Text & Non-Text Segmentation in Colored Images

Nitesh Kumar Singh, Avinash verma, Anurag kumar

Abstract— The purpose of this paper color images with complex background for text and non-text segmentation is to propose a new system. The existing text extraction methods in the case of images with complex background do not work efficiently. Locating text in case of variation in style, color, as well as complex image background makes text reading from images challenging. Here the approach used is based on pre-processing steps, edge detection, CC-analysis, bounding rectangles, segmentation and finally extraction of only those blobs which consist of textual part. This approach is tested successfully across various images taken manually and from internet.

Index Terms— Edge detection, Connected Component-analysis, Bounding rectangle, Segmentation, Smoothing, Binarization, Aspect Ratio.

INTRODUCTION

Text embedded in color images such **as** book covers, video frames, natural scene images and WWW images can provide some important information for peoples. Practically, with the rapid development of multimedia and internet nowadays, in many applications, we need to convert texts in images to electronic data for multimedia retrieval, book indexing, document analysis, OCR, industrial automation or recognize texts in images captured by camera.
This topic is a part of OCR when documents fed into OCR work efficiently if documents contain only text. A non text area keeps noise for OCR input and reduces the efficiency of OCR. OCR converts the scanned images of books, magazines and newspaper into machine readable text.
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.

Problem issues-
Due to the size, style, orientation and alignment as well as reduced image contrast and adaptation of text from the complex background images from differences in automatic text extraction is the problem**.**

1. Noise in Input Image:

In an image noise can be due to various sources. The main sources of noise in the input images are: **a)** .Due to the quality of the paper on which the printing noise. b) Because of the noise from the paper "Back Paper noise" to print on both sides of the cause. c) Noise and brightness sensors for the scanner source said. These all contribute to noise reduction in the accuracy of OCR systems do. As a result of this, in place of a noise correction routine is inevitable.

2. Skew in Input Image:

A multi-line text in the image having a slant creates problems in determining lines of text. In an image problem because of Hindi or Sanskrit Devanagari script ascenders and descenders to detect the presence and becomes more complex.

3. Images embedded with Text in Input Image:

Histogram analysis becomes impossible to use the line as a division within the text creates problems in terms of images. Because of the presence of irregular images are emerging that another problem to deal with variations in font and font size. Font size variation is large, it is the lack of accuracy, can be confused with irregular size image

RELATED WORK

In the present era , there are different methods in dealing with text detection and reorganization in images .Some approaches for text detection are classified into three categories. These three categories are**:-** texture-based methods, region based methods, and hybrid methods. **Texture based method** [7][8] is a structural approach to texture analysis and is based on several texture properties derived from Fourier Spectrum. Texture based analysis methods are classified as top down and bottom up methods. Performance of texture based methods is easily affected by the character size and background patterns. It fails in case of very small sized texts**. Region-based methods** [8] in a text field or the background color or gray scale differences in the properties or use the properties of alignment. The third category, **hybrid methods** [7] is a fusion-based and texture-based method.

Zhiyun Ren, Linlin Huang [6] proposed a fast and accurate method to detect text in color images with complex background. Firstly, fast color quantization algorithm is used to convert color images into many image layers, and then a CC-based analysis algorithm is applied to obtain candidate text regions in each layer.

Wu, Shao-lin qu, qing zhuo [5] has proposed a method for detecting and locating text in complex images using some more discriminative features such **as** color, edge and comer. This method uses color feature in spatial color quantized map and

edge feature in edge map to generate bounding blocks, which can reduce missing alarms in "**background-like**" text areas.

Although, these existing methods have some positive results, they still have some disadvantages in terms of computational complexity and low accuracy. For example, some heuristic rules-based methods that reduce reliance on texture, texture classification phase, even if they suffer from their computational complexity. On the other hand, region-based methods can identify texts at any scale, but it is very hard to segment text components accurately from a complex background.

Generalized complex background and appropriate proposals in text characters, scenery pictures can be embedded in text size, shapes and orientations are different document or Web, e-mail, images, text harder to find. Because of the complex background images directly through OCR software to recognize the text is impossible. Thus, we explore text and image areas that need the same inclination. Detection process described in this survey is consistent with the text extraction algorithm [2]. Yi grouping method and obtained the text line image or text in a candidate set through the line adjacent character grouping, gradient feature and color image segmentation based on connected component detection.

