International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

A new Connected Component Analysis based System for Text Segmentation in Degraded Historical Document Images

V. Sathya Narayanan, N. Kasthuri, T. Dharani, D. Deepa

Abstract: ***Historical documents contain valuable heritage information. These documents are preserved in the manuscript preservation center and archaeological departments. They are mostly degraded in nature and hence hard to read and understand the contents. So, there is a need for text segmentation and feature extraction to convert these manuscripts into machine editable format. In this work, we present an effective way to segment historical document images into characters. It is a challenging segmentation process due to complex background images. In this paper, horizontal histogram, vertical histogram and connected component analysis is used to segment text documents images. In this algorithm, the input image is converted to gray scale image, then gray image is converted into binary image [Otsu's method] and then all the objects containing fewer than desired pixels are removed. Line and word segmentation is implemented using horizontal and vertical histogram method respectively. Then the connected components are labeled and properties are measured for the image regions. Connected component analysis is used to segment the characters and the individual characters are extracted. The simulation result shows that the proposed segmentation method achieves an average accuracy of 93.37% for HDLAC 2011 DATASET. Moreover this method is more efficient and more suitable for real time tasks.***

Keywords : Otsu method, horizontal histogram method, vertical histogram method, connected component analysis, Bounding box segmentation, HDLAC Dataset.

I. INTRODUCTION

To preserve the cultural heritage present in the historical documents, libraries all around the world digitize massive number of manuscripts. These documents cannot be accessed until it is converted into digital format, so these manuscripts are scanned and digitized to preserve these documents from degradation and parchment.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Mr.V.Sathya Narayanan*, Assistant Professor, Electronics and Communication Engineering Department, Kongu Engineering College, Perundurai, Erode, India. Email: sathya198823@gmail.com
Dr.N.Kasthuri, Professor, Electronics and Communication Engineering Department, Kongu Engineering College, Perundurai, Erode, India. Email: kasthuri@kongu.ac.in
T.Dharani, Student, Electronics and Communication Engineering Department, Kongu Engineering College, Perundurai, Erode, India. Email: dharanithiruvelmani@gmail.com
D.Deepa, Student, Electronics and Communication Engineering Department, Kongu Engineering College, Perundurai, Erode, India. Email: deepadhanasekaran1998@gmail.com

Even after digitization, historical documents are difficult to access due to its degradation factors. The historical manuscripts need to be changed into machine editable one. Manual conversion is very difficult while considering huge amount of manuscripts. In modern printers, automatic, printed document images can be converted into machine understandable format. However, those manuscripts are not possible to extract because of low quality, due to age or faint typing and various disturbing elements such as holes, spots, various forms of writing, various font size and font style and overlapped characters. The four processes involved in digitization process are binarization, segmentation, feature extraction and classification. This paper primary focus is on front end (preprocessing and segmentation) of recognition process. The proposed method is evaluated based on standard performance metric.

A. Input Scanned Image

Firstly, input image is optically scanned. An image, which has been scanned could be a document of different dimensions, different style and with different font size. This scanned input picture is fed to pre-processing stage to process an image that was scanned.

B. Pre Processing

Pre-processing includes various operations over the input image, so that scanned Image becomes suitable for applying to further sub stages. Basically, the goal of pre-processing is to improve the quality of scanned input image. Noise removal, certain mathematical operations may be processed in this Pre-processing stage. It undergoes binarization, filtering, noise removal and thinning. These operations are performed over scanned input image to retrieve data.

C. Segmentation

Image segmentation is the process of separating a digital image into multiple segments or sets of pixels or super-pixels. The objective of segmentation is to partition or change the representation of an image into data that can be easy to analyze. Image segmentation is useful to locate image objects and its boundary.

II. LITERATURE REVIEW

Aladhahalli Shivegowda Kavitha, Palaiahnakote Shivakumara, Govindaraj Hemantha Kumar, Tong Lu introduced a new watershed model based system for character segmentation in degraded text lines(2017).

In this paper, noise is filtered out using sobel and laplacian values.

