

## **TEXT EXTRACTION FROM IMAGE AND DISPLAYING RELATED INFORMATION**

**<sup>1</sup>B.Malakonda Reddy,**

<sup>1</sup>Associate Professor, Department of ECE, Narayana Engineering College, Gudur, AP, 524101  
Email: [1bmalakondareddy@gmail.com](mailto:1bmalakondareddy@gmail.com),

**<sup>2</sup>P. Divya, <sup>2</sup>K. Keethe, <sup>2</sup>P. Bhargavi, <sup>2</sup>R. Harinadh**

<sup>2</sup>UG Student, Department of ECE, Narayana Engineering College, Gudur, AP,  
524101 [2divyaparuchuri1999@gmail.com](mailto:2divyaparuchuri1999@gmail.com)

**Abstract:** Image text is the image information embedded or written in a image of different forms image text can be found in capture images scan documents this image text is highly available. Now a days, and they are very important in the presenting and describing and transferring. And which helps in people communication solving problems cultural gap availability except for this information from which help people in communication solving problems are highly efficiency and easy of Access. If it is converted into text from the process by which the image converted into plain text is it test extraction. Different techniques such as connected component method mathematical morphological method edge detection-based method and test based method have been used previously in order to enhance performance and accuracy of text extraction is discussed and implementation is done integrated MATLAB code. Different techniques such as connected component mathematical morphological method edge detection based method and test based method have been used previously in order to enhance performance and accuracy of text extraction is discussed and implementation is done integrated MATLAB code.

**Keywords:** Connected Component Method, Mathematical Morphological Method, Edge Based Method, Test Based Method.

### **1. INTRODUCTION**

Image text is a text information embedded or written in images of different forms. Image text can be found in captured images, scan documents, magazines, newspapers, posters etc. This image text or a highly available now a days and they are very important in representing, describing, and exchanging information which helps people in communication solving problems, availability, creation of new types of job, cost-effectiveness. productive globalization and cultural gastric etc. This information from these image document would give higher efficiency and easy of Access if it is converted to text. The process by which image text converted into plain text that that compute can recognize it's a characters is a text extraction. The information from image document should be converted into text in order to get effective use and access of it like at reporting that are used in different image based application such as office work many methods have been given by different researchers and profession answer for text detection and extraction. Previously according to the methods used they are categorized into different methods such as connected component method document paper that need to digitalize and used for achieving, indexing, Information regain process or increasingly common today for example scanned documents. office work, magazines, advertisements, webpage and effective extraction of text from these

documents is a challenging problem the text extraction needs knowledge of text properties that is supposed to be extracted so first we should study the properties of text or characteristics can be categorized as

- 1) Colour:** Intensity of colour also affect the quality of text detection if all characters of same colour then extraction process become simpler and more effective.
- 2) Motion:** This property of text usually applied to the videos containing text, and this refers to the movement of text in vertical and horizontal direction.
- 3) Edge:** Images for reliable features of expect are compared with other features like colour layout on motion strong edges at the boundaries of text and background make the text detection.
- 4) Geometry:** There is a problem in size alignment Inter-Character distance.
- 5) Compression:** Compressed format digital files can be transferred and communicated in different media does text extraction systems. That extract text without the compression can be carried categorized as faster and better system variation of these text due to the difference in size, orientation style and alignment text is embedded in complex colour document images.

Degraded images and degraded document images low quality images as well as low image contrast. Complex background make problem in extraction of text from images documents highly difficult. Limitation of work is proposed to be done by MATLAB code integrated with simlink support package and it will be tested with different types of images text in the proposed. System of different types will be used first the original image is taken as a input for pre-processing. which the image is converted into the day colour noise and non-text objects of the image removed then image binarization enhancement text detection and extraction will be done by proposed algorithm. And passed to optical character recognize enjoyment for character recognize finally extracted and recognize text will be displayed and read by text.

## 2. LITERATURE SURVEY

C. P. Chaithanya, N. Manohar, Ajay Bazil Issac survey on " automatic text extraction and classification from natural images". Text detection is a method of locating areas in a picture. where ever text is present text detection and classification in natural pictures is very important for several computer version applications like optical character recognize distinguish between human and machine inputs and spam removal currently the challenge in text identifying is to detect the texted innaturally.

