Document Binarization carried out by converting RGB color space image to CMY color space further calculate the average of C, M & Y. Finally convert the grey scale image to binary by determining threshold value automatically. Segmentation is done by the projection profile method and paragraph skew correction recursively until height of the segmented line image is less than 7% of the input image, Connected Component Analysis is used to segment words/Characters. These segmented words can be feed to OCR for recognition. We have shown the comparisons between the proposed system with Feng thresholding method [12], the Sauvola method [13] and Haar Wavelet Decomposition Techniques [11] in experimental result section by showing the resulted

images; the proposed system results are encouraging for poor document images with different color foreground text, but if document image contains shadow.

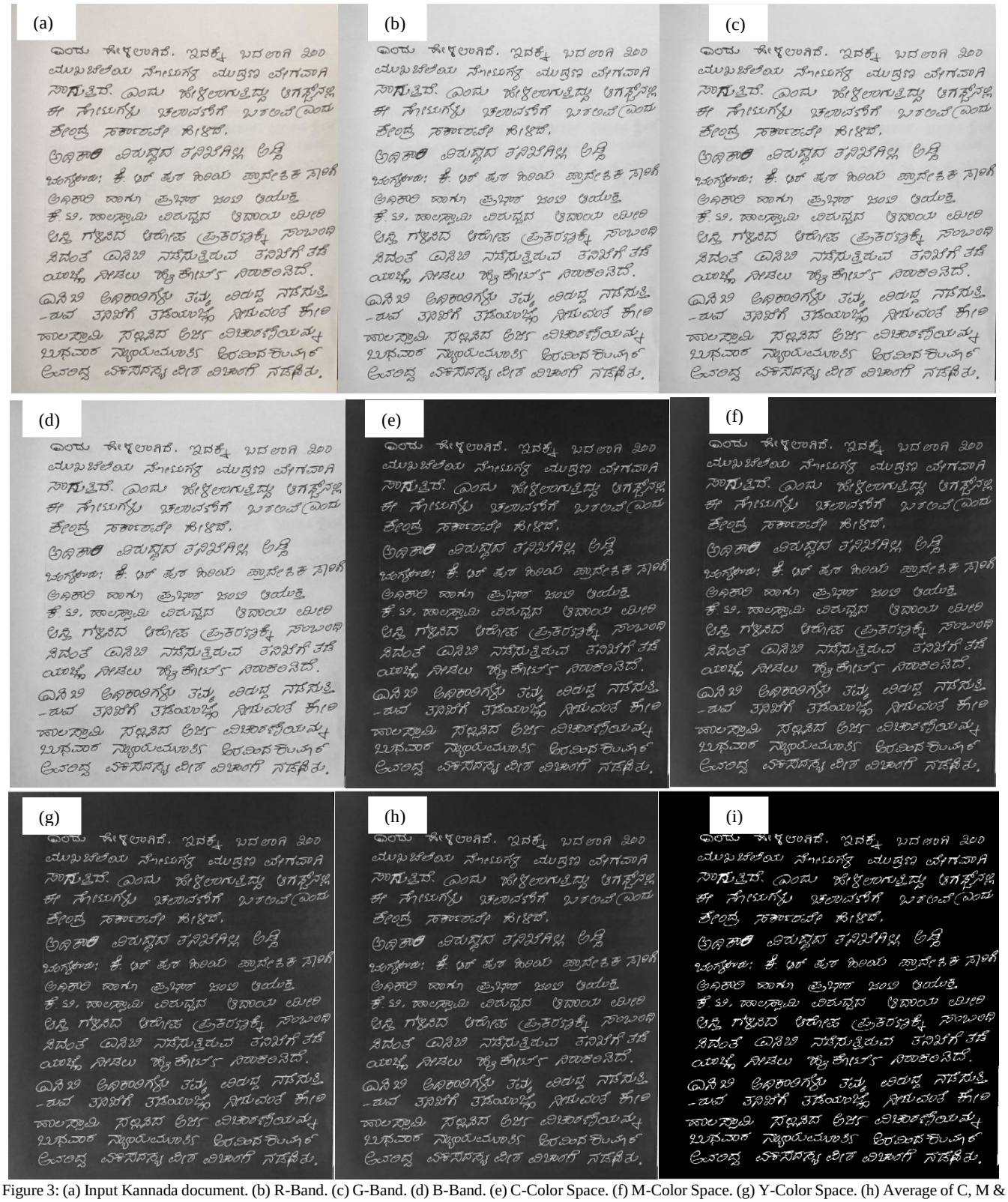

Figure 3: (a) Input Kannada document. (b) R-Band. (c) G-Band. (d) B-Band. (e) C-Color Space. (f) M-Color Space. (g) Y-Color Space. (h) Average of C, M & Y (AVG_{CMY}). (i) Binary Image.


