Text Extraction and Non Text Removal from Colored Images

Shivani Saluja
CDAC Noida

Tushar Patnaik
CDAC Noida

Tanvi Jain
CDAC Noida

ABSTRACT
The objective of this paper is to propose a new methodology for text extraction and non text removal from colored images. Text in web pages, library documents etc. is one powerful source of high level semantics. Existing text extraction methods do not work efficiently in case of images with complex background, different contrast or text embedded in a complex background. Documents fed into OCR work efficiently if they contain only text. The paper has taken into focus several images in several languages (English, Telgu, and Gurumukhi).Several existing text detection techniques have also been discussed in the paper. The approach used is based on preprocessing steps, adaptive thresholding, detecting connected components, generating blobs and finally extraction of only those blobs which consist of textual part.

General Terms
Image Segmentation, Digitizing libraries, Removal of graphics, OCR, pattern recognition

Keywords
Binarization, Pixel, Image, Text, Non text, text localization, Connected Component, Blobs, Color

1. INTRODUCTION
Text extraction is an important step for document image analysis and recognition pipeline. Detecting, segmenting, and extracting text in images which are part of documents is also a very important issue, since more and more document present text in images. Image Binarization is one of the steps that form one of the basic preprocessing steps before text extraction which converts an image (up to 256 gray levels) to black and white images (0 or 1). Document image binarization refers to the conversion of a gray-scale or colored image into a binary image. In most document processing systems, a binarization process precedes the analysis and recognition procedures. The use of two-level information (as 0/1) greatly reduces the computational load and the complexity of the analysis algorithms.

Documents with text embedded in complex colored and textured backgrounds are increasingly common today, for example in magazines, newspapers, magazines and web pages. Robust detection of text from these documents is a challenging problem. The approaches developed for ordinary documents, such as binarization by adaptive thresholding, are not generally applicable, because it seems to be impossible to find an optimal threshold or thresholds to preserve meaningful information and to eliminate unnecessary one.Text Binarization needs to be fast, efficient and robust in order to feed an OCR classifier with the correct input. The input fed into OCR generally required to be in textual form.Non text areas present are generally type of noise for OCR input and reduces the efficiency of OCR.Therefore we need to segment out likely regions of text from the image. This paper introduces a new technique for text binarization in colored document images. The proposed algorithm combines the advantages of several previous approaches for text detection, and utilizes a focus-of-attention approach for text finding.

Text extraction in mixed-type documents is a pre-processing and necessary stage for many document applications. In mixed-type color documents, text, drawings and graphics appear with millions of different colours. The presence of multiple colors increases the complexity of conversion. In spite of directly applying thresholding to input image several intermediate steps are introduced .These interims not only reduce complexity but also helps in generating better results.

2. PROBLEMS IN TEXT EXTRACTION
- Documents include text, line-drawings and graphics regions and can be considered as mixed type documents.
- Textured backgrounds
- Non uniformly illuminated documents
- Degraded documents (bad inking, old paper, with holes)
- Non uniform background (mixing colours and "reverse video")
- Documents with text and background in same color.
- Documents with rounded text or vertical text.

3. EXISTING TECHNIQUES
There are several classes of text detection methods.

One of the classes is ***connected component based*** which make use of spatial information of edges and color components which corresponds to text regions. Text is detected by analyzing connected components of dull colors with horizontal or vertical alignment and also on the basis of several geometric and shape regularity features.