Fig: a

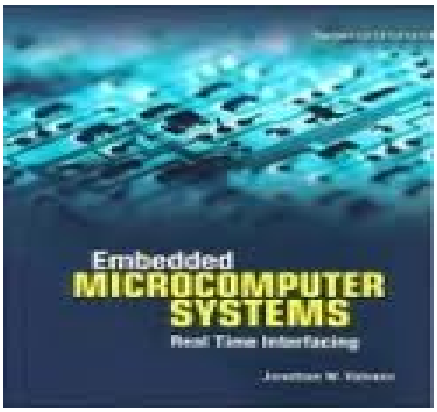


Fig: b

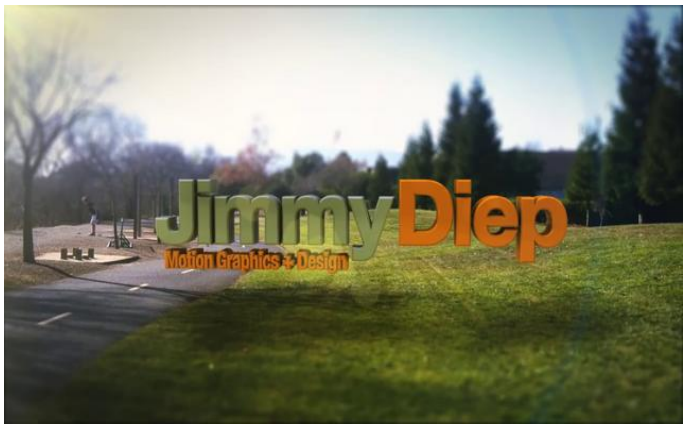


Fig: c

Fig: d

Fig 1: Sample Image

PROPOSED APPROACH

The proposed approach works in a sequential manner. The entire system is broken into three steps. First is pre processing step Second is text detection and third is finally segmentation. Pre-processing 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. 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 2: Initial process

1. Preprocessing-

Here, the input image background is removed using different algorithms and image is gray scaled,then binarized (In black and white) and at last stored in matrix of binary values. In preprocessing stage noise in image is removed by using following steps:

1.1 Color Transformation:

As the name suggest, color transformation 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 twostep procedure that comprise of gray scale conversion and binarization.

1.1.1 Gray scale conversion:

- Loaded Input image can be JPG or BMP
- Retrieves the properties of image like width, height and n channels.
- Get the pointer to access image data
- For each height and for each width of the image, convert image to grayscale by calculating average of r, g, b channels of the image convert to grayscale manually.

The input RGB color image to a gray-scale image intensity change as follows: Y = 0.299R + 0.587G + 0.114B

