Text and Non-Text Region Identification Using Texture and Connected Components

Ankit Vidyarthi[1], Namita Mittal[2], Ankita Kansal[3]
[1,2]Malaviya National Institute of Technology Jaipur, [3]BITS Plani
ankit.09303022@gmail.com, nmittal@mnit.ac.in, ankitakansal.com@gmail.com

ABSTRACT

Finding text area from document image i.e. an image which has text embedded with graphic is a challenging task. In the past few years, people are working on document images to extract the text from complex colored background images but results in the extraction of text with the loss of the existing graphics from the original image. However, it is a challenging problem to detect text and non-text region, because extraction of a text region from a document image has lower pixel intensity over graphics pixel intensity. In this paper, a new texture based method is proposed for extraction of Text and Non-Text area without losing the graphics from the document image using binarization and nearly connected component.

KEYWORDS - *Image Variance, Binarization, Morphological closing, Nearly Connected Component, Object Pixel.*

I. INTRODUCTION

Documents, in which text and non-text are intermixed to each other, are very commonly used in real life. More complexity to these documents comes when their backgrounds are complex colored, for example, in magazines, comics, advertisements and web pages. Image in the computer memory is stored in the pixel's form. So whenever, we are working with the document image all text and non-text region that are within the image, are considered as image. When we want to gather the detail about the text and non-text region in the document image, we have to deal with the pixel level details of the image to separating text and graphic in document image. Detection of text from these documents is a challenging problem. There are number of application areas where a text information extraction is being used, including document analysis [10], vehicle license plate extraction, technical paper analysis, Comics [9], assorted images [15] and so on. Text and non-text detection from document image is based on the spatial feature of the text and non-text region e.g. Spatial Cohesion, Frequency Component, Region and Orientation of text Region, entropy based, local variance, edge based, wavelet properties etc.

Spatial domain and frequency domain approaches are two ways of getting detail. Spatial domain approaches is related to directly working on the pixel values. While for frequency domain approaches, frequency domain representation of image is found and a feature vector is derived from this representation. There have being a lot of work done in past related to text detection in the document image but most of the work is related to only separation the text and non-text region only. There is less work related to graphic region that contain text in it. The objective of the paper is to deals with problems related to text and non-text region identification within a document image without losing the graphics. In this paper, effectiveness of nearly connected component method is explored for text extraction from the document image. Experimental results show that proposed approach produces effective results.

This paper is organized as follows. Related work and challenges are discussed in Section 2. In Section 3, proposed approach is described in detail. Further, experimental results are discussed in Section
4. Finally, section 5 presents the conclusion.

II. CHALLENGES AND RELATED WORK

For detection of text/non-text in document image, comics, natural scenery, historical documents, there are approaches based on spatial models.

For extraction of the text from natural scenery images author proposed an adaptive binarization and perceptual color based clustering method [11]. The verification of the true text was based on local and global relationship of single and multiple components. While in [18], color based segmentation followed by an extraction of several connected component based features with maximum likelihood estimation is being proposed for text extraction. In [16] extraction of text using ISEF edge detection

978-1-4799-3140-8/14/$31.00 ©2014 IEEE

technique has been proposed. For the text region identification using Morphological operations was given in [17] for natural scenery images.

For the extraction of the text region from historical documents author proposed a probabilistic model based on Markovian-Bayesian clustering method [12]. It's a three step process which includes clustering of the regions followed by the region merging results in a complete shape of the text. Finally binarization results in a complete text extraction from the image. Also pattern recognition based approach described in [14] uses pattern extraction from region-of- interest (ROI) with maximum likelihood classifier for text extraction.

Other application areas like text extraction from comics [9] uses region based text extraction methodology which includes connected components

and edge based detection approach for balloon detection, Text blob extraction, text recognition and text extraction. While OCR based approach [13] for searching and retrieval of biomedical images uses text extraction mechanism. The key idea is that the medical images has an text embedded with the image which was extracted and on the basis of such text the corresponding images are being searched and retrieved.