Watershed model for studying non-linear spacing between characters is used in this paper. They concentrated on line and word segmentation and not character segmentation.[1]

ZechengXie, Yaoxiong Huang, LianwenJin, Yuliang Liu, Yuanzhi Zhu, Liangcai Gao and Xiaode Zhang"Weakly supervised precise segmentation for historical document images" (2019). Here the challenging problem is historical document image segmentation from a Bayesian decision theory perspective. To achieve high-precision segmentation, three novel algorithms are used. That is boundary box segmentation to get a coarse segmentation result, recognition-guided boundary box segmentation to get fine segmentation result and recognition guided attention boundary box segmentation(Combination of both) to get fine segmentation result at required time. Furthermore, another proposal of an incremental weakly supervised learning strategy with judgment gate mechanism training had made. The result shows an improved performance of the character recognition as well as the final segmentation result. Their novel recognition guided and attention based boundary box segmentation method gives better performance metric result for high Intersection over Union (IoU) threshold values, which is critical for degraded historical images. The performance are low. It takes more computational time [2].

Di Lu, Xin Huang Li, Xue Sui introduced Binarization of degraded document images based on contrast enhancement (2018). In this paper, the difference of gray contrast between regions which divides significant areas and comparatively significant areas were used. For complicated regions, weak contrast enhancement is used to identify the difference between foreground and background regions of the image. Meanwhile, weak contrast enhancement is able to reduce noise in the results. For comparatively significant areas, strong contrast enhancement is used to adjust gray values so that the method can easily distinguish between foreground and background, and clear characters can be separated. This method have clear and complete characters as well as mostly noise-free backgrounds. This algorithm achieved the highest F-measure and PSNR for DIPCO 2016 dataset. The value of F-measure is 80.98 which are low when compared to some other existing methods [3]. A.S. Kavitha, P. Shivakumara, Kumar, Tong Lu "Text segmentation in degraded historical document images"(2016). In this paper, a new technique is used by combining sobel and laplacian operations. This method enchance low contrast pixels A grouping process, which involves the nearest neighbour criterion for merging text components have proposed. In this paper, they generate skeleton like structure for text pixels to reduce segmentation burdens and helps to study pixel structures in effective manner. Followed by this, their proposed clustering process helps them to segregate text and non text regions. This paper ends with text line extraction and does not deals with character extraction and recognition. The performance metrics Precision and Recall are discussed. Accuracy was not discussed [4]. Yuan wang Wei, Shifu Zhou, Zhijiang Zhang, Dan Zeng, Wei Shen, Mei Fang"Text detection in scene images based on exhaustive segmentation "(2017). In this paper exhaustic segmentation method is used. Exhaustic segmentation and two layer filtering method is used to segregate character candidate region and non character candidate region. Character candidate region is generated using parallel structure and then text line grouping stage, the edges of the fully connected graph of the remaining character candidate regions are cut by SVM classifier. This method is robust to the blurred image, low resolution and small sized text [5].

Hyung Jeong Yang, Quang Nhat Vo, Soo Hyung Kim, Gueesang Lee "Binarization of degraded document images based on hierarchical deep supervised network" (2018). In this paper, they talk about problems faced over binarizing degraded and aged document. So, they used full utilization of input-domain knowledge considerably limits distinguishing of background noises from the foreground. In this paper, a novel supervised-binarization method is proposed, in which hierarchical deep supervised network (DSN) architecture is learned for the prediction of the text pixels at different feature levels. With the help of high level features, the proposed DSN architecture can segregate foreground text pixels and background and severe degradations occur in historical images can be retrieved. At the mean time, foreground maps which are guessed using low level features provide superior vision at boundary area. In comparison with traditional segmentation methods, their network provides cleaner background and better-preserved strokes. The number of convolution layer is more which results in time consumption [6]. N. Nikolaou, M. Makridis, N. Papamarkos, B. Gatos, N. Stamatopoulos introduced segmentation of historical machine printed documents using adaptive run length smoothing and skeleton segmentation path(2010). This paper focus on digitizing historical machine printed document pages. Often, this kind of historical printed pages will be degraded by local skew, low quality and ink diffusion, and exhibit complex and dense layout. To meet these problems, we introduce the following innovative steps: (i) use of a novel Adaptive Run Length Smoothing Algorithm (ARLSA) in the way to meet the problem of complex and dark document layout, (ii) detection of noises, punctuation marks and certain small images that are usual in historical machine printed images, (iii) detection of possible obstacles formed from background areas in the way to segregate neighboring text columns or text lines, and (iv) use of skeleton segmentation paths in way to separate available connected characters. Similar experiments using lot of historical machine printed document pages provides good segmentation result.[7]

