K.N. Natei Journal of Engineering Research and Application www.ijera.com
ISSN : 2248-9622, Vol. 8, Issue5 (Part -V) May 2018, pp 27-33

RESEARCH ARTICLE OPEN ACCESS

Extracting Text from Image Document and Displaying Its Related Information

K.N. Natei*, J. Viradiya**, S. Sasikumar ***
**(Dept. Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, India*
*** ([3]Dept. Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, India*
*** ([3]Dept. Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, India*
Corresponding Auther: K.N. Natei

ABSTRACT
Image Text is the text information embedded or written in image of different form. Image text can be found in captured images, scanned documents, magazines, newspapers, posters etc. These image texts are highly available nowadays and they are very important in representing, describing and transferring information which help peoples in communication, solving problems, availability, creation of new types of jobs, cost effectiveness, productivity, globalization and cultural gap etc. The information from these image documents would give higher efficiency and ease of access if it is converted to text form. The process by which Image Text converted into plain text is Text Extraction. Text Extraction is useful in information retrieving, searching, editing, documenting, archiving or reporting of image text. However, variation of these texts due to differences in size, orientation style, and alignment, text is embedded in complex colored document images, degraded documents image, low quality image, as well as low image contrast and complex background make problem text extraction extremely difficult and challenging one. Different techniques such as Connected Component Method, Mathematical Morphology Method, Edged Based Method and Texture Based Method have been used previously, but those all have their own limitations when measured by different parameters like precision, recall and f-score. In this paper, text extraction from image documents, using combination of the two powerful methods Connected Component and Edge Based Method, in order to enhance performance and accuracy of text extraction is discussed and implementation is done by integrated MATLAB code with MATLAB/Simulink tool and the proposed system is tested by Digital Image Binarization Competition (DIBCO) 2017 dataset. Finally, the extracted and recognized is converted to speech for proper use for visually impaired people.

Keywords - Connected Component, Edge Based, Text Extraction, Image Text

---

Date of Submission: 14-05-2018 Date of acceptance: 29-05-2018

I. INTRODUCTION

Image Text is the text information embedded or written in image of different form. Image text can be found in captured images, scanned documents, magazines, newspapers, posters etc. These image texts are highly available nowadays and they are very important in representing, describing and transferring information which help peoples in communication, solving problems, availability, creation of new types of jobs, cost effectiveness, productivity, globalization and cultural gap etc. The information from these image documents would give higher efficiency and ease of access if it is converted to text form. The process by which Image Text

converted into plain text that computer can recognize its ASCII character is Text Extraction [1, 2]. The information from image documents should be converted into text in order to get efficient use and access of it like archiving or reporting that are used in different image based applications such as office works.

Many methods have been given by different researchers and professionals for text detection and extraction previously. According to the methods used; they are categorized into different methods, such as; Connected Component Based Method,

Edge Based Methods, Region Based Method, Texture Based Method and Mathematical Morphology Based Method [4]. These all methods and techniques used in image text extraction have their own advantages and limitations based on different parameters such as precision rate, recall rate, accuracy etc [3].

Document papers that need to be digitized and used for archiving or indexing or information retrieval process are increasingly common today, for example scanned documents of office works, in magazines, advertisements and web pages. Robust and efficient extraction of text from these documents is a challenging problem [1, 3] due to different properties of text in image.

Text extraction has a vast number of applications that are related with computer vision, image processing, computer based works all. Some basic applications are listed below:

- Text searches in Images - Currently, it is difficult to search image using text and deliver intended image content. Text extraction would enable better searching by extracting the content of an image.
- Content based Indexing - For the purpose of archiving and indexing documents, the content of the document is required in the digital format. Knowledge about the text content of documents can help in the building of an intelligent system which archives and indexes the printed documents.
- Wearable or portable computers: With the rapid development and advancement of computer hardware technology, wearable computers are now on a boom.
- License container plate recognition: There is still a development needed in the phase of vehicle license plate and container plate recognition.
- Texts in WWW images: The extraction of text from WWW images can provide relevant information on the Internet.
- Reading foreign language text - One of the common problems faced by a person in foreign land is that of communication, understanding road signs, signboards etc.
- Archiving documents - Archives of paper documents in offices or other printed material like magazines and newspapers can be electronically converted for more efficient storage and instant delivery to home or office computers.