In the literature all of the approaches described extract the text from image with the loss of graphics. The proposed approach overcomes this drawback.

III. APPROACH USED

Proposed method for text and non-text identification is consisting of several digital image processing steps. These steps are related to processing of the pixel level detail. The objective is the find the relationship among the pixels to get the text and non-text regions' detail. Firstly, properties of the text/non-text region are collected, and then these properties are work as base of text/non-text identification. This approach takes a color/gray image as the input. For the color image color to gray conversion is required. A gray level histogram is plated for this gray image that shows the number of pixel at a particular gray level. Work flow of the proposed approach is described in Figure 2.

This histogram is used in Otsu threshold method [2] to calculate the global threshold value for the given image. The image is binarized by using this threshold value. The binarized image is used to find the connected components [4]. These components are assigned unique label number.

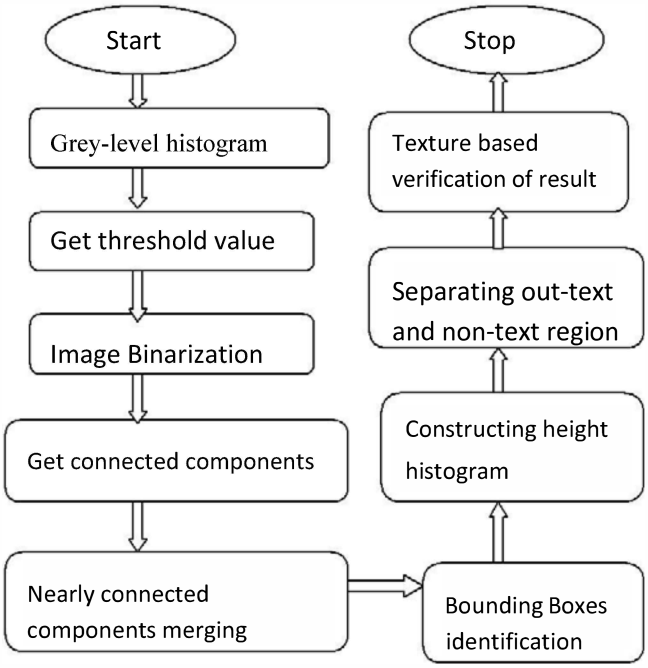
Start
Grey-level histogram
Get threshold value
Image Binarization
Get connected components
Nearly connected components merging
Bounding Boxes identification
Constructing height histogram
Separating out-text and non-text region
Texture based verification of result
Stop

Fig 2. Work flow for the approach

The components that are nearly connected to each other are merged into one component. As a result, the new label number is assigned to the resultant components.

These components are used to find the bounding rectangle. The heights of the bounding rectangles are used to construct the height histogram. Height histogram is used in the heuristic approach to find the height of the acceptable bounding boxes. The acceptable height is used for filtering the components in the text and non-text.

The filter components are verified on the bases of the texture property that is being used for the text and non-text identification. Here variance is selected as the texture property for the text and non-text verification purpose. So finally filtered components are verified a text or non-text. In the final result, the text, non-text and background is displayed in single image but text, non-text and backgrounds are in black, white and gray color respectively.

A. COLOR TO GRAY SCALE CONVERSION

Color image have a lot of details. These images have detail about the color and gray level (i.e.

brightness level). Color details play important role in the skin identification, medical fields etc. But including the color detail for the text and non-text identification is not useful. In other words, it only increases the processing overhead. So the best way to reduce the overhead is RGB to GRAY conversion of the color image. Color to gray level conversion reduces the complexity of OCR.

B. GRAY-LEVEL HISTOGRAM

Gray level images have the pixels of different gray level. It has the pixels whose gray level value is in between maximum gray level (white color) and the minimum gray level (black color). Gray level histogram (fig 1.) gives the information about how many image's pixels are present at a particular gray level.

