International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-4, Issue-3, Mar- 2017]*
https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

Text Detection in Document Images: Highlight on using FAST algorithm

Geetika Mathur[1], Ms. Suneetha Rikhari[2]

[1]Student, Department of E.C.E., College of Engineering and Technology, Mody University, Lakshmangarh, Sikar, India
[2]Assistant Professor, Department of E.C.E., College of Engineering and Technology, Mody University, Lakshmangarh, Sikar, India

Abstract*—In recent years, text extraction from document images is one of the most widely studied topics in Image Analysis and Optical Character Recognition. These extractions of document images can be used for document analysis, content analysis, document retrieval and many more. Many complex text extracting processes Maximization Likelihood (ML), Edge point detection, Corner point detection etc. are used to extract text documents from images. In this article, the corner point approach was used. To extract document from images we used a very simple approach based on FAST algorithm. Firstly, we divided the image into blocks and their density in each block was checked. The denser blocks were labeled as text blocks and the less dense were the image region or noise. Then we check the connectivity of the blocks to group the blocks so that the text part can be isolated from the image. This method is very fast and versatile, it can be used to detect various languages, handwriting and even images with a lot of noise and blur. Even though it is a very simple program the precision of this method is closer or higher than 90%. In conclusion, this method helps in more accurate and less complex detection of text from document images.*

Keywords—Corner point,FAST(Features from Accelerated Segment Test), OCR,multilingual documents, handwritten documents.

I. INTRODUCTION

In recent years, the trend to digitalize documents has emerged. With digitalization of the world the paper based documents need to be converted into digital to make them handier, searchable and for preserving of the documents. Optical Character recognition is used for this process. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text [2]. For over a half century research in this area is ongoing and character recognition rate in modern OCR is above 99% on a high-quality document and 90% on handwritten documents. For degraded documents and books the efficiency of OCR comes down to 80%. In recent times, many techniques have been used for text extraction in document images. Here we will use a very simple approach based on FAST point's algorithm. Firstly, we divide the document image into smaller non-overlapping blocks of a fixed size. We then check the density in each block using FAST corner detection technique. The denser blocks were labeled as text blocks and the less dense were the image region or noise region. Then we check the connectivity of the blocks to group the blocks so that the text part can be isolated from the image. We then build the text region and save it.

This method is very fast and versatile, it can be used to detect various languages, handwriting and even images with a lot of noise and blur. Even though it is a very simple program the precision of this method is closer or higher than 90%. In conclusion, this method helps in more accurate and less complex detection of text from document images.

II. OPTICAL CHARCTER RECOGNITION

The development of character recognition in last decade is remarkable and the method for character detection is vast. The advancements of Character Recognition are evident in Optical Character Recognition (OCR), Document Classification, Computer Vision, Data Mining, Shape Recognition, and Biometric Authentication [2]. Character recognition is the process to classify the input character per the predefined character class [1]. Character recognition has its application in identification of text in images. The text maybe a scanned document or a handwritten text.

A. Text from Images:

In recent years, the trend to digitalize documents has emerged. With digitalization of the world the paper based documents need to be converted into digital for more handy,

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

searchable and preserving of the documents. Optical Character recognition is used for this process. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text [2]. For over a half century research in this area is ongoing and character recognition rate in modern OCR is above 99% on a high-quality document and 90% on handwritten documents. For degraded documents and books the efficiency of OCR comes down to 80%. In recent times, many organizations depend on OCR for better performance and more efficiency. OCR can be performed offline and/or online. Online recognition the OCR processor recognizes the character as they are given. In offline method, the processor may recognize both document as well as handwritten characters but recognition in offline mode highly depends on the quality of the scanned images.[10]