Text extraction needs knowledge of text properties that is supposed to be extracted. So first we should study the properties of text. Text properties or characteristics can be categorized as follows [3]:

1) Geometry:

- Size: there is a problem on text extraction because of differences in text size, but it can be minimized by assuming specific data during text region detection process.
- Alignment: Text on image may be aligned in vertical, horizontal, or in different shape alignments that may bring difficulty while recognizing. Thus it can be solved by transformation algorithms on character recognition.
- Inter-character distance: characters in a text line have a uniform distance between them, thus space maybe eliminated.

2) Color: intensity of color also affects the quality of text extraction. If all characters of same color then extraction process become more simple and effective.

3) Motion: this property of text usually applied to the videos containing text, and this refers to the movement of text in vertical or horizontal direction.

4) Edge: edges are reliable feature of text as compared with the other features like color layout or motion, strong edges at the boundaries of text and background makes the text detection easy.

5) Compression: compressed format digital files can be transferred and communicated in different Medias. Thus, Text Extraction system that extract text without decompression can be categorized as faster and best system.

Variation of these texts due to differences in size, orientation style, and alignment, text is embedded in complex colored document images, degraded documents image, low quality image, as well as low image contrast and complex background make problem in extraction of text from image documents highly difficult.

The proposed method, aims to alleviate these problems by reading the text information from the image documents which are captured by a camera, scanner or any other devices using combination of Connected Component and Edged Based Methods which can extract and recognize a text from a complex image and as well as efficient and robust by improving existing performance.

The implementation of the work is proposed to be done by MATLAB code integrated with Simulink Support Package and it will be tested with different types of image texts. In the proposed system different steps will be used. First, the original image is taken as input for preprocess in which the image is converted to gray color, 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 Recognition (OCR) engine for character recognition.

Finally, extracted and recognized text will be displayed and read by text to speech (tts) tool [10].

II. RELATED WORK

Many researchers have done their work on extraction of image text and retrieving the information though there are many challenges. These researches are based on different image text detection and extracting techniques which have their own advantages as well as limitations. Review of these literatures is given by the following table in summarized way.

Chowdhury Md. Mizan, Tridib Chakraborty and Suparna Karmakar [9], proposed algorithm to recognize printed hard copy and convert to required format text using OCR (Optical Character Recognizing) and Image Processing techniques. The algorithm recognizes the character offline, is efficient to extract bimodal images and is applicable in retrieval of image, video, web page text etc. They suggested that future researchers have to done on OCR area.

Akhilesh A. Panchal, Shrugal Varde, M.S. Panse [10] , used combination of two approaches, Connected component and Region Based to provide access of computer vision technology for visually impaired people by extracting and converting image text into speech in approved accuracy and speed. Combination of approach techniques results faster and better system. The system is not checked with complex image and text of small size and varied alignment. The authors suggest combination of techniques for improvement of accuracy and speed.

Najwa-Maria Chidiac, Pascal Damien, Charles Yaacoub [11], used MSER (Maximally Stable Extremal Regions) and Stroke Width Detectors to detect and extract text from natural scene regardless of orientation with improved accuracy on blurred and noisy image. But the proposed system couldn't detect image having text with small size or thin width and shadow effect.

Jack Greenhalgh and Majid Mirmehdi [15] created novel system for detection and recognition of text in traffic signs automatically using MSER (Maximally Stable Extremal Regions) and HSV (Hue-Saturation-Value). The result improved accuracy of recognition F-measure of 87%. The image on sign must be capture when it is larger size.

Rashedul Islam, Md. Rafiqul Islam,Kamrul Hasan Talukder[12], proposed hybrid techniques (Edge-Based and Connected Component Based) that enables to increase accuracy of text area detection and extraction techniques by combining. In this algorithm accuracy of the extraction system is improved (87.25%). They test using only 08 images to evaluate, but not considered degraded images and small size text, not checked by OCR to recognize characters. Future work is to create database for sake of training.

Arvind, Mohamed Rafi [2] used connected component method to maximize detection and extraction of text from image and categorizing it. They tried to improve performance in precision (65.06%) and recall rate (89.25%), and presented the results in graph.