Histogram is the best way to show the image detail. It is a graphical representation of the information about how many gray levels that particular image have, and how many pixel are present at these gray level. Histogram processing gives the way to find out proper threshold value to binarize the image.

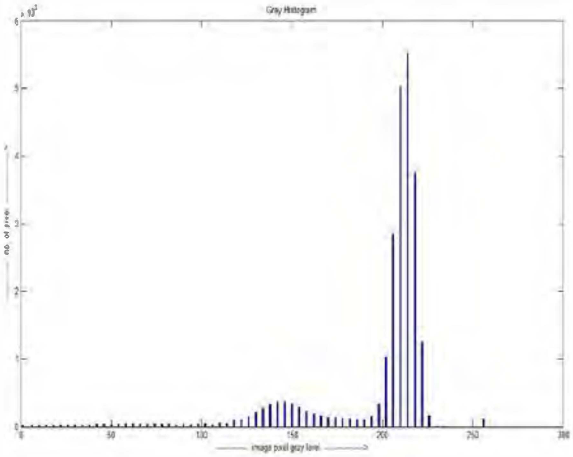

Fig1. Gray level histogram

C. THRESHOLD VALUE

Threshold value is needed to separate the object points from the background. Calculation of the threshold value in depends on the type of image, detail of the objects present in the image, intensity of the overall image. There are many method of the threshold that already exist. Suppose we have an image f(x,y), that have light objects on the dark background. The object and background have gray levels grouped into two dominant modes. Threshold T is a gray level value that separates object pixels from background pixels. A point (x,y) for which f(x,y)>= T is called an object point, else the point is referred a background point. Hence threshold is mathematically defined as function T of the form

$$T = T\,[x, y, p(x, y), f(x, y)] \qquad (1)$$

A threshold image g(x, y) is defined as

$$g(x, y) = \begin{cases} 1 & \text{if } g(x, y) >= t \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Here pixels labeled 1 denote object pixels and pixels 0 denote background pixels. Threshold value can be global threshold, adaptive threshold, optimal global and adaptive threshold or thresholds based on several variables. Otsu threshold method can be used to find out the global threshold value for a given image. It operates directly on the gray level histogram. Basic idea behind the Otsu method is to maximize the between-class variance.

D. BINARIZATION

Binarization is a process of decreasing the gray level resolution up to two gray levels [8]. As color to gray level conversion is the process to decreasing the color depth. An image is binarized using the threshold value. Binarized image have two gray levels, one for the object points and another for the background points.

E. CONNECTED COMPONENTS

In the binarized image, the connected component are found on the bases of 4 –neighbors (N_4 (p)) or 8-neighbors (N_8 (p)). These connected components are the object points that are connected to each other. Finally each connect component is assigned a label, that label uniquely denotes a particular connected component. These labels show the number of object that object points of the given image have. Background of the image is assigned the 0 label.

F. NEARLY CONNECTED COMPONENTS

The connected components that are very near to each other are merging into one by using Morphological closing Operation. The reason of merging is that they don't have useful information in isolation. In result of merging the nearly connected components, a new image is obtained. Now for this image connected component are found and labeled accordingly.

G. BOUNDING RECTANGLES

Connected components may have different region properties attached to them. These properties are used for getting the detailed information of the connected components. There are ‘Area', 'Bounding Box', 'Centroid', 'Filled Area', 'Filled Image', 'Image', 'Pixel Idx List', 'Pixel List', and 'Sub array Idx' region properties that are attached with the connected components. Bounding Box property is used for finding the bounding boxes. Connected components are now enclosed into bounding rectangles that are only of the size to cover the connected component only.

H. HEIGHT HISTOGRAM.

Bounding boxes’ height is collected and this information is stored for further uses. Height histogram is used to represent height information of bounding boxes and used for initial filtering of connected components into text and non-text regions.