III. EXISTING SYSTEM

A. Bounding Box Segmentation

In existing method, boundary box segmentation is used to detect text in historical images. Bounding Box for connected components are the properties of the labeled connected component regions. A bounding box of a labeled region is a rectangle that just encloses the region completely. When a specific bounding box is determined for a connected region, the co-ordinates of the corners of the bounding box and its width and height are available. A bounding box completely specifies the boundaries of the corresponding connected component. Filled bounding boxes completely cover the corresponding connected components.

B. Limitations

Segmentation result is coarse. Segment the characters without consulting with the character recognizer.

C. Advantages

Provide good segmentation result for individual characters. Computational time is less.

IV. PROPOSED SYSTEM

The process flow of our proposed system is given in above Fig. 1.

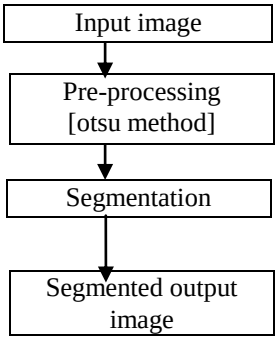

Fig. 1 Proposed system design.

A. Input Image

Firstly, image of input data is historical document image. The historical document image can be of any document but it should be in English and letters may be uppercase or lowercase letters or digits 0-9 or punctuation marks used in English. And these characters may be of any font style and font size but the one and only condition is that the characters in the image should be visible and clear in the normal view and not when it is zoomed. This historical input image is fed to pre-processing section so as to process over that historical image.

B. Pre-Processing

Preprocessing is nothing but, binarizing the historical document image and removing the noise in the historical image. So that, the image will provides exact result, when the image is segmented. Operations involved in pre-processing are given in Fig. 2.

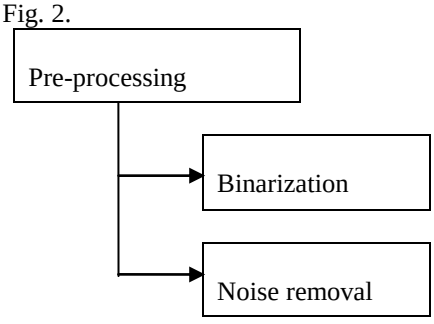

Fig. 2 Pre-processing module which involves binarization and noise removal.

- *Binarization:*

  Processes involved in pre-processing are given in Fig. 3.

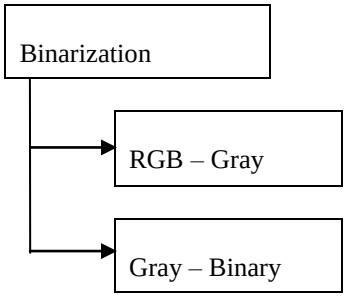

Fig. 3 Binarization module

- *RGB to Gray Conversion:*

  At first, a color image or RGB format image is 24 bit and 3 dimensional. It should to be converted into binary image which is 2 bit (0's and 1's). This direct conversion is not possible. To convert RGB image into binary image, the RGB image need to be converted into grayscale image which is 8-bit. A grayscale image is composed of different shades of gray color that is from 0-255 pixel value range. RGB image is converted into grayscale image by conserving the luminance or brightness of color image. Grayscale image is a combination of 30 percent red, 11 percent blue and 60 percent green. This rgb2gray function in MATLAB removes the hue and saturation information from the original image and then it will retain luminance or brightness of the image from 8 bit gray scale format. And this, 8bit scale format is gray scale image and it contains different shades of black and white color. This gray scale image is feed to gray to binary conversion to get a binary image.
  Algorithm:
  i). Convert the original image from RGB to gray scale using rgb2gray function in MATLAB.
  ii). Gray value can be estimated as in (1),

