A Study: Extraction of Text on Image

Ranjith Kumar T[1], M Jagdeeshwar[2] , Pala Prathima[3]

[1] [1]*Research Scholar, Dept. of Computer Science,Chaitanya Deemed to be University,Warangal Urban,Telengana,India.*
[2]*Professor,Dept.of Computer Science,Chaitanya Deemed to be University,Warangal Urban,Telengana,India.*
[3]*Assistant Professor,Dept.of Computer Science,Chaitanya Deemed to be University,Warangal Urban,Telengana,India.*

ABSTRACT

The aim of the paper is to get idea that how text is extracted from image. While extracting the text from image there are so many variations like text from shaded image or different font styles or different colors or different alignments and sizes. These many variations make difficult to extract the text from image. So to overcome these problem and extract text from image. We process the extraction in four steps, first step the color image is converted to grayscale. Second after that remove the noise and identify the edges. Third convert to binary image and fourth using OCR the text is identified form image.

Keyword : *canny edge, OCR, binary image, color image*

1. INTRODUCTION

In today technologies, most of the data is in unstructured format like text, image, audio, video etc. So, if data is in text format and converting the unstructured text data into structure format we use Natural Language Processing and sentimental analysis algorithm for analysis. But if the data is present in image format then extracting the text data from the image is very complex. The text on the image will be different size, style, orientation and alignment as well as the image color is low or high. To handle this problem, first the text data presented on image is to be extracted after extracting the text we can convert it into structure format and can analyses easily. So this paper gives an idea about how the text is extracted from the image. To detect text from image various techniques are implemented. Many authors explained in their papers with different techniques to extract the text data from image

2. LITERATURE REVIEW

Some of the methods used to extract the text from images are Connected Component Based Method, Edge Based Method, Region Based Method, Mathematical Morphology Based Method and Optical Character Recognition(OCR).

Talukder et al.,[1] proposed a Connected Component Based Method to identify small or large texts in the image. In this method first text extraction starts with the conversion of the color image to gray scale image after that converts the gray scale image to binary image. From the binary Image each text region is marked and the text is extracted from the image.

Rampurkar *et al.*,[2] proposed two different techniques using morphological Based Method. First technique is adjacent Character grouping and second techniques is text line grouping. These techniques can find text by using structured-based partition and grouping method morphological operations.

Grover *et al.*,[3] proposed a simple edge based feature to extract text from image using five steps. First conversion of color image to gray scale. Second, Edge detection by sobel operator, non-maximum suppression, and thresholding. Third, partitioning edge image into small non overlapping blocks and calculating an edge based feature for each block. Fourth, based on the value edge based feature satisfies either as text or non-text based. At last, post processing of the result obtained for improvement of text detection.

Liu [4] proposed a multiscale edge based text extraction algorithm. This algorithm is used to detect and extract text automatically. In this algorithm to extract the text the method consists of three stages: Candidate Text Region Detection, Text Region Localization and Character extraction.

Saravanan [5] discuss the conversion of color image to grayscale image complicated process. To convert the color image to grayscale a new algorithm is proposed by author that algorithm contains 18 steps. This algorithm performs RGB approximation, reduction and addition of chrominance and luminance.

Noel *et al.,* [6] presents a paper on edge detection in grayscale image using grid Smoothing. In this image is represented as a graph and in it each node is represented as pixels and edges are detected using an adequate cost function and density of points in region. The result image is a sub graph representing the edges present in original image. The algorithm is tested on original image and result image.

Coates *et al.*,[7] discuss reading text from scene image is challenging problem. For text detection from image and character recognition different techniques are implemented. Authors proposed a new model in his paper i.e. large-scale algorithm. It learns the features automatically from unlabeled data.