I. FILTERING COMPONENTS

A height histogram is created to show the detail about the height of the histogram. The heuristic approach is applied to find out the height of the acceptable bounding rectangles. Initially the text and non-text are identified on the bases of the acceptable rectangles height range. But to get the more accurate results, we apply the variance texture property for the text and non-text identification out of document image.

Connected components are filtered out into text and non-text based on the acceptable height of the bounding boxes. The acceptable height is found out from the height histogram using approach.

J. VERIFING TEXT AND NON-TEXT

Connected components [1] that are filtered as the text and non-text region in the previous step is verified using the texture selected texture property. This texture property verifies whether a connected component that is selected as text is text in real or otherwise. This approach used variance as a texture property for this purpose. Text, non-text area and background that are identified in the approach are finally assigned the different color in the final output.

IV. EXPERIMENT RESULTS

Present approach is based on the connected component analysis and texture analysis. It uses the best part or the other approaches such as bounding box approach, variance based text detection in document image. This approach is advance than the previous approaches because previous approaches are related to detection of the text region only. They do not deal with the issue of the text region that is embedded within the graphic region. So present approach is good in respect of the fact that is tried to identify the text embedded into graphics region or reverse.

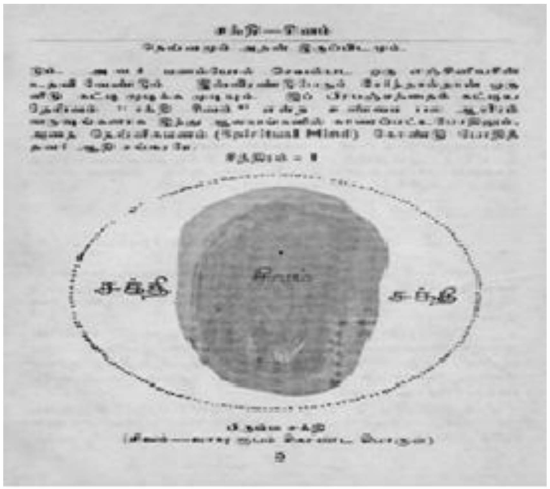

fig 3. Original Document image

Gray level image in fig 3 is given as input the image texture properties are ploted in the fig 5. These texture properties are used for the verification of the text and non-text.