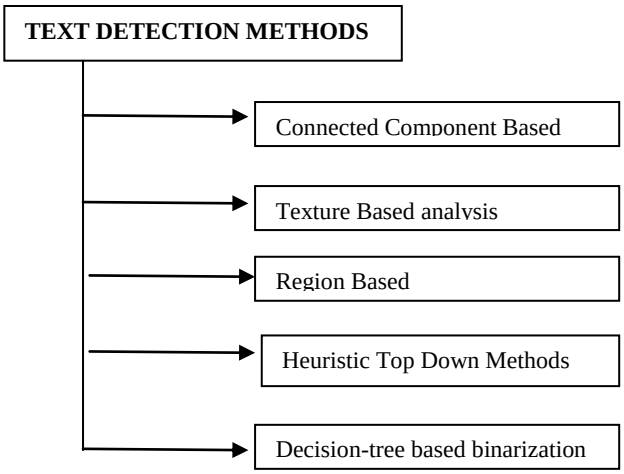

Fig1: Classification of text detection techniques

Second class is ***texture based analysis method***. It is a structural approach to texture analysis and is based on several texture properties derived from Fourier Spectrum.
Gabor filters, variance, etc.techniques can be used to detect the textural properties of a text region in an image. Texture based analysis methods are classified as top down and bottom up methods. This approach is basically used at places where text is assumed to represent a different texture than non-text. With this approach, text is often considered as a periodic texture characters on a text line form a more or less periodic structure in the horizontal direction, while the text lines in turn create periodic structure in vertical direction. Performance of texture based methods is easily affected by the character size and background patterns. It fails in case of very small sized texts.

Third Class is ***region based approach.*** It basically divides images into small regions and then group regions which have characters to get words. Region grouping technique is applied to combine pixels which belong to the common cluster.
It also works on pixel level to identify character regions through color clustering algorithm. It also works on pixel level to identify character region by making use of several color clustering algorithm.

Fourth class is ***Heuristics top down*** approach. Heuristic filters are used to identify text areas in image. It identifies text area in images by using heuristic filters. The image regions are split based on their horizontal or vertical alignment. Baseline detection concept is used based on the fact that text strings are characterized by fixed top and bottom baselines.

Fifth Class is ***Decision-tree based binarization method*** selects various color features to binarize color document images. First, if the document image colours are concentrated within a limited range, saturation is employed. Second, if the image foreground colours are significant, luminance is adopted. Third, if the image background colours are concentrated within a limited range, luminance is also applied.

All different approaches and methods for detecting text, do not solve the problem completely as there are several factors which effect their performance e.g. language, font alignment and size, color and contrast with the background color. The paper throws a light on such approach that will solve all the existing problems.

4. SAMPLE IMAGES

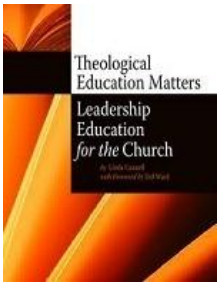

Fig 2: multiple background color

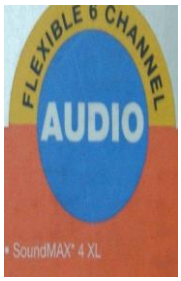

Fig 3: Rounded Text

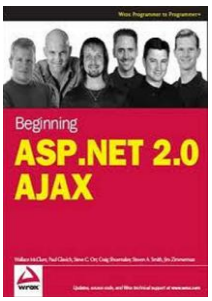

Fig 4: Cover page

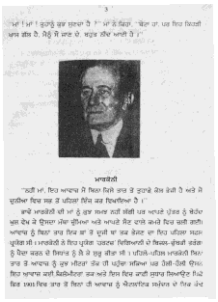

Fig 5: Gurumukhi

Fig 6: Telgu Image

Fig 7: Complex Cover Page

5. PROPOSED APPROACH

The proposed approach works in a sequential manner. The entire system is broken into three steps. *First is preprocessing step Second* is *text detection* and third is finally non text removal from images. Preprocessing basically includes color transformation, noise removal and joining of nearest components. Text Detection deals with detecting connected component, generating blobs or bounding boxes around the connected components .Several conditions and features (geometric, textural, shape regularity) enables us to separate the text blobs. *Third* is *non text removal or background elimination*. This is done by comparing with the original image points which are present in the localized text boxes are retained and rest are removed.

Text extraction enables us to segment our image into two parts textual and non textual which if required can also be processed separately. Text part describes the actual content of the document.