(a)

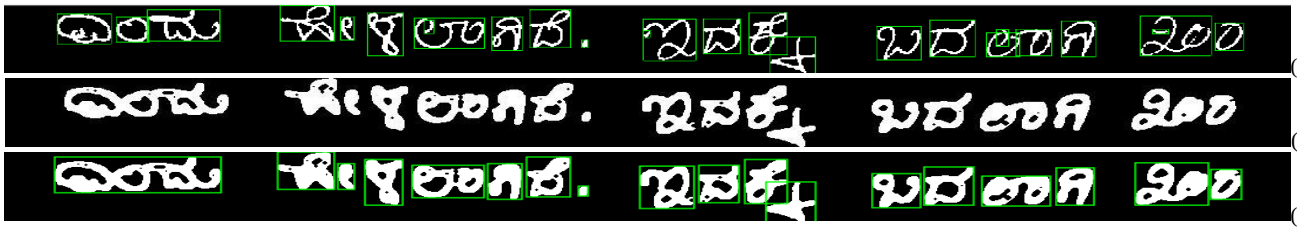

Figure 4: (a) 1st segmented line using proposed binarization and segmentation technique. (b) Connected component technique on proposed system. (c) 1st segmented line using Haar Wavelet Decomposition technique binarization and segmentation (Previous proposed technique). (d) Connected component technique on previous binarization technique.

Figure 5: (a) 8th segmented line using proposed binarization and segmentation technique. (b) Connected component technique on proposed system. (c) 8th segmented line using Haar Wavelet Decomposition technique binarization and segmentation (Previous proposed technique). (d) Connected component technique on previous binarization technique.