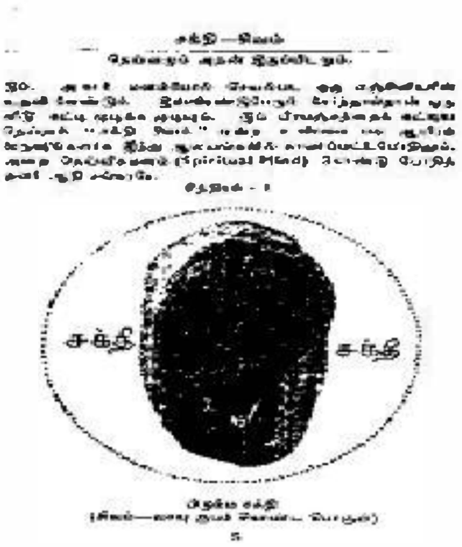

Fig 4.Binary Image

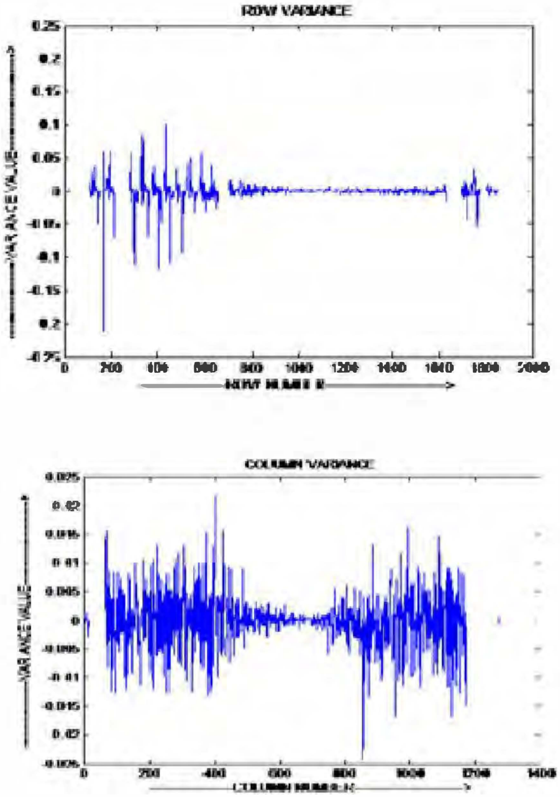

Fig 5. Texture properties

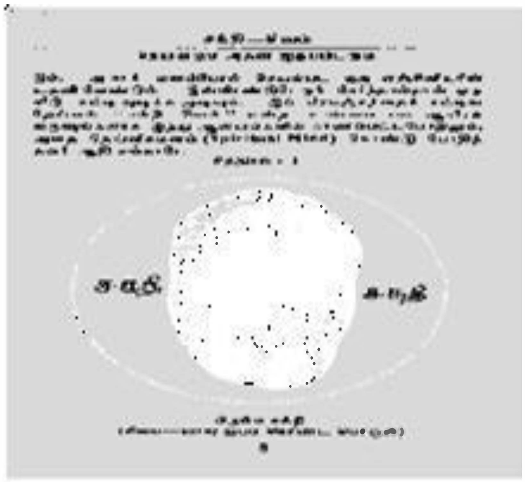

fig 6. Output image text (black), non-text (white) and balckground (gray)

Fig 7. is the input image, that have text and non text in it. fig 8 show the result of the approach.

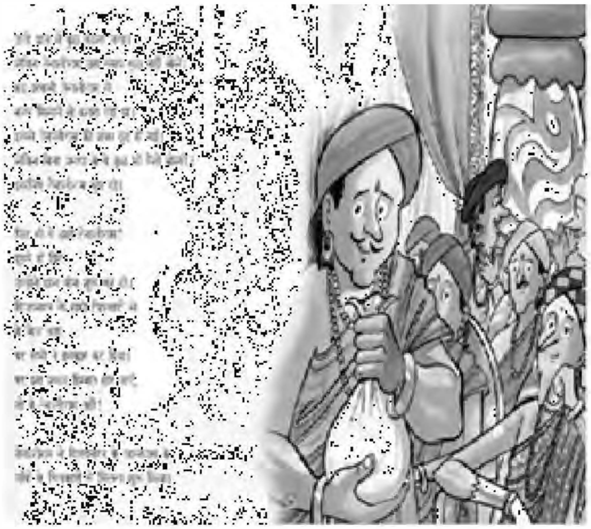

Fig 7. Original image

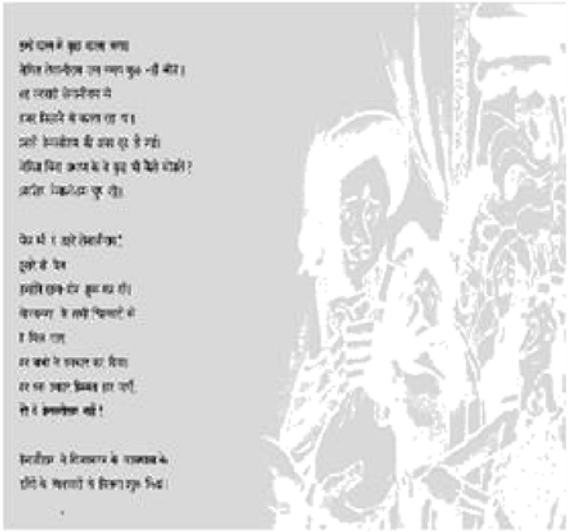

Fig 8. Result image

V. CONCLUSION

Finding text area from the image is a challenging task because text embedded in the image has low pixel intensity which suppresses it. In this paper, a new texture based method for extraction of Text and Non-Text area from the image based on binarization of gray level image using nearly connected components is being proposed. Experimental results show that proposed approach produces effective results which were further verified through texture verification.