  $$Gray=0.299*R+0.587*G+0.114*B \quad (1)$$

- *Noise removal:*

  Binarized image is then feed into bwareaopen function in MATLAB to remove the noise in the image. Binary area open function(bwareaopen) determines the connected components and then calculate the area of connected components and removes all the connected components less than certain pixels in a binary image and produces another binary image. Bwareaopen function in MATLAB acts as a simple noise filter and it removes the noise in the image.
  Algorithm:
  i. Determine the connected components.
  ii. Compute the area of each component.
  iii. Remove all small objects.
  iv. Small objects mean pixel length less than 30 using bwareaopen function.

v. This acts as a simple noise filter.

C. Text Segmentation

Text segmentation is nothing but subdividing the text into meaningful lines, words and characters. The same was implemented in proposed system line segmentation which extracts the line from text or an image or a paragraph. This extracted line is fed for word segmentation and character segmentation. Word segmentation extracts each words from the line based on spacing. Character segmentation extracts each and every character from the line. These segmented individual characters are then fed for feature extraction. Individual characters are extracted from the text based on the segmentations given in Fig. 4.

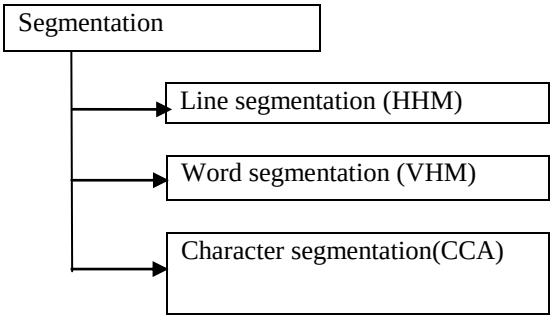

Fig. 4 Segmentation module which involves various segmentation process

- *Line Segmentation(HHM)*

For line segmentation, horizontal histogram segmentation method is used. To segment a line, horizontal histogram is generated in order to segment the text line from the input noise removed binary image. The space between the lines is considered to separate the lines. The space is determined by 1's and 0's. 1's is the text line which is white color in binary image and 0's is space which is black color in binary image. Find all the ones regions in which value is non zero. These rows contain text. Iterate the same across the rows and point flag one for non zero that is text region and flag zero for all zeroes that is non text regions. Then histogram is plotted for flag 1 and flag zero values and outputs from histograms are chosen as threshold value. And based on these threshold values, rows from the image are segmented which is lines. These lines are stored and feed for word segmentation and character segmentation.

Algorithm:

i. Calculate the sum of intensities across all columns for each row of pixels in the image.

ii. Find all the row regions in which this value is non zero. These are the rows in which text is present. Flag these rows with value 1 and those without text with value 0.

iii. Iterate across all rows. Points where flag value goes from 0 to 1 represents the starting of a line. Points where flag value goes from 1 to 0 represents the ending of a line.

iv. Plot the horizontal histogram.

v. Separate rows from text based on histogram value.

vi. Store the rows in which text is present. These will be used in word segmentation.

- *Word Segmentation(VVM):*

For word segmentation, vertical histogram segmentation method is used. To segment a line into word, vertical histogram is generated in order to segment the word from the text line. The spaces between the words are considered to separate the lines. The space is determined by 1's and 0's. 1's are words which are represented by white color in binary image and 0's are spaces which are represented by black color in binary image. The histogram is plotted for ones region which is word and those histograms are chosen as threshold value. Find all the ones regions in which value is non zero. These columns contain text. Iterate the same across the columns and point flag one for non zero that is text region and flag zero for all zeroes that is non text regions. Then histogram is plotted for flag 1 and flag zero values and outputs from histograms are chosen as threshold value. And based on these threshold value, columns from the line image are segmented which are words. These words are stored and feed for character segmentation.