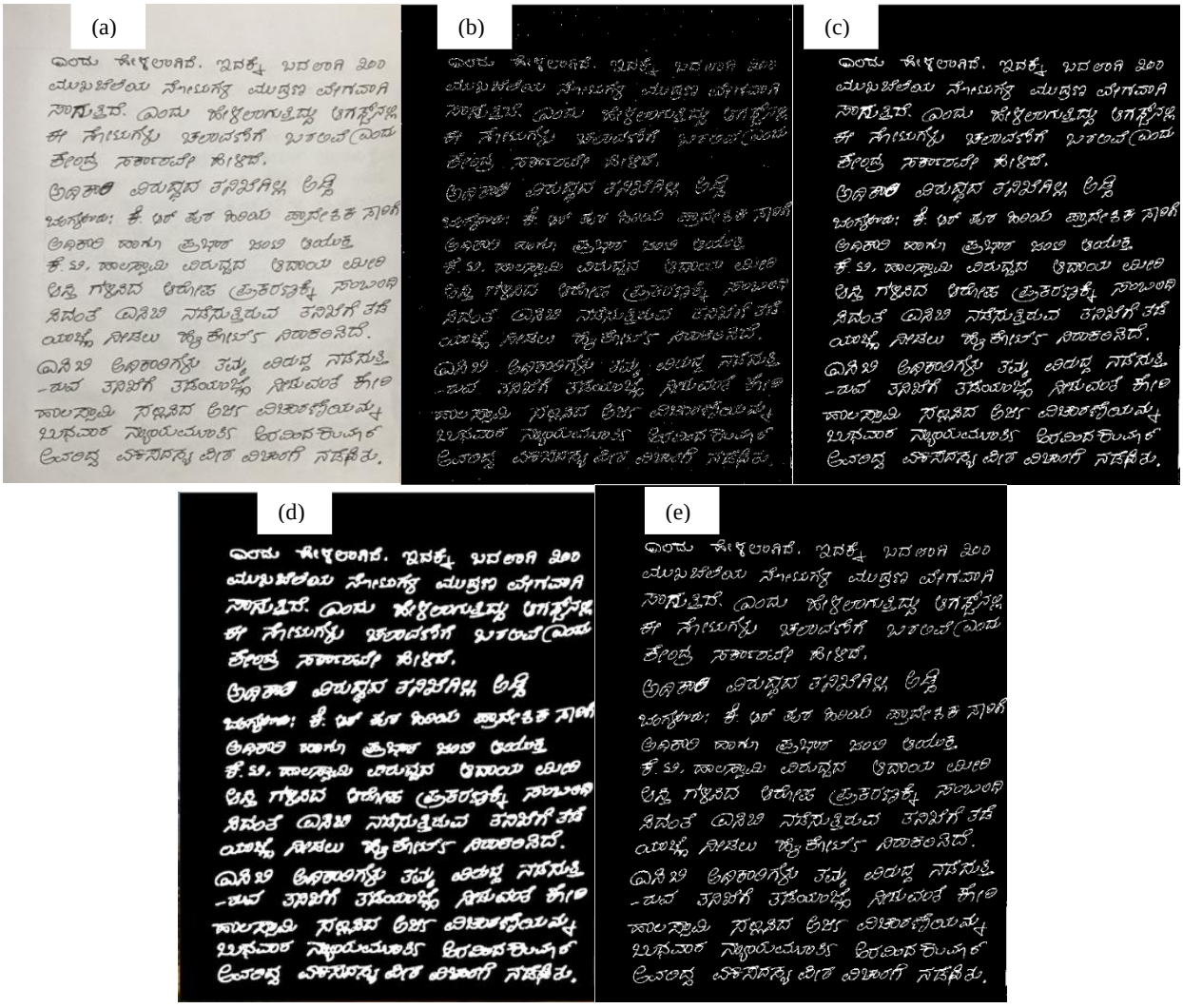

Figure 6: Binarization results: (a) Input Image. (b) Feng thresholding method. (c) Sauvola method. (d) Haar wavelet decomposition method. (e) Proposed System.

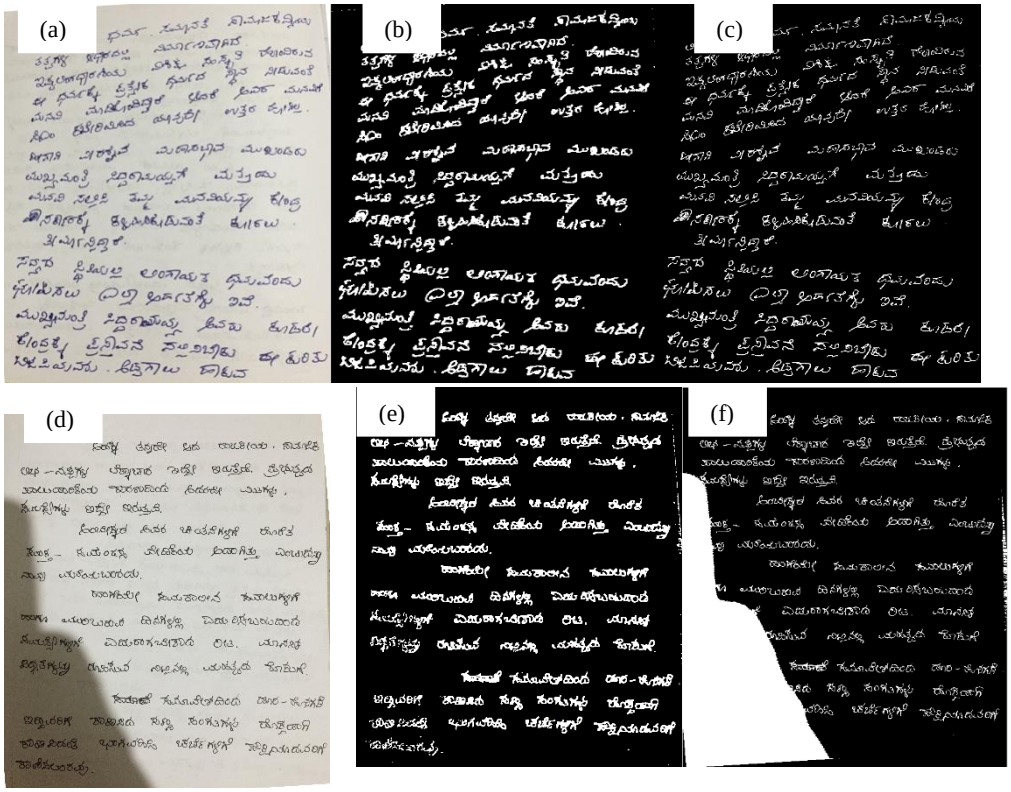

Figure 7: Comparing Proposed System with Haar Wavelet Decomposition [11]: (a) & (d) Input document image. (b)&(e) Haar wavelet decomposition technique [11]. (c)& (f) Proposed System.