REFERENCES

1. Haralick, Robert M., And linda G. Shapiro, “Computer And Robot Vision”, Volume I, Addison-Wesley, PP. 28-48, 1992
2. Otsu, N., "A Threshold Selection Method From Gray-Level Histograms", IEEE Transactions On Systems, Man, And Cybernetics, Vol. 9, No. 1, PP. 62-66, 1979.
3. C.Strouthopoulos, N.Papamarkos And C.Chamzas “Identification Of Text –Only Area In Mixed Type Documents” Engng Applic. Artif. Vol. 10, No. PP. 387-401, 1999 Elsevier.
4. Q. Yuan, C. L. Tan “Text Extraction From Gray Scale Document Images Using Edge Information” Dept. Of Computer Science, School Of Computing National University Of Singapore 3 Science Drive 2, Singapore 2004.
5. Oleg Okun, Yu Yan And Matti Pietik Ainen “Robust Text Detection From Binarized Document Images” Machine Vision Group, Infotech Oulu And Department Of Electrical Engineering, University Of Oulu, Finland January 2005
6. Anoop M. Namboodiri And Anil K. Jain “Document Structure And Layout Analysis” International Institute Of Information Technology, Hyderabad, 500 019, India, Michigan State University, East Lansing, Mi - 48824, USA 2007
7. Zhixin Shi And Venu Govindaraju “Line Separation For Complex Document Images Using Fuzzy Runlength” Center Of Excellence For Document Analysis And Recognition(Cedar) State University Of New York At Buffalo, Buffalo, Ny 14228, U.S.A.
8. Jonghyun Park, Toan Nguyen Dinh, and Gueesang Lee “Binarization Of Text Region Based On Fuzzy Clustering And Histogram Distribution In Signboards” World Academy Of Science, Engineering And Technology 43, 2008.
9. M.Sundaresan, S.Ranjini,”Text Extraction from Digital English Comic Image using Two Blobs Extraction Method”, in proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME) March 21-23, 2012.
10. Samit Biswas, Amit kumar das,”Text Extraction from Scanned Land Map Images”, IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision, 2012.
11. SeongHun Lee, JaeHyun Seok, KyungMin Min, JinHyung Kim, "Scene Text Extraction using Image Intensity and Color Information", Chinese Conference on Pattern Recognition, IEEE 2009
12. Rachid Hedjam, Reza Farrahi Moghaddam and Mohamed Cheriet, "Text Extraction from Degraded Document Images", second Europian workshop on visual information processing (EUVIP) 2010
13. Songhua Xu, Michael Krauthammer, "Boosting Text Extraction From Biomedical Images using Text Region Detection", Biomedical Sciences and Engineering Conference (BSEC), 2011
14. Rachid Hedjam, Mohamed Cheriet, "Novel data representation for text extraction from multispectral historical document images" ,International Conference on Document Analysis and Recognition, IEEE 2011
15. C.P.Sumathi, N.Priya, "Text Extraction from Assorted Images using Morphological-Region, Texture and Multiscale Techniques - A Comparative Study", International Conference on Information Communication and Embedded Systems (ICICES), 2013
16. Sanjay Shah, Chintan Modi, Manisha Patel, "Novel Approach for Text Extraction from Natural Images Using ISEF Edge Detection", International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), 2011
17. Mohammad ShorifUddin, Madeena Sultana, Tanzila Rahman, and Umme Sayma Busra, "Extraction of Texts from a Scene Image using Morphology Based Approach", International Conference on Informatics, Electronics & Vision, IEEE 2012
18. Ranjit Ghoshal, Anandarup Roy, Swapan K. Parui, "Text Extraction from Scene Images using Statistical Distributions", Third International Conference on Emerging Applications of Information Technology (EAIT), IEEE 2012