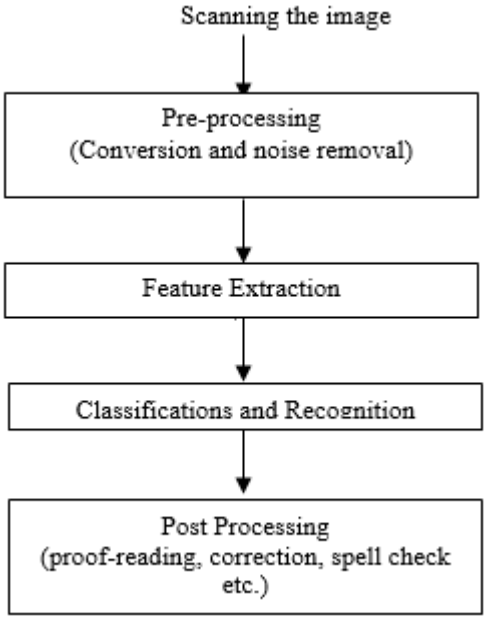

Fig.1: Stages for OCR

OCR consists of many phases such as Scanning of image, Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition, Post Processing. The task of preprocessing relates to the removal of noise and variation in the image [3]. In scanning step the image is acquired. The quality of image depends highly on the scanner being used. In practical applications, the scanned images are not perfect there may be some noise due to some unnecessary details in the image which can cause a disruption in the detection of the characters in the image. Preprocessing involves removal of noise (applying filters like Gaussian filter, Gabor filter etc.) and proper conversion of image like a colored image can be converted into gray scale or binary image for further processing of image. Feature extraction involves recognizing the feature required. Classifications and Recognition phase is the extraction phase of the process. After finishing the OCR process several postprocessing steps are necessary depending on the application, e.g. tagging the documents with meta-data (author, year, etc.) or proof-reading the documents for correcting OCR errors and spelling mistakes [4].

OCR is still in research and much advancement need to be made in this technology. The future scope of this is OCR in mobile devices, handwriting recognition, recognition of various languages except English (like Arabic, Devanagari, Telugu text), extraction and processing of images from video, processing and restoration of old documents and many more.

B. Document Images:

A document image contains various information such as texts, pictures and graphics [5]. These images are obtained by scanning handwritten documents, old documents, printed documents, journals etc. Many challenges are faces for recognizing scanned documents like low contrast, low resolution, color bleeding, complex background and unknown text color, size, position, orientation, layout etc. Even if the OCR system is of supreme quality the system can still not give proper output due to the problems discussed above. Generally, the process of OCR works best if the background of the image is clean and the image is free of any noise.[6]

C. Extraction from Document Images:

Many techniques have been used for text extraction in document images. In this article, we will use a very simple approach based on FAST point's algorithm. Firstly, we divide the document image into smaller non-overlapping blocks of a fixed size. We then check the density in each block using FAST corner detection technique. The denser blocks were labeled as text blocks and the less dense were the image region or noise region. Then we check the connectivity of the blocks to group the blocks so that the text part can be isolated from the image. We then build the text region and save it.

This method is very fast and versatile, it can be used to detect various languages, handwriting and even images with a lot of noise and blur. Even though it is a very simple program the precision of this method is closer or higher

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

than 90%. In conclusion, this method helps in more accurate and less complex detection of text from document images.

III. COMPONENTS OF AN OCR SYSTEM

A distinctive OCR system consists of various components for OCR systems. OCR consists of many phases such as Scanning of image, Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition, Post Processing. The task of preprocessing relates to the removal of noise and variation in the image [3]. In scanning step the image is attained and the image is digitalized. The quality of image depends highly on the scanner being used. In practical applications, the scanned images are not perfect there may be some noise due to some unnecessary details in the image which can cause a disruption in the detection of the characters in the image. Preprocessing involves removal of noise (applying filters like Gaussian filter, Gabor filter etc.) and proper conversion of image like a colored image can be converted into gray scale or binary image for further processing of image. Feature extraction involves recognizing the feature required. Classifications and Recognition phase is the extraction phase of the process. After finishing the OCR process several postprocessing steps are necessary depending on the application, e.g. tagging the documents with secondary data like author, year, etc.or proof-reading the documents for correcting OCR errors and spelling mistakes [4].

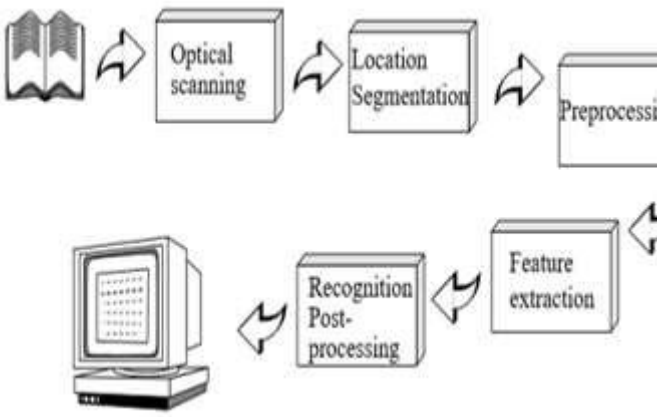

Fig.2: Components of an OCR system [7]

1. Optical Scanning:

In the scanning process the digital image of the document is captured. A scanner is used to scan the documents. The quality of the document depends highly on the scanner being used. So, a scanner with high speed and good color quality is necessary for proper processing of the image.