Vaishnav Ganesh, Dr. L. G. Malik [8] analyzed Big Data by Google Apis and proposed the framework for Big Data Image used Color based partition method and Text line grouping method using Canny edge detector and Hough transform methods respectively. By applying trained classifiers either time or efficiency will be improved.

Harpreet Singh, Deepinder Singh [14] used mathematical morphology for Extraction of image text using resulting improved performance 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.

Partha Sarathi Giri [13] compared two basic approaches for extracting text region in images: edge-based and connected-component based using a set of images that vary along the dimensions of lighting, scale and orientation. Proposed future work to design the verifying extraction text region by SVM and HMM, and then to design recognizer system for extraction text regions.

Niti Syal, Naresh Kumar Garg [7] the paper is based on integration of Daubechies DWT, Gradient Difference and SVM, resulting extract text region effectively. Future work proposed is implementation of OCR system to recognize the text, use better method in non-text removal.

III. METHODOLOGY

Various methods and tools will be used in order to meet the general and specific objectives of this research work, these are listed as follows:

1. Literature Review

For better understanding of text extraction from image, numerous literatures on different aspects of this dissertation work have been studied previously. Basic concepts, applications methods or techniques as well as algorithms used in extraction of text in image documents developed and analyzed by different experts and researchers should be discussed. Study on characteristics of text that may have impact on text extraction has been conducted. Review of different methods of text extraction from image conducted. Connected Component Based, Edge Based Method, Region Based Method, Mathematical Morphology Based Method, and digital Image Processing are most used examples of those techniques [4, 16].

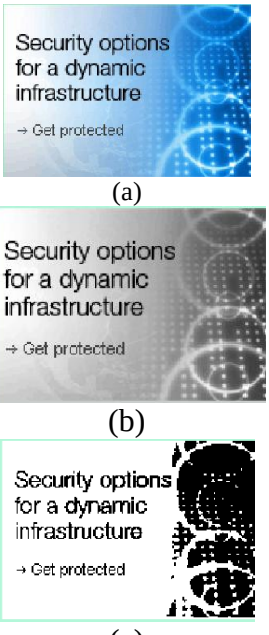
2. Methods

The whole work of this dissertation is categorized into three major parts or stages as discussed below [5]:

- Pre-processing Stage
- Processing Stage
- Post Processing Stage

A. Pre-processing Stage

The Pre-processing stage eliminates challenges created by noise, blurring effect and uneven lighting which makes performing text detection, extraction and recognition that are embedded in image documents simple and better. In this stage, the image input (scanned or captured by different devices or browsed from drive) is processed to remove any noise that may affect the image during the time of acquisition or during the time of transmission. A colored (RGB) image will be converted to a grayscale image and then thresholding, edge thinning, and noise removal process will be done. The image is then converted to a binary image with suitable threshold in order to simplify extraction process [17].



(a) (b) (c)

Figure 1. Pre-processing stage representation: a) Original Image [17]
b) Grayscale Image c) Binary Image

The above Figure 1, illustrates the general representation of preprocessing step of text extraction system. The figure 1 (a) is original image (RGB) of book cover captured by camera, Figure 1 (b) grayscale image used for further process of extraction and figure 1 (c) is binary image that computer can recognize. This process is very important step in extracting text from image, since RGB image may have noise, and could not identify text and non-text objects of the image.

B. Processing Stage

Processing stage have different steps in which the image is checked whether it contains text or not, identify to locate the text area on image, and differentiating foreground and background of that image text is done. The following steps are most common:

- Text Detection: takes enhanced image as input and decides it contains text or not and identifies the text regions in an image.
- Text Localization: merges the text regions to formulate the text objects and define the tight bounds around the text objects. Figure 2 shows Architecture of Processing stage. Text detection, localization and tracking modules are closely related to each [5].
- Text Tracking: is also used to speed up the text extraction process by not applying the Binarization and recognition step to every detected object.
- Text Binarization: used to segment the text object from the background in the bounded text objects. It converts grayscale image to binary image, where text pixels and background pixels appear in two different binary levels like white text on dark background or vice versa. Binarization can also be done before the other steps.