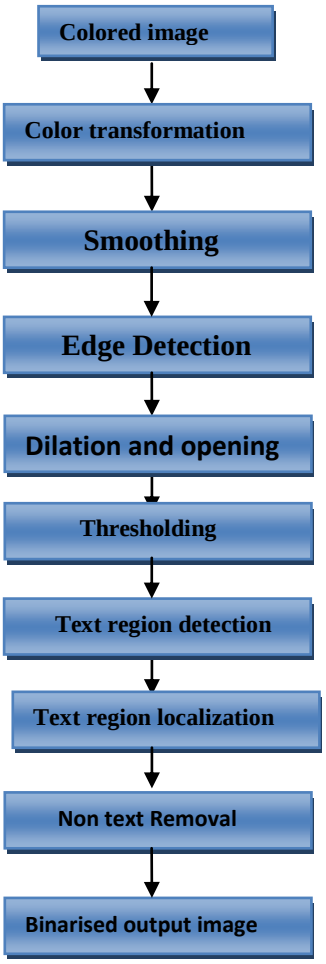

Fig 8: Proposed Approach

5.1) Color Transformation

Color transformation as the name suggest involves transformation of colored document image. It is very cumbersome to handle colored documents because of high variation in intensity of pixels of document. A color transformation is a *two step* procedure that comprise of color reduction and gray scale conversion.

Color reduction is accomplished using *pyrmeansegmentation* algorithm. It merges the low contrast areas in the background by making use of dominant colors present in the image and creates large compact areas.

- Divide the entire image into small cubes
- Calculate intensity value of each pixel in each cube
- Each pixel is replaced by the mean value of color intensity of its 9 nearest neighbors

Gray Scale Conversion:

Initially the image is split into three color planes red, green and blue. Feature based techniques like saturation, hue and luminance are used. Luminance corresponds to brightness in the color and Saturation is measure of amount of white in color. Colored image is converted into gray scale using standard conversion Grey(lum)=0.3R+.59G+.11B where R,G and B corresponds to red, blue and green color respectively with values ranging from 0 to 255.

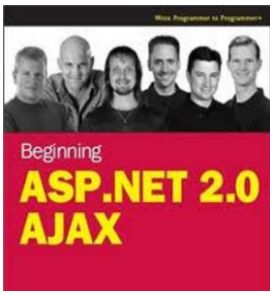

Fig9: Input Image

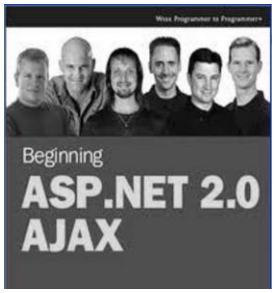

Fig10: Color Transformation

5.2) Smoothing

Images taken with both digital cameras and conventional film cameras will pick up noise from a variety of sources. Smoothing, also called blurring, is a simple and frequently used image processing operation to remove noise from images. In salt and pepper noise (sparse light and dark disturbances), pixels in the image are very different in color or intensity from their surrounding pixels. In Gaussian noise, each pixel in the image will be changed from its original value by a (usually) small amount. Low pass filtering is used for noise removal. Several filters like arithmetic mean filter, geometric mean filter, harmonic mean filter etc can be employed. In this approach a Gaussian smooth over a 3 × 3 area centered on each pixel is used. Gaussian filtering is done by convolving each point in the input array with a Gaussian kernel.

Fig11:Smoothing

5.3)Edge Detection:

Edges are the boundaries separating regions with different brightness or color. It is one of the preprocessing steps which are introduced to improve efficiency and accuracy of the final output. It takes grayscale image on input and returns bi-level image where non-zero pixels mark detected edges. The resultant edge image obtained is sharpened in order to increase contrast between the detected edges and its background, making it easier to extract text regions. With good edge detection result, the time complexity of later text detection will be reduced and the detection accuracy will be improved.

5.4)Dilation And Opening:

Dilation and closing are used in most of the more sophisticated connected-component algorithms to reduce unwanted or noise-driven segments.

Dilation adds pixels to the boundaries of objects in the image. The number of pixels added depends on the size and shape of the structuring element. In this approach a structuring element consisting of zeroes and minus ones is taken and a pointer is maintained for the structuring element. Minus ones determine neighbors of each pixel from which maximum is taken and put to the corresponding destination pixel.

Opening generally smoothes the contours of the image breaks down narrow bridges and eliminates thin protrusion. It isolates objects which may be just touching one another. Opening operation is used to connect nearby large regions.