2. Location and segmentation:

This process locates the places where contents are present. The process that determines the constituents of an image is segmentation. It is essential to locate the regions of the document that have data printed and distinguish them from noise and pictures. For example, during automatic mail-sorting, the address is located and separated from other constituents of the envelope like stamps or logos, before recognition process.

Segmentation is the separation of characters or words from image which is performed on text. Most optical character recognition systems segment the words into isolated characters which are documented individually. This technique is easy to device, but problems occurs if the characters' touch or if characters are disjointed and consist of several parts. The main problems in segmentation may be divided into four groups:

1. Extraction of touching and disjointed characters.
2. Distinguishing noise from text. Dots and accents may be mistaken for noise, and vice versa.
3. Mistaking graphics or geometry for text. This leads to nontext being sent to recognition.
4. Mistaking text for graphics or geometry. In this case the text will not be passed to the recognition stage. This often happens if characters are connected to graphics [7].

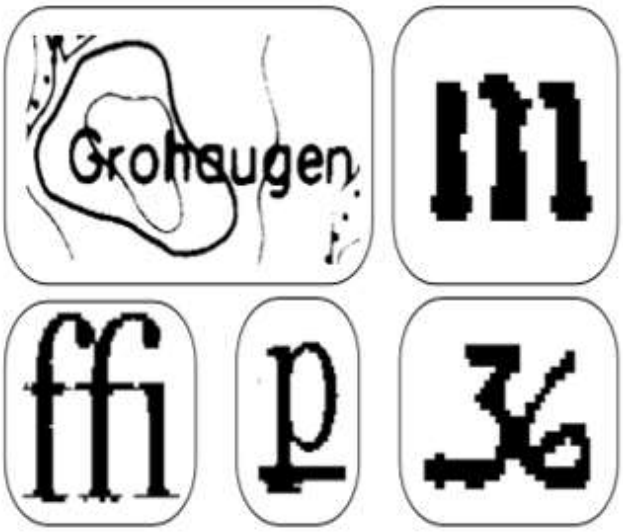

Fig.3: Example of Degraded symbols[7]

3. Pre-Processing:

The image is scanned and is converted into gray scale. The gray scale image maybe converted to binary image. This process is called Digitization of image (Binarization). In practical applications, a scanner is not perfect; the image

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

that is scanned may have some noise. This may be due to some redundant details present in the image. The denoised image is produced by applying some appropriate methods. This denoised image is saved for further processing [2].
Depending on the resolution on the scanner and the success of the applied technique for thresholding, the characters may not be perfectly scanned.

4. Feature extraction:

The pre-processed image serves as the input to this and each single character in the image is found out [2]. The image from the extraction stage is matched with all the preloaded characters in the system. Once the matching is completed, the template with the maximum correlated value is declared as the character present in the image. [1]
The objective of feature extraction is to detect the essential characteristics of the characters, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The best way of describing a character is by the actual image. The techniques for extraction of such features are often divided into three main groups, where the features are found from:

- The distribution of the points.
- Transformations and series expansions.
- Structural analysis.[14]

5. Post Processing :

After feature extraction stage, there might be some unrecognized characters, those characters may get defined in the post-processing step. [2] Character grouping to make a meaningful text and error detection and correction is done in this step.

IV. PROPOSED WORK

In the proposed approach to extract document from images we used a very simple FAST algorithm. Firstly, we divided the image into blocks and their density in each block was checked. The denser blocks were labeled as text blocks and the less dense were the image region or noise. Then we check the connectivity of the blocks to group the blocks so that the text part can be isolated from the image.This method is very fast and versatile, it can be used to detect various languages, handwriting and even images with a lot of noise and blur. Even though it is a very simple program the precision of this method is closer or higher than 80%. In conclusion, this method helps in more accurate and less complex detection of text from document images.
The flowchart in figure 4 shows the steps involved in the proposed approach. The details of the steps are given below:

Step 1: The image is scanned and is converted into gray scale. The gray scale image maybe converted to binary image. This process is called Digitization of image (Binarization) The noise is due to the scanner. In the project we have used Gaussian filter .Gaussian filtering is used to blur images , remove noise and remove unwanted details in the image.[12][13]

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Step 2: The corner points are determined by FAST algorithm[9]
Step 3: Divide the image in non-overlapping blocks and calculate the number of corner points.
Step 4: From the block find the block which has the maximum number of corner points (Nmax), define a threshold using the selected block, threshold used as T=0.2*Nmax. (20% of maximum value)
Step 5: Divide the blocks having more number of corners than the threshold belong to text regions, and blocks having less threshold belong to image or background region.
Step 6: After detecting text blocks from corner point, check for connectivity of these blocks (8-connected regions) to rebuild text regions. [15]