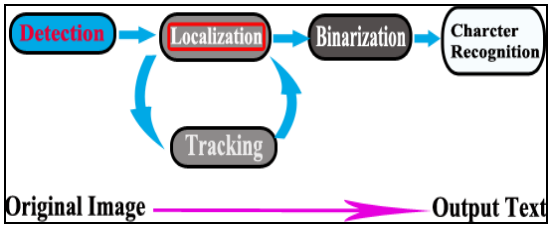


Figure 2. Processing stage of text extraction [5]

The Figure 3 below shows the localized text area of the original image on Figure 1 above done in processing stage. The area that recognized as text is bounded by rectangle boxes. This is done by localization step, that its output can be used as input for next step of segmenting text and non-text objects that simplifies character recognition process.

Figure 3. Text Area Localization

III. POST PROCESSING STAGE

This stage contains the further processes that would be done on extracted text such as segmentation, character recognition and text to speech.

- Segmentation: is concerned with dividing an image into meaningful regions. Image segmentation can be broadly classified into two types [6]. For character segmentation, first the image must be segmented row-wise (line segmentation), then each row must be segmented column-wise (word segmentation).
- Character Recognition: The last stage is the character recognition. This module converts the binary text object into the ASCII text using OCR tool. Figure 4. shows how 'PARUL UNIVERSITY' word wrapped in image gets recognized through Text Extraction process.
- Text to Speech (tts): this process gives audio or speech output of character or word recognized by OCR.

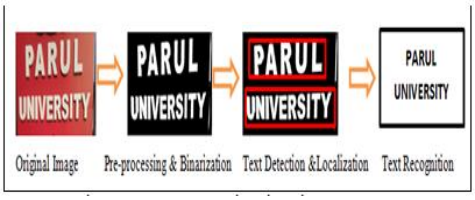


Figure 4. Post processing that shows output

IV. RESULTS AND DISCUSSION

The proposed system have been done on text extraction from image techniques tested with different types of image input by combining the two algorithms, Connected Components and Edge Based Methods by improving their limitations by measuring with some parameters such as precision, recall rate and accuracy based on variety of images input (image size, quality or alignment) used for test. The analysis on these algorithms is discussed as below.

1. Edge Based Algorithm

Edge-based algorithm is one of the popular Methods for text extraction which focus on the high contrast between the text of image and its background. This method identifies and merges the edges of the text boundary while processing. The figure 5 below can represent architecture of Edge Based text extraction algorithm. The input image fed to processing engine which includes sub processes like text region detection, text localization and text extraction. Finally, monochrome output image which is white for ground and black background.

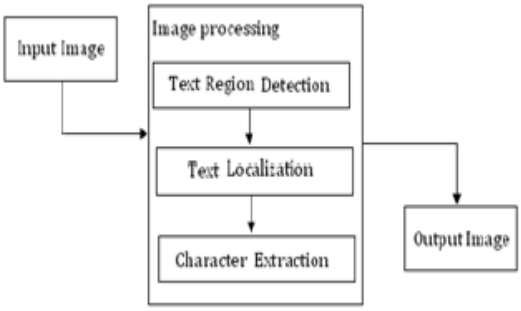


Figure 5. Block Diagram of Edge Based Algorithm [13].

After testing edge based algorithm with input images having different properties, figure 6 (a) clear image but with large size and figure 6(c) a roughly degraded and with small size. Thus, the result of these input images is given as figure 6(b) and figure 6 (d) respectively.

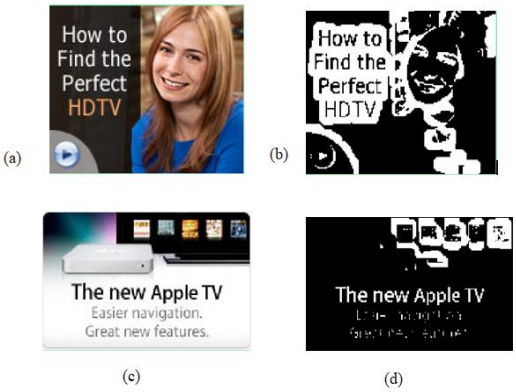


Figure 6. Results of Edge Based Algorithm: (a) and (c) are original input images [17], (b) and (d) output

2. Connected Component Algorithm

Connected Component based method merges small components into successively large components until all the region of image is identified. It uses geometrical analysis to merge these text components which alters out non-text components and set the boundaries of text contained regions [13]. The figure 7 below can represent the architecture of connected components text extraction algorithm.