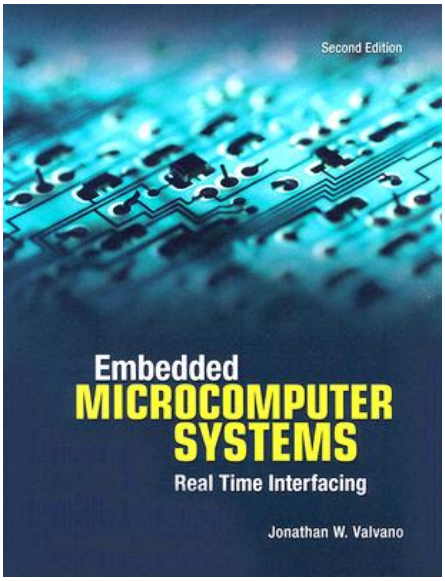


Fig 3: Input Image

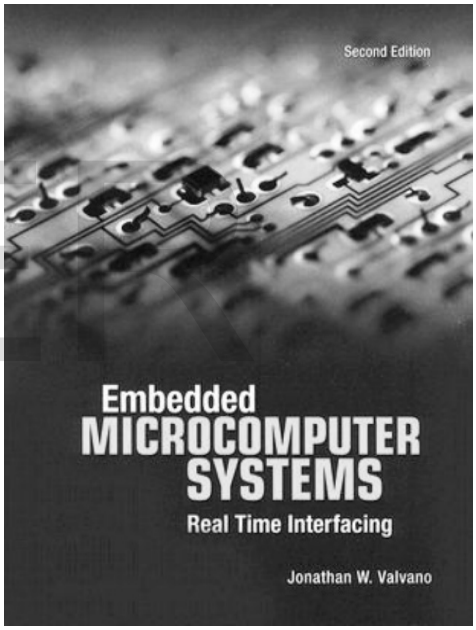


Fig 4: Gray Scale Image

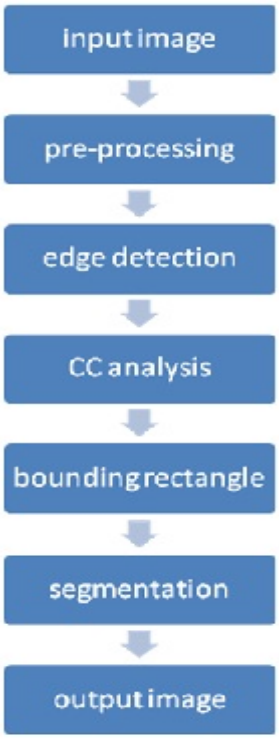


Fig 5: Proposed Approach

1.1.2 Binarization:

Then grayscale image (Y) is converted into binary image (Black and White) The binary image characteristic function are as shown below.

b(x, y) = 1 if g(x, y) < T
= 0 if g(x, y) > T , T= Intensity mean

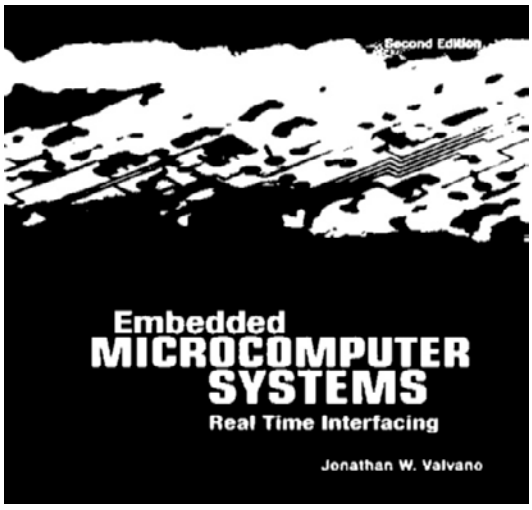


Fig 6: Binarized image

1.2 Smoothing:

Both digital cameras and conventional film cameras will pick up noise from images taken with a variety of sources. , Also called blurring, smoothing to remove noise from images in a simple and frequently used image processing operations. Salt and pepper noise (sparse light and dark disturbances), the image pixels from the pixels around them are very different in color or intensity. In Gaussian noise, each pixel in the image is a (usually) small amount will change from its original value. Low-pass filtering is used to remove noise. Filters such as math, geometric mean, mean filter, harmonic mean filter, etc. can be employed. The approach used is a Gaussian centred at each pixel smooth over a 3 × 3 area. Gaussian filter with a Gaussian kernel is done by convolving each point in the input array.

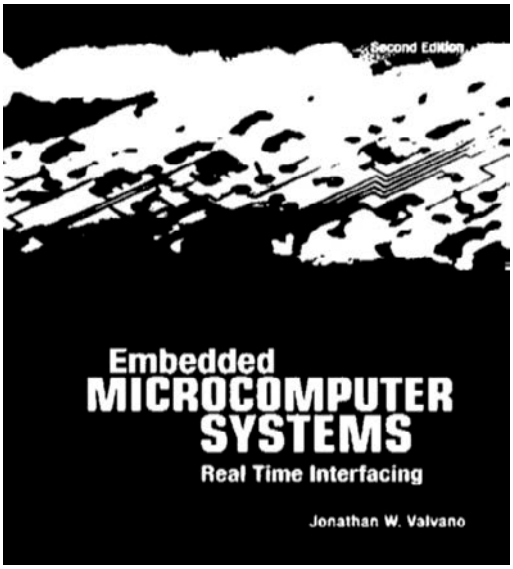


Fig 7: Smoothing

2. Edge Detection-

Edges separating regions with different brightness or color limitations. The final production capacity and to improve the accuracy of the pre-processing steps are introduced to one. It takes input gray scale image and the non-zero pixels mark detected edges, which allows bi-level image. The resulting gain an edge detected image edges and the background, making it easier to extract text fields in order to increase the contrast between fast. With good edge detection result, the detection time of the text will reduce complexity and detection accuracy will be improved.