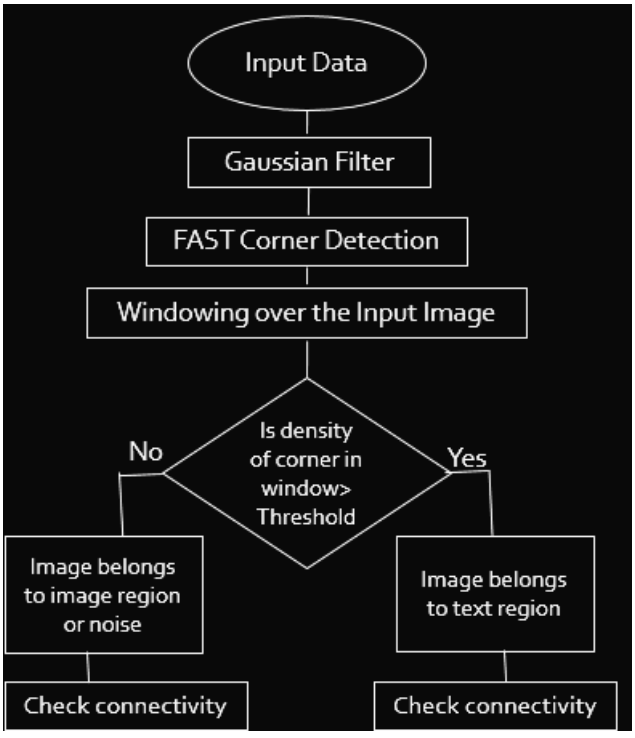

Fig.4: Flow chart for the algorithm[8]

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

V. EXPERIMENTAL RESULT

This is a simple method with precision and recall are over 90% and often with 95% on an average. However, this technique is not very effective for big size fonts as well as for some specific pictures for which corners are responding too much. Despite of these problems it is fast (and can be parallelized) and less complex as compared to other OCR tools and could be further improved in the future. Finally, this method seems also to be very efficient in extracting more complex layouts such as paragraphs, and lines.