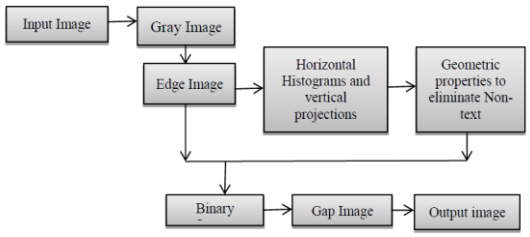


Figure 7. Basic Block Diagram of Connected Component Method Algorithm [2]

We have used some test images which are vary in properties and after processed by the above connected components algorithm, different output are obtained. The figure 8 (a) and (c) are two of input images, figure 8 (b) and (d) are respective output images of the algorithm.

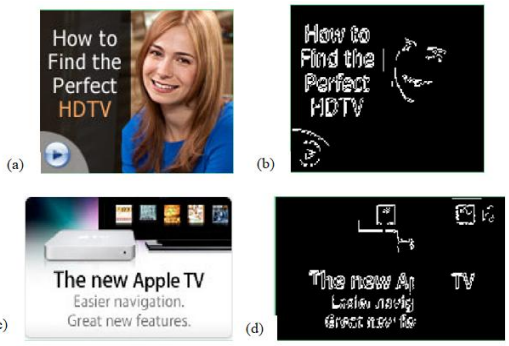


Figure 8. Extraction process of Connected Component Algorithm, (a) and (c) are original images, (b) and (d) respective output

3. The proposed Algorithm

The proposed algorithm is combination of the two algorithms, Edged Based Algorithm and Connected Components Algorithm discussed above in order to get higher performance text extraction. Using the combination of the two algorithms has an advantage not for text extraction also for recognition process it will give most enhanced input for OCR (Optical Character Recognition) which convert image text to plain text that computer can recognize. Finally, the output text will be enabled to be read. The architecture of the proposed algorithm and whole text from image extraction system is given in figure 9 below. The figure 10 show the last extracted text result of the proposed algorithm.

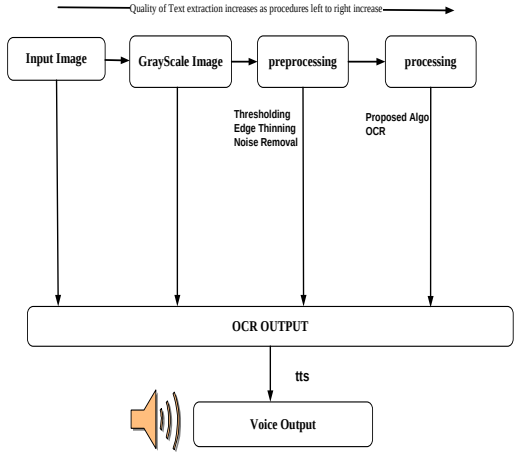


Figure 9. Architecture of the proposed system.

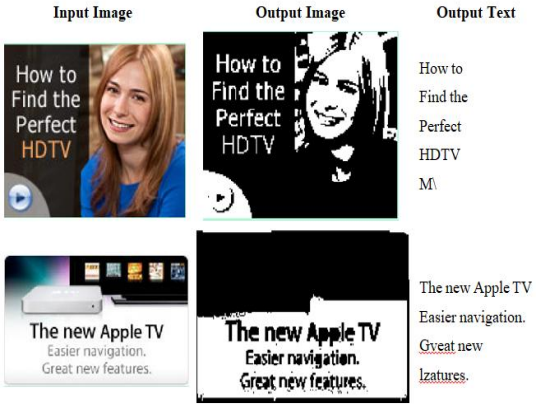


Figure 10. The last result of proposed algorithm

4. Performance Measure

In order to get better performance of text extraction the combination of powerful algorithms is used; Edge Based Methods and Connected Component Method [5]. These algorithms performance can be analyzed 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 DIBCO 2017 (ICDAR 2017 Document Image Binarization COmpetition) [19].

Both edge based and connected component methods have their own strengths as well as weaknesses. In the proposed system these limitations would be decreased and the performance of text extraction from image documents. The table 1 below shows the overall result of the two algorithms Edge based and Connected Component using different images with respective output comparing with proposed one.