.

Fig 8: Edge detection

3. Connected Component Analysis

At this stage, we detect all the connected components in the binary image. Then we satisfy conditions to the candidate text regions and merge them as a record of the connected equipment. Most of the text is displayed in a color image areas are always arranged in horizontal lines. Similar to the character of the same text line connected component must meet several conditions. The horizontal distance between them is quite low, whereas, for example, the top and bottom of the connected devices should be roughly the same pixel on the line. CCS such as text fields are marked candidate. An example of a text line is shown in the image below:

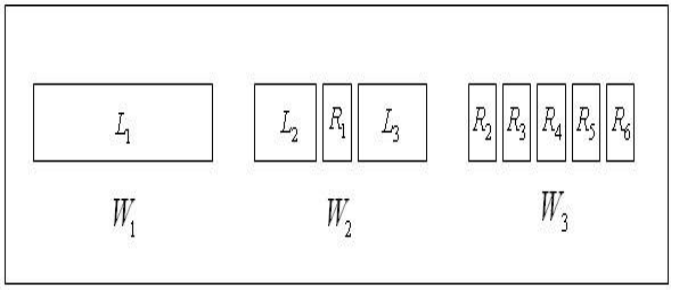


Fig: 9 an example of a text line

W1, W2 and W3 are three English words in the same text line. W3 represents that all the letters in a word are separated into different CCs which should have almost the same width.W2 shows that some letters of a word may be assigned into a same connected component while some letters are alone in their own CCs. W1 indicates the situation that the word is assigned into just one CC. Our method first will merge the connected components represented by W2 and W3. Then the whole text line will be marked as a whole.

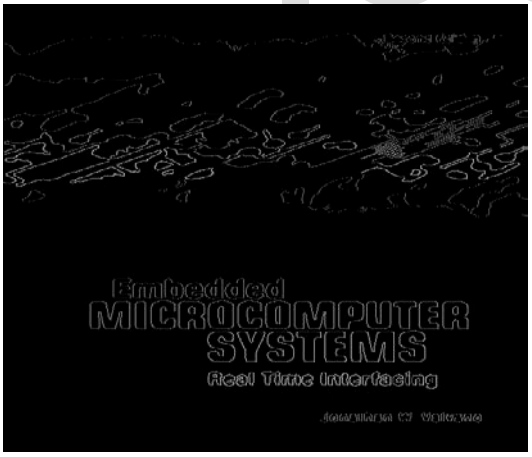


Fig 10: CC detection

4. Bounding Rectangle 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}

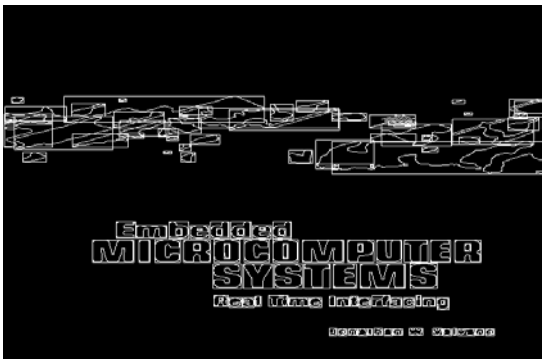


Fig 11: Bounding box generation

5. Segmentation

This stage tries to segment the input image using different segmentation techniques, the filling segmentation which uses geometric and shape regularity features. This is the useful segmentation technique because it allows us to segment the text letters in image even if they are tilted and not contiguous. Each segmentated character thinned and scaled. The segments are processed individually to make the recognition easier.
So this means that if any sentence or word contains such character which has vertical straight line; it would be detected by this step:

5.1 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) Throughout the text, the sentence may be approximately the same height.
3) Characters and text cannot be represented only some pixels in an image. Process for the pixel text form that is connected to a sufficient number of other finds of pixels, the pixels as it considers part of the text.
4) The inter- character spacing: characters in a text line are the same distance between them.
5**) Aspect Ratio:** The bounding box is defined as the ratio of width to height. The textual region between 1.33 and 20 is constrained to lie

5.2 Shape Regularity Features:

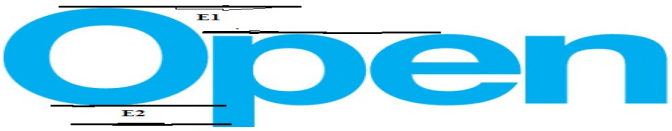
1) Capture Ratio: The ratio is covered by CC defines how to capture the bounding box area.
2) Compactness: the CC copy from the perimeter of the square is a feature that divides the area. It is generally less than 800.