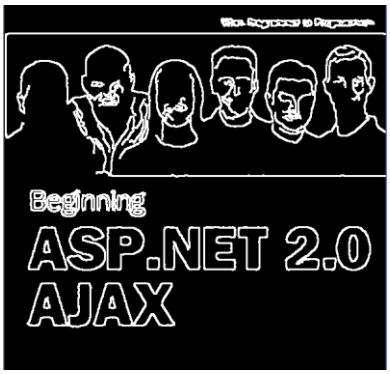

Fig12: Dilation &Opening

5.5)Thresholding:

The simplest and earliest method is the *global thresholding technique* that uses a single threshold to classify image pixels into foreground or background classes. Global thresholding techniques are generally based on histogram analysis. It works well for images with well separated foreground and background intensities. The global thresholding algorithm chooses a fixed intensity threshold value T (from 0 to 255). If the intensity value of any pixel of an input image is more than T, the pixel is set to white otherwise it is black.

Most of the document images do not meet this condition and hence the application of global thresholding methods is limited. In order to deal with such type of images, *adaptive thresholding* is used in this approach. The adaptive threshold T(x, y) is set on a pixel-by-pixel basis by computing a weighted average of the b*b region around each pixel location. Pixels in the b* b region are classified on the basis of threshold. Robust binarization gives the possibility of a correct extraction of the sketched line drawing or text from its background.

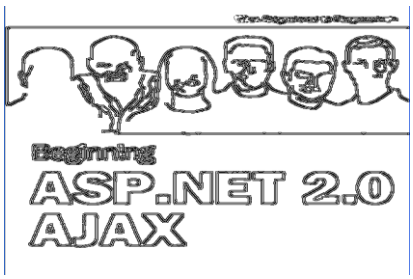

Fig13: Thresholding

5.5)Text Region Detection and Localization:

The text detection stage seeks to detect the presence of text in a given image. It is one of the most important steps in text binarization. It includes connected component detection, bounding box formation, detecting overlapped boxes, applying internal box constraints, verification of several textual properties(geometric and shape) like size, height, compactness ratio, aspect ratio; density etc.This leads to removal of few of the bounding box drawn around the non textual region. Further horizontal and vertical profiling is done in the boxes to divide boxes containing more than one line of textual words. After text region localization the area of image not included within the identified blobs is retained and rest is removed.

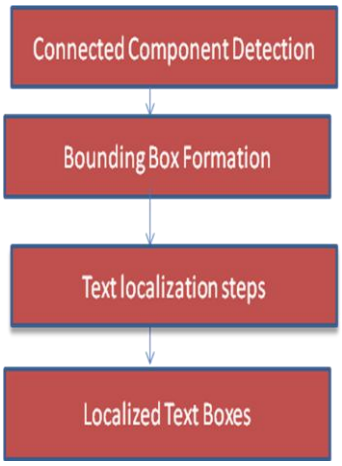

Fig14: Text Localization

5.5.1)Connected component Detection:

It is a two step process involving detecting of connected component as well as grouping similar type of connected component.Connected Component are detected in the image based on the analysis of geometric arrangement of edges, homogeneous color, pixel connectivity (4 or 8 connected) or gray scale components that belong to several characters present in the textual region. C*onnected component* is classified as *small*, *medium* and *big* by the following rules. A connected-component is labeled as *small* if the number of pixels *NP* in the component is less than four and greater than one, *big* if *NP* is greater than image area multiplied by a .0065 and otherwise, it is *labeled as medium.* Moreover connected-*component ratio* is the number of connected components divided by the total number of connected-components in the detected sequence.

Algorithm for detecting connected component

- Scan the image along the row until it comes to a point p(p:pixel to be labeled)
- Examine the eight neighbors of p.
- Label p as follows:
  - If all neighbors are 0 ,assign a new label to p,else
  - If only one neighbor has value 1,assign its label to p,else
  - If more than one of the neighbor has value 1,assign one of the labels to p and make a note of equivalences

Fig13: Connected Component Detection

5.5.2)Bounding Box Formation

After identification of connected components bounding boxes are generated around them.Connected components are also identified in the non textual region,in order to eliminate small non textual region boxes are pushed only around those group of connected component whose area is greater than 80 , perimeter is greater than 20 and also no of connected components is greater than 8.

Bounding boxes can be defined as simply rectangles enclosing the group of connected components.Each and every bounding box has its unique width and height.