Pictures due to many factors like low quality images and unclear words, typing font image having a lot of colour stock then the background colour blood pictures due to do some natural problems like rain, Sunny, snow accept the main aim of the work is to identify and classify the text in natural pictures here the system detect the text and find the connected regions change them together in the related position uses the text classification in engine to filter change with low classification confidence scores.

Harpreet singh, Deepindersingh used mathematical morphological for extraction of image text using resulting improved performances and low noise but not detecting small text from Complex background the paper States the future work to be extracting small text and converting into editable form.

Sarathi Giri compare the two basic approaches for extracting text regions in images edge based and connected component based using a set of images that varies along the dimensions of lighting and orientation proposed future work to design the verifying extraction text regions by SVM and HMM and then to design recognized text for extracting text regions.

Nitisyal, Naresh Kumar Garg the paper based on integration of Daubechines DWT, SVM Gradient SVM differences and resulting extract text regions clearly future work proposed implementation of OCR system to recognize the text use better method is known text removal.

## 3. PROPOSED METHODOLOGY

The proposed algorithm is combination of two algorithm gift based algorithm and connected component algorithm discussed above in order to get higher performance text extraction using the combination of two organism has an advantage not for the extraction also for recognition process it will give most enhanced input for OCR (optical character recognition) which convert the image text to plain text that computer can recognize.

Finally, the output will text will be enhanced to use the architecture of the proposed algorithm and on the combination of powerful algorithms is used; Edge Based Methods and Connected Component Method. These algorithms performance can be analysed and evaluated by precision rate, recall rate, f-score and accuracy. The result of the measures differs based on the input image type means scale or size difference, lighting variance, and orientation or direction variances. We have used 410 images dataset from existing resources that other researchers used. The dataset is selected from Document Image Binarization Competition) Both edge based and connected component methods. which text from images is given the lost extracted text result of the proposed algorithm suitable threshold in order to simplify text extraction process general representation of pre-processing step of text extraction system this is a original image of book cover captured by camera grayscale image used for further process of extraction and it is binary image that computer can recognize this process is very important step in extracting text from images since RGB image may have noise and they could not identify the text on non-textual objects of the image.

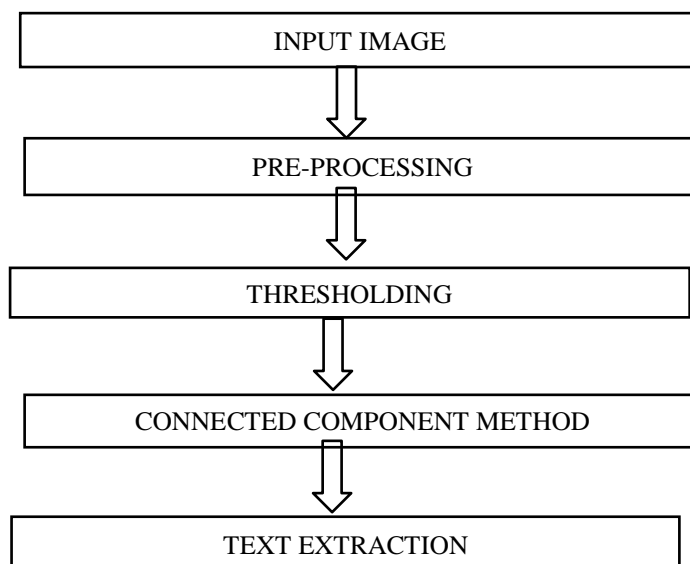


Fig 1. Proposed Algorithm

### **1) Pre-Processing Stage**

Pre-processing state eliminates challenges by creating by noise blurring effect uneven lightning which makes performance text detection extraction and liberalization. That are embedded in image document simple and better in this stage the image input is process to remove any noise that may affect the image during the time of acquisition or during the time of Transmission. A coloured RGB image will be converted to gray scale image that are sold at evening and noise removal process will be done. This image is converted to the binary image with suitable threshold in order to simplified extraction process. Generally, the representation of pre-processing steps of text extraction system the original image RGB of book cover captured by camera. Gray scale image used to further process of extraction and it is binding image the computer can organizer this process is very important procedure extracting text from images. Since, RGB image we have noise and they could not identify text and non-textual objects of the image.