3. PROPOSED METHOD

The proposed method is based on the concepts referred, edge detection is best for text extraction. In this method First step, the color image is converted into gray color image. Second step is to take input as gray color image and find out the edges using canny edge detection method. Third Step change after edge detection converts into binary image. Last step using Optical Character recognization text is extracted. The block diagram of our proposed method is shown in Fig. 1:

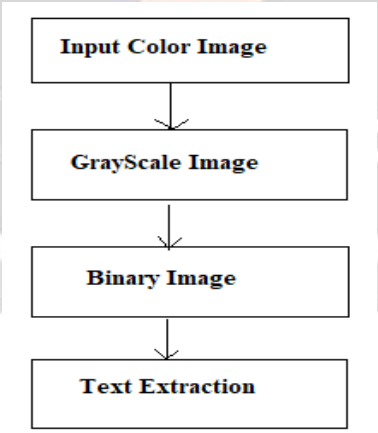

Fig.1: Block diagram of proposed method

Step 1. To convert the color image to grayscale image we have two possible methods

i. Average Method
ii. Weighted Method

Average Method

This method is most simple method to convert the color image to grayscale image. First most of the image will be in the combination of RGB colors. So, in this method we find the average of these three colors. i.e., we add red, green, blue and divide it by 3

$$Grayscale = (Red + Green + Blue)/3$$

But there is a problem in this method all the three colors have different wavelength and have their own contribution in the formation of image. But in the three colors one have more effect and other have less effect for appearance. So while conversion we are taking average of three colors i.e. 33.33% red, 33.33% green and 33.33% blue by taking average the image does not occurs properly in grayscale image. To overcome this problem, we use Weighted Method.

Weighted Method

To solve the problem facing in average method, we use another method i.e. weighted method or luminosity method. In this we use an equation to convert the color image to grayscale image.

The equation is

$$X= (0.3* R) + (0.59*G) + (0.11*B)$$

Here,

X is a new Grayscale image
R is Red Color
G is Green Color
B is Blue Color

In the above equation, R is take as 30%, G is taken as 59% and B is taken as 11%. Using this equation, the converted color images to grayscale images are shown in Fig.2:

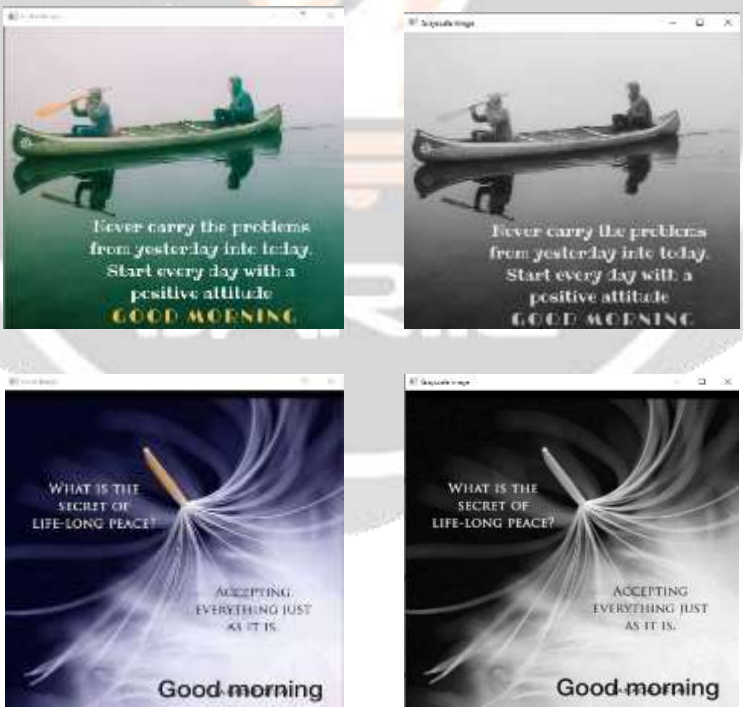

Fig.2 : Color Image to Grayscale Image

Step 2 Now in this step we use canny edge detection. In this method we follow these steps for edge detection