Algorithm:

i. Calculate the sum of intensities across all rows for each column of pixels in the line.

ii. Find all the column regions in which this value is zero and the region is considerably large (greater than 85 columns). The minimum length of the region is set to ensure that gaps between characters do not lead to segmentation. These are the columns in which words are not present. Flag these columns with value 0 and those with words with value 1.

iii. Iterate across all columns. Points where flag value goes from 0 to 1 represents the starting of a word. Points where flag value goes from 1 to 0 represents the ending of a word.

iv. Plot the vertical histogram.

v. Separate columns from line text based on histogram value.

vi. Store the columns in which text is present (words). These will be the input to character segmentation.

- *Character Segmentation(CCA):*

Character segmentation is based on connected component analysis method. In connected component analysis method, connected components are labeled uniquely based on a given heuristic. Connected components are components which are not separated by a boundary. These connected component partition an image into segments.

Algorithm:

i. Identify all the connected components in the input word.

ii. Label the connected components.

iii. Iterate through the list of connected components and save each of them into a separate image.

iii. These images are the characters that will be passed to the character recognition algorithm.

V. RESULT AND ANALYSIS

A. Simulation analysis

Various dataset images are taken and analysed by using Connected Component Analysis Method. This proposed system improved the accuracy of segmenting historical document images taken from dataset HDLAC 2011 and it make use of MATLAB software for segmenting it.

B. Dataset Details

Datasets are taken from PRIMA research lab. The PRIMA (Pattern Recognition and Image Analysis Research Lab) Dataset contains various images like machine printed image, handwritten image, color image, grayscale and real time images.

Dataset Description:

- HDLAC 2011 Dataset.
- The dataset contains machine printed document images of various types like novels, newspapers, books and old magazines.
- 17 different languages and 11 scripts from the 17th to the early 20th century.

C. Performance metrics – accuracy:

Ratio of total detected to the total available in the text.

D. HDLAC 2011 datasets

Some images from HDLAC [Historical Document Layout Analysis Competition] 2011 datasets are given in Fig. 5.

Fig. 5 HDLAC 2011 dataset images.

E. Simulation result

Simulation result for an image from HDLAC 2011 dataset is shown in Fig. 6. In Fig. 6, first image is input image, followed by grayscale image and final image is binary image.

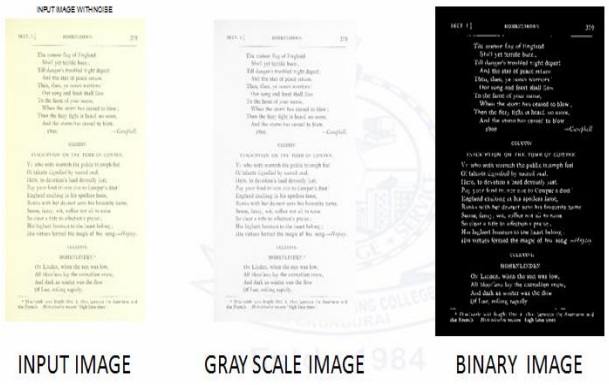

Fig. 6 Pre-processed output

Segregation of text image into separate lines is called line segmentation. Line segmentation was done using horizontal histogram segmentation method. Fig. 7 shows line segmented output.

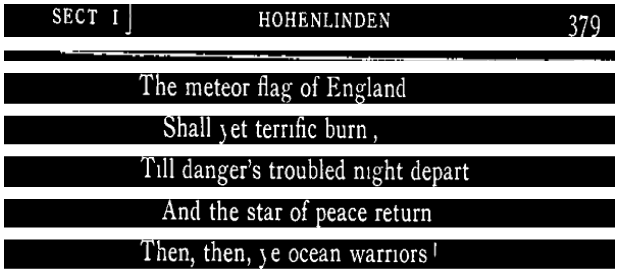

Fig. 7 Line segmented output.

Segregation of text line into separate words is called word segmentation. Word segmentation was done using vertical histogram segmentation method. Fig. 8 is word segmented output.