### **2) Thresholding**

Image thresholding is a simple at effective way of partitioning and image into foreground and background. This image analysis technique is a type of image segmentation does isolates object by converting grayscale image into the binary image binary thresholding is most effective image with high level of contrast.

#### **Otsu method**

Otsu thresholding method involve iterating through all possible threshold values and calculating a measure of speed for pixel values. Each side of the threshold that means pixel that area fall in foreground and background the aim is to find the threshold value with the sum of the foreground and background spreads is at its minimum.

### **3) Connected Component Method**

Component based method major small components into success fully lodge components until they all the regions of the image identified. It uses geometric analysis merge these text components which alters our non-textual components and set the boundaries of text contained regions representing. The architecture of connected components text extraction algorithm we have used some text images which are very in properties and offer process by the above connected algorithm.

### **4) Text Detection**

In this process the connected component based approach is used to make that possible. text region standing out as compared to non-text regions every pixel in the edge image is assigned a weight with respect to its neighbour in each Direction. This weight value is maximum value between the pixel and its neighbour pixel in left and upper and upper right directions the algorithm uses these three labour values to detect the edges in horizontal vertical and diagonal directions the resultant its image obtaining is sharpened in order to increase the contact between the detected edges and its background making it easier to text extract regions.

### **5) Text Extraction**

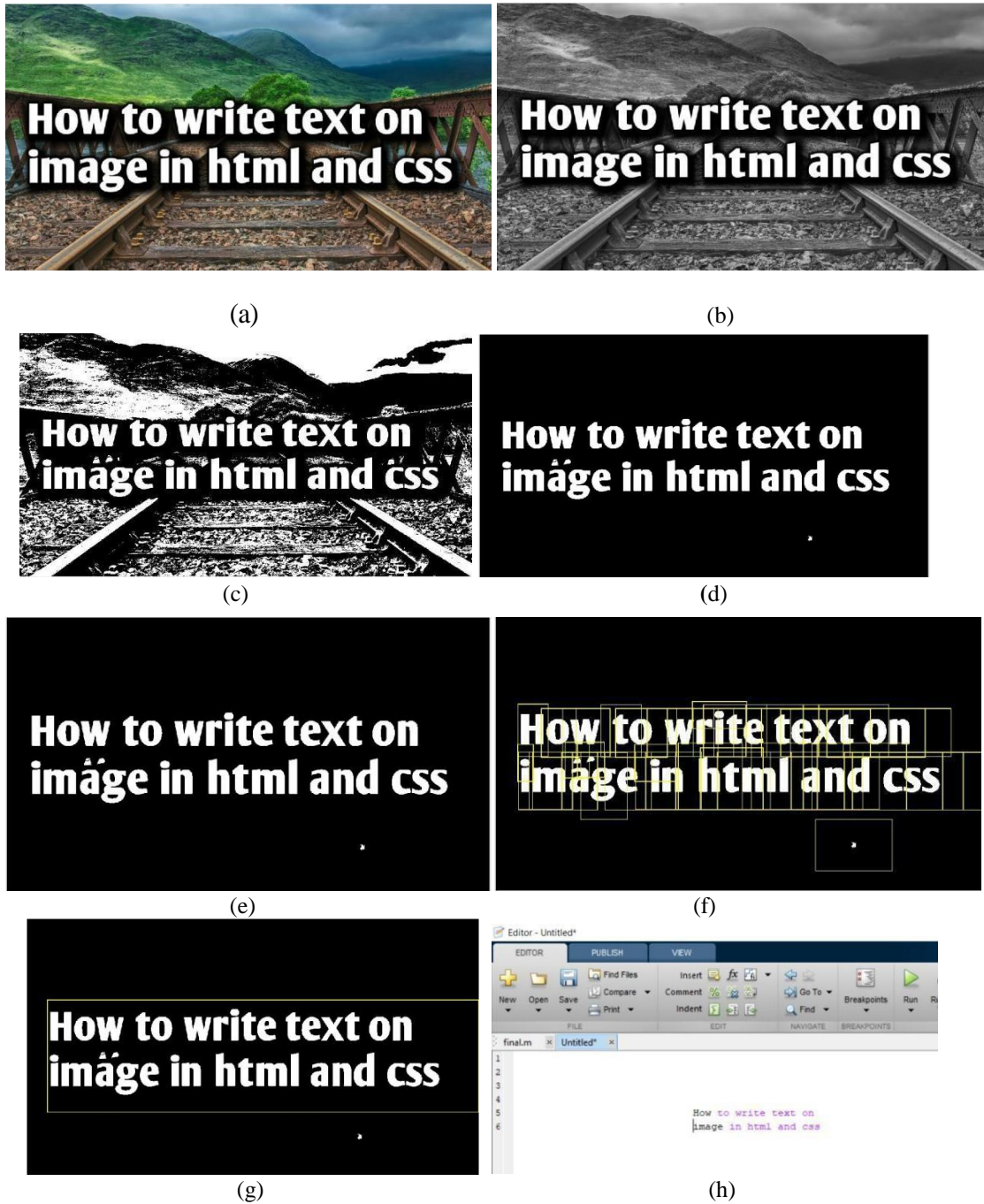
Text extraction and enhancement for GRE generating the input to an OCR method.