1. Gaussian Blur operator
2. Intensity Gradiant Calculation
3. Finding the edge directions

4. Non-maximum suppression
5. Thresholding
6. Edge Tracking
7. Final Cleaning

- Gaussian Blur Operator is used to remove the noise from the grayscale image. This noise removed image shall enable further smooth processing and flawless.
- Intensity Gradiant Calculation, in finding calculation of gradiant sobel edge detection is used in this process. Sobel filter has two kernel (3x3 matrix) one of them correspond the X (horizontal) and another shall be used for the Y(vertical) direction. These two kernels shall be convoluted with the original image under process and through which the edge points are calculated with ease. The kernel values shown below are fixed for sobel filter and cannot be altered.

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Gradient Approximation can be calculated with

$$G = \sqrt{Gx^2 + Gy^2}$$

Then G will be compared against the threshold and with which, one can understand the taken point is an edge or not. The formula for finding edge direction is

$$\Theta = \tan^{-1}(Gy/Gx)$$

- Non-Maximum suppression is the next step. The Gradiant magnitude operator obtains thick edges but the final images is expecetd to have thin edges. So non-maximum suppression process thick edges to thin edges through edge direction already available.There can be only 4 possibe directions fro any pixels. They are 0 degrees,45 degrees, 90 degrees and 127 degrees.
- Thresholding is the next step after non-maximum suppression. Even after non maximum in the image there may be still some noise, some of the edges shown may not real and some edges could miss. To overcome this problem we need to go with double thresholding. In this process we set two thresholds one is high and other is low. If a value is above then high the it is to be seen as stronger edge and if value is below the low then it is not an edge.
- Edge Tracking – In this we remove the weak edges which are not connected to stronger edge.
- In the final cleaning all the weak edges can be removed from it.

The below Fig.3, show the edge detected image from grayscale image

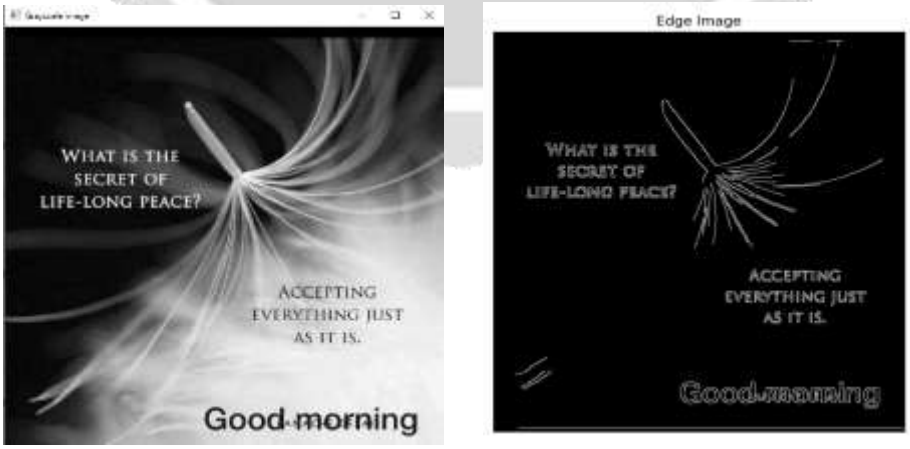

Fig. 3: Grayscale image to Edge detection

Step 3: After removing noise from the grayscale image using canny edge based detection, next it is converted to binary image. A binary image is a digital image that can have only two possible values for each pixel. Each pixel is stored with a single bit either 0 or 1. To form a binary image we select a threshold intensity value. If intensity value is greater than the threshold value that pixel is changed to 0(black) and if lesser than the threshold value that pixel is changed to 1(white). Thus the image is changed to binary image it is shown in the Fig.4.