Fig. 8 Word segmented output

Segregation of word into separate characters is called character segmentation. Character segmentation was done using connected component analysis segmentation method. Fig. 9 is character segmented output.

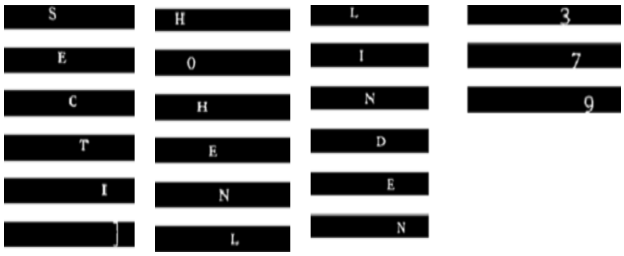

Fig. 9 Character segmented output

F. Comparison table

Table-I shows the comparison of accuracy between BBS-Boundary Box Segmentation and CCA-Connected component analysis segmentation.

Image	Existing Method (BBS) Accuracy %	Proposed Method (CCA) Accuracy%
1	98	99
2	76	88.32
3	96.04	96.47
4	96	96.91
5	96.44	97.08
6	96	96.45
7	96	96.74
8	91	92.43
9	97	98.57
10	85	85.97
11	94	97.59

12	93	95.21
13	91	91.85
14	98	98.11
15	99	99.8
16	95.1	95.8
17	0	90.29
18	94	98.18
19	93	94.42
20	91	73.34
21	93	94.41
22	95	99.4
23	97	96.41
24	89	65.98
25	97	95.52
26	0	75.08
27	98	98.84
28	84	84.84
29	90	91.57
30	96	97.15
31	95	99.4
32	98	99.07
33	95	95.9
34	99.5	99.53
35	96	96.9
Average	**88.308**	**93.34**

Table-I : Accuracy comparison table

G. Inference

The proposed connected component analysis method has been tested on Dataset - HDLA 2011 Dataset. The simulation result out performs most segmentation methods in terms of Accuracy. The Table.1 shows comparison results of the connected component analysis with bounding box segmentation method. The Fig. 10 shows accuracy comparison between existing system – bounding box segmentation and connected component analysis segmentation. Accuracy of proposed method is greater than the existing method and this accuracy is desirable.

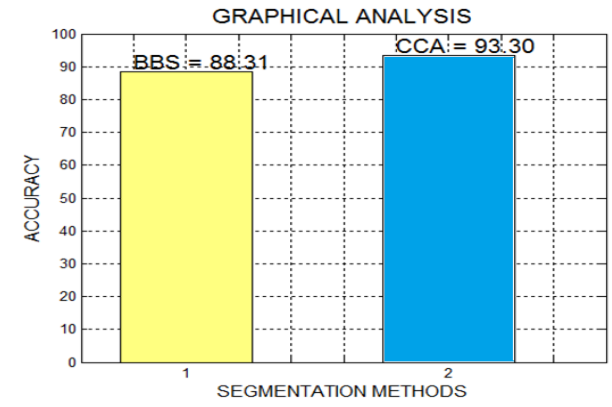

Fig. 10 Graphical Analysis

VI. Conclusion And Future Scope

A. Conclusion

Challenging problem in degraded historical document images were solved using connected component analysis segmentation method. To obtain high-precision segmentation line segmentation using horizontal histogram method, word segmentation using vertical histogram method and character segmentation using connected component analysis were done progressively. Experiments show that the proposed had improved the performance of the segmentation result. Compared with traditional methods, the proposed connected component analysis exhibits superior performance and accuracy for historical image from Pattern Recognition And Image Analysis (PRIMA) research lab.

B. Future work

The proposed method connected component analysis has done certain segmentation process over HDLAC DATASET 2011 images and obtained good accuracy result. For segmentation, the proposed method has performed processing over blur images, multiple touching character component images and has obtained a good result. But the exact digitization of the image was not obtained and it can be obtained by feature extraction and classification which will be explored in future.