Table 1. The comparison of all algorithms results done using dataset images

Measures	Edge Based	Connected Components	Proposed
Precision	80.78%	88.05%	96.11%
Recall Rate	86.87%	86.91%	93.58%
F-Score	83.71%	87.48%	94.83%
Accuracy	72.24%	76.55%	90.42%

V. CONCLUSION AND FUTURE SCOPE

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]. S. Grover, K. Arora, S. K. Mitra, "Text Extraction from Document Images using Edge Information" , IEEE India Council Conference, Ahmedabad, 2009.
[2]. Arvind, M. Rafi, "Text Extraction from Images Using Connected Component Method" JoAIRA, STM Journal, 2014, 13-18.
[3]. D. Gera, N. Jain, "Comparison of Text Extraction Techniques- A Review", International Journal of Innovative Research in Computer and Communication Engineering, 2015, pp. 621-626.
[4]. Y. Gupta, Sh. Sharma, T. Bedwal, "Text Extraction Techniques", International Journal of Computer Application, NSFTICE, 2015, pp. 10-12
[5]. A. Panchal, Sh. Varde, Dr.Prof.M.S.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.
[6]. Shivani, D. Bansal, "Techniques of Text Detection and Recognition: A Survey", International Journal of Emerging Research in Management &Technology (IJERMT), 2017, pp. 83-87.
[7]. N. Syal, N. K. Garg, "Text Extraction in Images Using DWT, Gradient Method And SVM Classifier", 2014, pp.477-481
[8]. V. Ganesh, Dr. L. G. Malik, "Extraction of Text from Images of Big Data" International Journal of Advance Research in Computer Science and Management Studies, IJARCSMS, 2014, pp.40-46.
[9]. 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.
[10]. A. A. Panchal, Sh. Varde, M.S. Panse, "Character Detection and Recognition System for Visually Impaired People", IEEE, International Conference On Recent Trends In Electronics Information Communication Technology, 2016, pp.1492-1496.
[11]. Najwa-Maria Chidiac, P. Damien, Ch.Yaacoub, "A Robust Algorithm for Text Extraction from Images", IEEE, 2o16, pp.493-497.
[12]. R. Islam, Md. Rafiqul Islam,Kamrul Hasan Talukder "An Approach To Extract Text Regions from Scene Image", IEEE, International Conference on Computing, Analytics and Security Trends (CAST), 2016, pp.138-143.
[13]. P. S. Giri, "Text Information Extraction and Analysis from Images Using Digital Image Processing Techniques," International Journal on Advanced Computer Theory and Engineering (IJACTE), 2013, pp.66-71.
[14]. H. Singh, D. Singh, "Text Confining and Extraction in Image Using Mathematical Morphology," International Journal of Science and Research (IJSR), 2012, pp.288-290.
[15]. J. Greenhalgh and M. Mirmehdi, "Recognizing Text-Based Traffic Signs," IEEE Transactions On Intelligent Transportation Systems, 2015, pp.1360-1369.
[16]. R. Hedjam, R. F. Moghaddam and M. Cheriet, "TEXT EXTRACTION FROM DEGRADED DOCUMENT IMAGES", April 2014, https://www.researchgate.net/publication/224214156_Text_extraction_from_degraded_document_images , August 20, 2017.
[17]. MathWorks, Inc., "Image Processing Toolbox™User'sGuide", http://in.mathworks.com, September 24, 2017.
[18]. MathWorks, Inc., "Getting Started with Android™Devices", http://in.mathworks.com/help/supportpkg/android/examples/getting-started-with-android-devices.html , October 02, 2017.
[19]. DIBCO 2017, "ICDAR 2017 Document ImageBinarizationCOmpetition", https://vc.ee.duth.gr/dibco2017/ ,March 25, 2018
[20]. Agam Kumar Tyagi. MATLAB and SIMULINK for Engineers; 2nd edition, Oxford University Press, New Delhi, 2012, pp 1, 65.

Authors Profile

Mr. K. N. Natei pursed Bachelor of Science in Computer Science from Ambo University, Ethiopia in 2014 and was working as Graduate Assistant II and Assistant Lecturer in Department of Computer Science at Wachemo University, Ethiopia from 2014 to 2016. He is currently pursuing M.Tech in Computer Science and Engineering Department from Parul University, India. He has published 1 survey paper on international conference prepared by Parul University (PiCET-2018) and doing his graduation research paper on the area of Image Processing. He has 2 years of teaching experience in different cources of Computer Science.

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