Steps for bounding box generation:

- Find starting and end point of connected component
- Height of bounding box=
  {y coordinate of last pixel of connected component-
  y coordinate of first pixel of connected component}
- Width of bounding box
  {x coordinate of last pixel of connected component-
  x coordinate of first pixel of connected component }

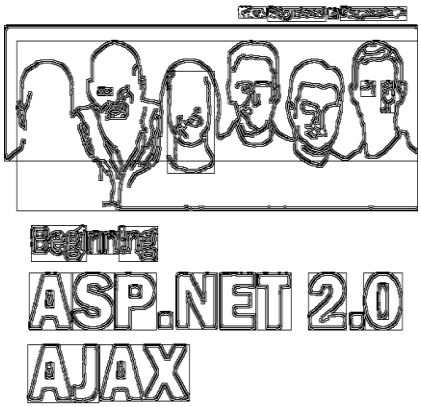

Fig14:Bounding Box generation

5.5.3)Text Localization:

This step is responsible for locating text blocks in the image generated by following steps.

- Remove the outer boundaries and rectangles from the image if any
- Detect overlapped boxes
- Removal of irrelevant boxes by using internal boxes count
- Horizontal profiling
- Check for text alignment ;horizontal or vertical
- Apply geometric and shape regularity features
- Eliminate area of the image whose coordinates do not lie in the derived bounding boxes.

Remove the outer boundaries and rectangles.

Prior to applying several localization steps, image is scanned for rectangles. Rectangles are searched in the image by making use of template matching. Outer boundaries are removed on the basis of largest sized rectangles or rectangle whose height and width is greater than three-fourth of the image width and height.

Detect overlapped boxes

In the generated bounding boxes there might be some boxes present that overlap each other. Such overlapped boxes lead to detection of localization of inappropriate boxes.

- HBD(i,j)=MAX(XL(i),XL(j))-MIN(XR(i),XR(j))
- VBD(i,j)=MAX(YL(i),YL(j))-MIN(YL(i),YL(j))
- If HBD(i, j) < 0, CCi and CCj overlap in the vertical direction
- If VBD(i, j) < 0 , CCi and CCj overlap in the horizontal direction
  HBD (i, j) >.=0AND VBD (i, j)>=No connection between CCi and CCj

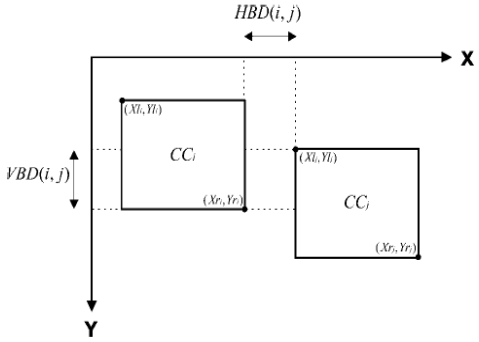

Fig15: Overlapped Boxes Detection

Detection internal boxes

It is usually observed that textual regions contain less than three internal boxes .Non textual region comprise of more than 3internal boxes. Internal boxes are detected by means of following conditions.

- Scan the entire image
- Read the x co-ordinate, y co-ordinate, height and width of bounding box.
  For each bounding box B1 check whether
- B1.x>B2.x andB1.x+B1.width-1<B2.x+B2.width-1
- B1.y>B2.y andB1.y+B1.height-1<B2.y+B2.height-1

If 1 and 2 are satisfied then box is considered internal

After detection of internal box the number of internal boxes within each box is counted. If the number of internal boxes is less than 3 then outer bounding box is accepted and inner ones are rejected. Otherwise inner ones are accepted and outer bounding box is rejected

.

Horizontal profiling

Horizontal profiling takes into consideration the no of black pixel in each row. The no of black pixels in each row are counted and the row in which the count falls below the threshold level its index is stored.

This is basically done to divide boxes which comprise of more than one text line.