REFERENCES

1. Aladhahalli Shivegowda Kavitha, Palaiahnakote Shivakumara, Govindaraj Hemantha Kumar, Tong Lu, A new watershed model based system for character segmentation in degraded text lines, 2017.
2. ZechengXie, Yaoxiong Huang, LianwenJin, Yuliang Liu, Yuanzhi Zhu, Liangcai Gao and Xiaode Zhang, Weakly supervised precise segmentation for historical document images, 2019.
3. Di Lu,Xin Huang Li, XueSu, Binarization of degraded document images based on contrast enhancement, 2018.
4. A.S. Kavitha, P. Shivakumara, Kumar, Tong Lu , Text segmentation in degraded historical document images, 2016.
5. Yuanwang Wei, Zhijiang Zhang, Wei Shen, Dan Zeng, Mei Fang, Shifu Zhou, Text detection in scene images based on exhaustive segmentation , 2017.
6. Quang Nhat Vo, Soo Hyung Kim, Hyung Jeong Yang, Gueesang Lee, Binarization of degraded document images based on hierarchical deep supervised network, 2018.
7. Nikolaou N, Makridis M, Gatos B, Stamatopoulos N, papamarkos N, Segmentation of historical machine printed documents using adaptive run length smoothing and skeleton segmentation path, 2010.
8. Yiping Chena, Liansheng Wangi, Recognition based segmentation algorithm for online Arabic handwriting,2017
9. A.S.Kavitha, P.Shivakumar, G.H.Kumar, Tong Lu, Text Segmentation in degraded Historical document images,2016
10. Wei Xiong, Jingjing Xu, ZijieXiong, Juan Wang, Min Liu, A new scheme for text line and character segmentation from gray scale images of palm leaf manuscript, Optik,vol. ED-164,pp. 218-223, 2016.
11. Valdon,Boban, Efficient character segmentation approach for machine printed documents,2017
12. Bag S. Harit G. Bhowmick P., 'Recognition of Bangla compound characters using structural decomposition'- Elsevier Vol 47 No. 3, 2014.
13. Bannigidad P. Gudada C. , 'Restoration of degraded Kannada handwritten paper inscriptions (Hastaprati) using image enhancement techniques'- International Conference on Computer Communication and Informatics ,2017.
14. Belagali N. Angadi S. A., 'OCR for handwritten Kannada language script'- Int. J. Recent Trends Eng. Res. (IJRTER) Vol 2 No. 8, 2017.

15. Kale Karbhari V., 'Zernike moment feature extraction for handwritten devanagari (Marathi) compound character recognition' - IJARAI Vol 3, 2014.
16. Karthik S. Srikanta Murthy K., 'Segmentation and recognition of handwritten kannada text using relevance feedback and histogram of oriented gradients-a novel approach' - Int. J. Adv. Comput. Sci. Appl. (IJACSA) Vol 7 No. 1, 2016.
17. Mohana H.S. Navya K. Srikanth P.C. Shivakumar G., 'Stone in scripted Kannada Character matching Using SIFT' - Proceedings of the IRF International Conference pp.126–131.
18. Soumya A. Hemantha Kumar G. (2015) 'Recognition of historical records using Gabor and zonal features' - SIPIJ Vol-6, 2014.

AUTHORS PROFILE

Mr.V. Sathya Narayanan is presently working as Assistant Professor a, Department of Electronics and Communication Engineering, Kongu Engineering College, Erode, India (affiliated to Anna University). Presented 5 Papers in Journals and organized several events.

Dr.N. Kasthuri is presently working as Professor a, Department of Electronics and Communication Engineering, Kongu Engineering College, Erode, India (affiliated to Anna University). Published more than 50 papers in Journals and presented more than 50 papers in conferences. And organized more than 30 events (Seminars, Workshops, Conferences and Hands-on training, etc;).

T. Dharani is currently pursuing bachelor of engineering degree in a Department of Electronics and Communication Engineering, Kongu Engineering College, Erode, India (affiliated to Anna University).

D. Deepa is currently pursuing bachelor of engineering degree in a Department of Electronics and Communication Engineering, Kongu Engineering College, Erode, India (affiliated to Anna University).