REFERENCES

1. Minoru Yokobayashi and Toru Wakahara, "Binarization and Recognition of Degraded Characters Using a Maximum Separability Axis in Color Space and GAT Correlation", The 18th International Conference on Pattern Recognition (ICPR'06), IEEE-2016.
2. Florian Westphal, Niklas Lavesson, Hakan Grahn, "Document Image Binarization Using Recurrent Neural Networks", 13th IAPR International Workshop on Document Analysis Systems, IEEE-2018.
3. Ekta Vats, Anders Hast and Prashant Singh, "Automatic Document Image Binarization using Bayesian Optimization", arXiv:1709.01782v3 [cs.IR], 21 Oct 2017.
4. Mohammed Mudhsh, Shengwu Xiong, Mohamed Abd El Aziz, Aboul Ella Hassanien, Pengfei Duan, "Hybrid Swarm Optimization for Document Image Binarization Based on Otsu Function", CASA 2017.
5. Boran Şekeroğlu and Adnan Khashman , "Performance Evaluation of Binarization Methods for Document Images", ICAIP 2017, August 25–27, 2017, Bangkok, Thailand.
6. M-W Lin, J-R Tapamo, B Ndovie, "A Texture-based Method for Document Segmentation and Classification", Joint Special Issue — Advances in end-user data-mining techniques ARIMA/SACJ, No. 36., 2006.
7. Christoph Wick and Frank Puppe, "Fully Convolutional Neural Networks for Page Segmentation of Historical Document Images", arXiv:1711.07695v1, 21 November 2017.
8. Nasid Habib Barna, Tisa Islam Erana, Shabbir Ahmed and Hasnain Heickal, "Segmentation of Heterogeneous Documents into Homogeneous Components using Morphological Operations", 17th International Conference on Computer and Information Science (ICIS) June 6-8, 2018, Singapore, IEEE-2018.
9. A. Zahour, B. Taconet, P. Mercy, S. Ramdane, "Arabic Hand-Written Text-Line Extraction", IEEE-2001.
10. Vinod H.C., Niranjan S.K. (2018) Multi-level Skew Correction Approach for Hand Written Kannada Documents. In: Rocha Á., Guarda T. (eds) Proceedings of the International Conference on Information Technology & Systems (ICITS 2018). ICITS 2018. Advances in Intelligent Systems and Computing, vol 721. Springer, Cham.
11. Vinod H.C., Niranjan S.K. "Binarization & Segmentation of Kannada Handwritten Document Images", Second International Conference on Green Computing & Internet of Things (ICGCIoT 2018), at Global Academy of Technology, Bengaluru, India during 16-18 August 2018, IEEE-2018.
12. Meng-Ling Feng and Yap-Peng Tan, "Contrast adaptive binarization of low quality document images", IEICE Electron. Express, Vol. 1, No. 16, pp.501-506, (2004).
13. J. Sauvola and M. Pietikainen, "Adaptive document image binarization", Pattern Recognition Society. Published by Elsevier Science Ltd. Pattern Recognition 33, 2000.

AUTHORS PROFILE

Vinod H C. Is presently working as Assistant Professor in the Department of Information Science and Engineering at SJB Institute of Technology, Bangalore, Karnataka, India. He received his Bachelor of Engineering (B.E), Master of Technology (M.Tech) from **Visvesvaraya Technological University**. Currently he is perusing Ph.D in Visvesvaraya Technological University, Belagavi, Karnataka, India. His research interests include Image Processing and Pattern Recognition.

Dr. S K Niranjan Is working as professor in the department of Master of Computer Applications at JSS Science and Technology, Mysuru, Karnataka, India. He has more than 25 years of experience both in teaching and industry. He received his Bachelor of Business Management (BBM), Master of Computer Applications (MCA), Master of Technology in Software Engineering (M.Tech [SE]) and Ph.D in Computer Science from University of Mysore. Areas of interest are Image Processing, Pattern Recognition, Software Engineering, Business Analytics, Business Intelligence, Cloud Computing etc. Published more than 50 technical papers in reputed books, journals and conference proceedings. Invited as General Chair for more than six intentional conferences conducted in different parts of India. Delivered keynotes in many national and international conferences in India. Involved as Technical Program Committee member for many national and international conferences and also editor for many national and international journals and conference proceedings. Actively involved in many professional organizations like IEEE, ACM, CSI, ISTE etc. Executive Committee (EXECOM) member of IEEE Computer Society, India Council for year 2015 & 2016. At present actively involved in the IEEE Bangalore Section and member of Executive Committee (EXECOM) for the year 2017 and 2018.