#### **Optical Character Recognition**

OCR is a mechanical electronic translation of image of handwritten, typewritten printed text into mission editable text widely used as a form of data entry for printed paper data records. whether the passport documentation invoice blank statements computerized receipts business cards made printouts of static data or any suitable documentation it is a column method of digitalize printed text. So, that we can be electrically searched stored more compactly displayed on line and used in machine process such as cognitive computing, machine translation Text-to-Speech data and text minimization OCR is a field of research in pattern recognition, Artificial Intelligence and computer vision earlier version needed to be with images of each character and worked on one front of the time at once and Systems capable of producing the high degree of recognition activity for most of wants for now, and with the support for varieties of digital images file format in some systems are

capable of producing formatted output and closely approximates the original page including images columns and their non textual components.

#### 4. RESULTS AND DISCUSSION



**Fig 2. Recognition of Text :** a) Original image b) Gray image c) Binary image d) Remove non-textual image e) Bounding Box f) Connected component image g) Detected text, h) Output image

## **5. CONCLUSION**

Extraction of text from image documents is very important in different areas nowadays. In this we proposed the algorithm which gives good performance in text extraction by combining two algorithms, Edge Based and Connected Components. The extracted text recognition done by OCR with better accuracy and finally audio output produced. The paper does not include handwritten and complex font text which can be future work.

## **REFERENCES**

1. Reza Sarshogh and Keegan Hines proposed “Computer vision extraction text from images” International Journal of Advanced Research in Computer Science(IJARCS), 2019, pp.795-810
2. Holly Rogers a journal on “Transmitting thresholds in the work of natural images”, IEEE, International Conference on Computing, Analytic and Security Trends (CAST), January 2018, pp.138-143 .
3. K.N. Natei "Extracting Text from Image Document and Displaying Its Related Information "International Journal of Engineering Research and Applications (IJERA), vol. 8, no.5, 2018, pp.27-33
4. Ch.Md Mizan, T. Chakraborty and S. Karmakar, “Text Recognition using Image Processing”, International Journal of Advanced Research in Computer Science(IJARCS), 2017, pp.765-768
5. Shivani, D. Bansal, “Techniques of Text Detection and Recognition: A Survey”, International Journal of Emerging Research in Management &Technology (IJERMT), 2017, pp. 83-87.
6. R. Islam, Md. Rafiqul Islam, Kamrul Hasan Talukder “An Approach To Extract Text Regions from Scene Image”, IEEE, International Conference on Computing, Analytic and Security Trends (CAST), 2016, pp.138-143.
7. Panchal, Sh. Varde, Dr. Prof. MS. Panse, “Comparative study of Image processing techniques used for Scene text detection and extraction”, International Journal of Engineering Research and General Science (IJERGS), 2016, pp. 183-188.