Check for text alignment; horizontal or vertical

This step tries to locate any possible text based on its height. It is based on the fact that half of the characters have vertical straight line and around 90%of English words contain such

character. This step focuses on finding such vertical straight lines which are part of text and then expanding their scan area on both sides of these lines to determine the lower and upper limit of candidate text blocks. So this means that if any sentence or word contains such character which has vertical straight line; it would be detected by this step

Apply geometric and shape regularity features
Remaining bounding boxes are presumed to contain only textual area .Several geometric and shape regularity features are applied in order to eliminate any of the boxes which consist of non textual region.
Geometric Features:
1) Min size: 20 pixels
Detecting any text which is smaller than certain threshold value is of no use for the system as this text block cannot be recognized later by the OCR package
2) Text can be approximately of same height throughout the sentence.
3) Characters and text cannot be represented in an image by only few pixels. If the process finds that this pixel is connected to other sufficient number of pixels to form text, it considers that pixel as part of text.
4) Inter-character distance: characters in a text line have a uniform distance between them.
5) Aspect Ratio: It is defined as ratio of width to height of bounding box. It is constrained to lie between 1.33 and 20 for textual region.

Shape Regularity Features:
1) Occupy Ratio
Occupy Ratio defines how much of the bounding box region is covered by CC's.
2) Compactness:
It is a feature which divides the area of CC's by the square of CC's perimeter. It is generally less than 800.
3) Filled Area:
Most of texts do not have large gaps in proportion to their areas.

5.5)Non text Removal(Background Elimination)
The newly formed image from the previous step is compared to the original image the region lying within the bounding box is retained. This is done by subtracting original image from the newly formed image. The output of this step is the image comprising of only textual region.

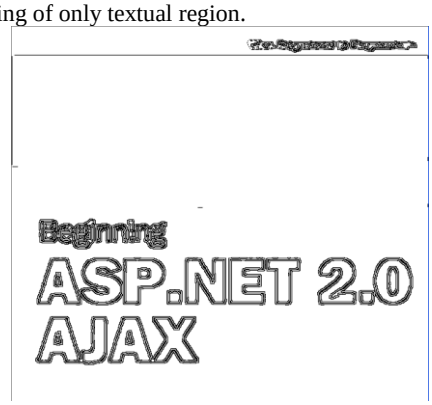

Fig16:Output Image with only textual area

Table 1:Result Analysis

Type of Image	No of Images	Text Detected completely	Text detected along With non text	Text not detected
Colored image (textual)	**50**	**50**	**0**	**0**
Colored Image from text books with both text and non text (English)	**100**	**85**	**15**	**0**
Images with both text and non text(Gurumukhi)	**30**	**22**	**8**	**0**
Images with both text and non text(Telgu)	**90**	**80**	**10**	**0**
Cover images	**70**	**32**	**30**	**8**

6. ACKNOWLEDGEMENTS
Special thanks to my respected guide Mr.Tushar Patnaik who has encouraged me to write this paper and helped me in each and every thins and thick of my project.

I also want to thank my parents, my sister and my brother in law who have been with me during this project and have constantly supported me.

7. REFERENCES
[1] Chen D, H. Bourlard, 2001. And J. -P. Thiran, "Text identification in Complex Background using SVM, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 621-626.*

[2]Tushar Patnaik,Shalu Gupta,Deepak Arya "Comparison of Binarization Algorithm in Indian Language OCR", *Proceedings of ASCNT – 2010, CDAC, Noida, India, pp. 61 – 69*

[3] Efthimios Badekas, Nikos Nikolaou, Nikos Papamarkos "Text Binarization in Color Documents", *2007 Wiley Periodicals, Inc.*

[4] Efthimios Badekas, Nikos Nikolaou, Nikos Papamarkos "Font and Background Color Independent Text Binarization", *2007 Wiley Periodicals, Inc*

[5]J. He, Q. D. M. Do, A. C. Downton, J. H. Kim, "A Comparison of Binarization Methods for Historical Archive Documents," *Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp.538542, 2005*

[6] Huang, Huadong Ma, He Zhang, "A New Video Text Extraction Approach" *IEEE International Conference on Multimedia and Expo, 2009. ICME 2009*

.[7]Keechul Jung, Kwang In Kim, Anil K. Jain "Text Information Extraction in Images and Video*: A Survey*

8. AUTHOR'S PROFILE

Shivani ,done Btech in Computer Science with 79.42% in 2010 from UPTU(Lucknow).Currently pursuing MTech in Information Technology from CDAC Noida(GGSIPU New Delhi).I am doing project on Text Extraction and non text removal from colored document images.