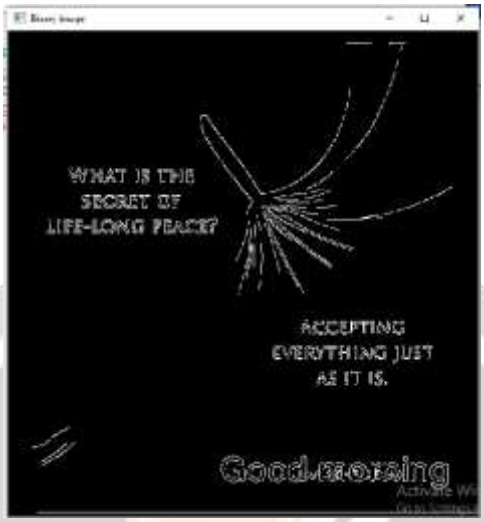

Fig. 4: Binary Image

Step 4: Using OCR the text is identified from binary image it is show in Fig.5.

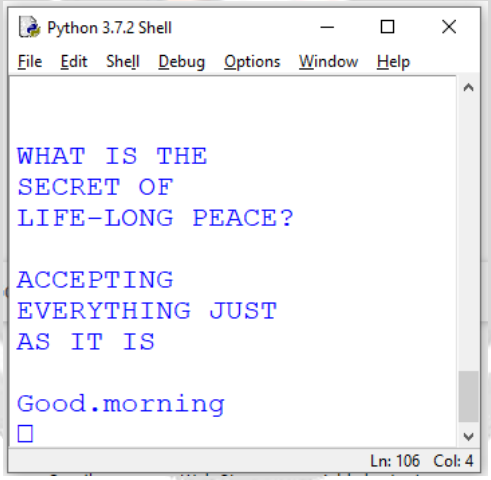

Fig.5: Extracted text from Binary Image

4. CONCLUSIONS

Now a day, so many different methods are implemented for extracting text on image. In this paper, the proposed method will give an idea and procedure to extract the text form image with simple four steps. Using equations first step converting color image to gray scale image. Next step using canny edge detection method edges are identified and noise is reduced from image. Next step, the edges identified gray scale image is converted into binary image. Last step, from the binary image using OCR text is extracted.

5. ACKNOWLEDGEMENT

We are grateful to God to give us the opportunity and capability to complete this research. Then we remember our parents whose are always thinking our successive. While doing this work we faced so many problems and to solve the problem we got help from various sources. We would like to express our sincere regard all the teachers of CSE Discipline, Chaitanya Deemed to Be University

6. REFERENCES

[1] K. H. Talukder and T. Mallick, "Connected component based approach for text extraction from color image," *2014 17th Int. Conf. Comput. Inf. Technol. ICCIT 2014*, pp. 204–209, 2003.

[2] V. V. Rampurkar, S. K. Shah, G. J. Chhajed, and S. K. Biswash, "An approach towards text detection from complex images using morphological techniques," *Proc. 2nd Int. Conf. Inven. Syst. Control. ICISC 2018*, no. Icisc, pp. 969–973, 2018.

[3] S. Grover, K. Arora, and S. K. Mitra, "Text extraction from document images using edge information," *Proc. INDICON 2009 - An IEEE India Counc. Conf.*, no. 5409409, pp. 23–26, 2009.

[4] X. Liu, "MULTISCALE EDGE-BASED TEXT EXTRACTION FROM COMPLEX IMAGES Xiaoqing Liu and Jagath Samarabandu The University of Western Ontario Department of Electrical & Computer Engineering," pp. 1721–1724, 2006.

[5] C. Saravanan, "Color image to grayscale image conversion," *2010 2nd Int. Conf. Comput. Eng. Appl. ICCEA 2010*, vol. 2, no. January, pp. 196–199, 2010.

[6] G. Noel, K. Djouani, and Y. Hamam, "Edge detection in grayscale images using grid smoothing," *Commun. Comput. Inf. Sci.*, vol. 150 CCIS, no. PART 1, pp. 110–119, 2011.

[7] A. Coates *et al.*, "Text detection and character recognition in scene images with unsupervised feature learning," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 440–445, 2011.