3) Field filled: Most of the texts are not large gaps in proportion to their fields

5.3 Segmentation of CC:

5.3.1 Merging of single letters:

Considering the characteristics of English letters, not all the single letters have the upper or lower borders on the same pixel line as shown in Fig 3. In order to make all the letters of a word contained in a same CC, E1and E 2 shown **in below Fig: 12**



Moreover, the difference between the two adjacent letters should be shorter than a threshold which is set to be 15 pixels in our method.

5.3.2 Merging of English word:

The distance between the two nearing English words is much longer than that of two nearing letters. We assume that such distance is less than twice the width of a single letter. Since it's harder to obtain the accurate width of a letter, we take the height as the parameter to set the rules instead. In that case d1, d2 and d3 shown **in below Fig: 13**

Same|d1|Same|d2|But|d3|Different

Fig: 14 Output image

CONCLUSION

Text detection from any kind of images like document, digital camera based and web, email is challenging due to the random text appearances and complex backgrounds. In this paper a new method for text localization and extraction is proposed. The text detection based on the CC analysis reduces the computation time. Currently for robust text detection and recognition for a particular text image as input is done by applying different preprocessing method to remove complex background. First we detect connected components in image and then character grouping is performed to detect text characters then, text recognition is performed to identify text in input image. Complex background for recognition and OCR text extraction from text normalization based methods for developing this approach to learning.
In the future work, special technique should be investigated to segment the character from their background before putting them into OCR software.

ACKNOWLEDGMENTS

Special thanks to my respected guide **Mr. Avinash Verma** who has encouraged me to write this paper and helped me in each and every thins and thick of my project and also thanks to my friend **Anurag Kumar** who helped me in my research work.

REFERENCES

[1] R.C. Gonzalez, R.E. Woods, Digital Image Processing second edition, Prentice Hall, 2002.

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

[3] Learning OpenCV by Gary Bradski and Adria Kaehler Copyright © 2008 Gary Bradski and Adrian Kaehler. All rights reserved.Printed in the United States of America.

[4] **Digital Image Processing** Third Edition Rafael C. Gonzalez University of Tennessee Richard E. Woods NledData Interactive Pearson International Edition prepared by Pearson Education PEARSON PrenticeHall.

[5] wu, shao-lin qu, qing zhuo. wen-yuan wang, "automatic text detection in complex color image", Department of Automation, Tsinghua University, Beijing, 100084, P. R. China, 4-5 november 2002.

[6] Miss. Poonam B.Kadam, Mrs. Latika R. Desai, "A Hybrid Approach to Detect and Recognize Texts", Assistant Professor,Computer Detartment ,D.Y.P.I.E.T, Pimpri, Pune,Indiain Images,2013.

[7] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben "A Robust Algorithm for Text Detection in Images", Dept. of Math. & Computer Science, University of Marburg, D-35032 Marburg, Germany,2006.

[8] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, Andrew Y. Ng "Text

Detection and Character Recognition in Scene Images with Unsupervised Feature Learning", Computer Science Department Stanford University, 2007.

[9] Narasimha Murthy K N, Dr. Y S Kumaraswamy, Professor, Dept "A Novel Method for Efficient Text Extraction from Real Time Images with Diversified Background using Haar Discrete Wavelet Transform and K-Means Clustering", Dept of Information Science and Engineering, V T U, City Engineering College, Bangalore, Karnataka, 2011.

[10] Gang.Zhou, Yuehu.Liu, Zhiqiang.Tian "SCENE TEXT DETECTION WITH SUPERPIXELS AND HIERARCHICAL MODEL", Institute of AI and Robotics, Xi'an Jiaotong University, Xi'an, P.R.China, 710049, 2012.