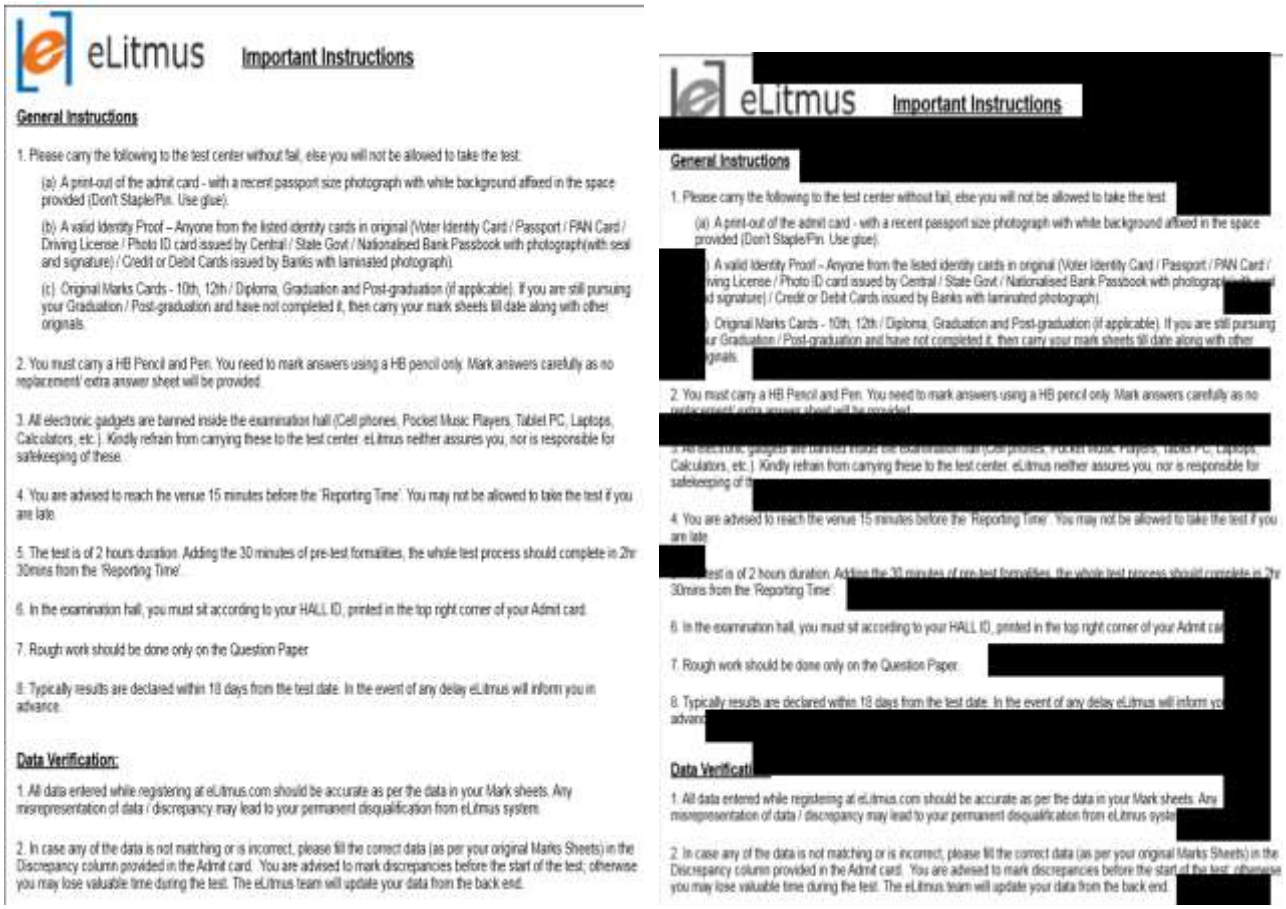

Fig.5: English document image (Original image, Detection of text image)

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

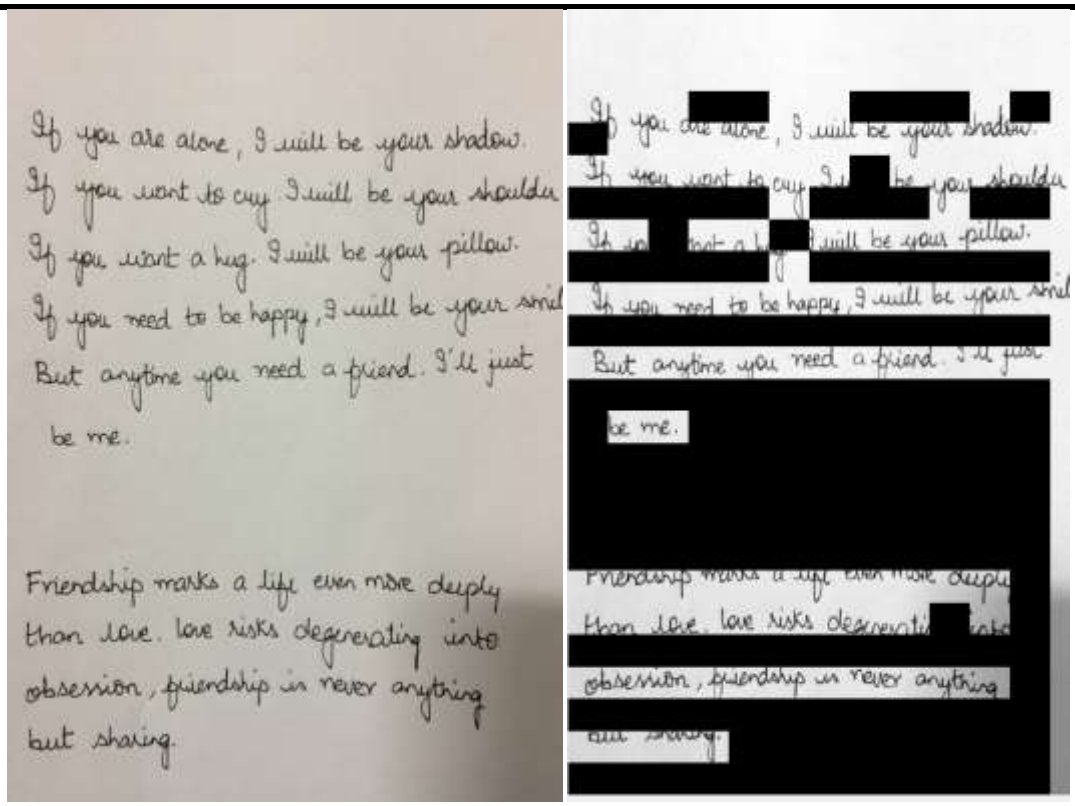

Fig.6: English handwritten image (Original image, Detection of text image)

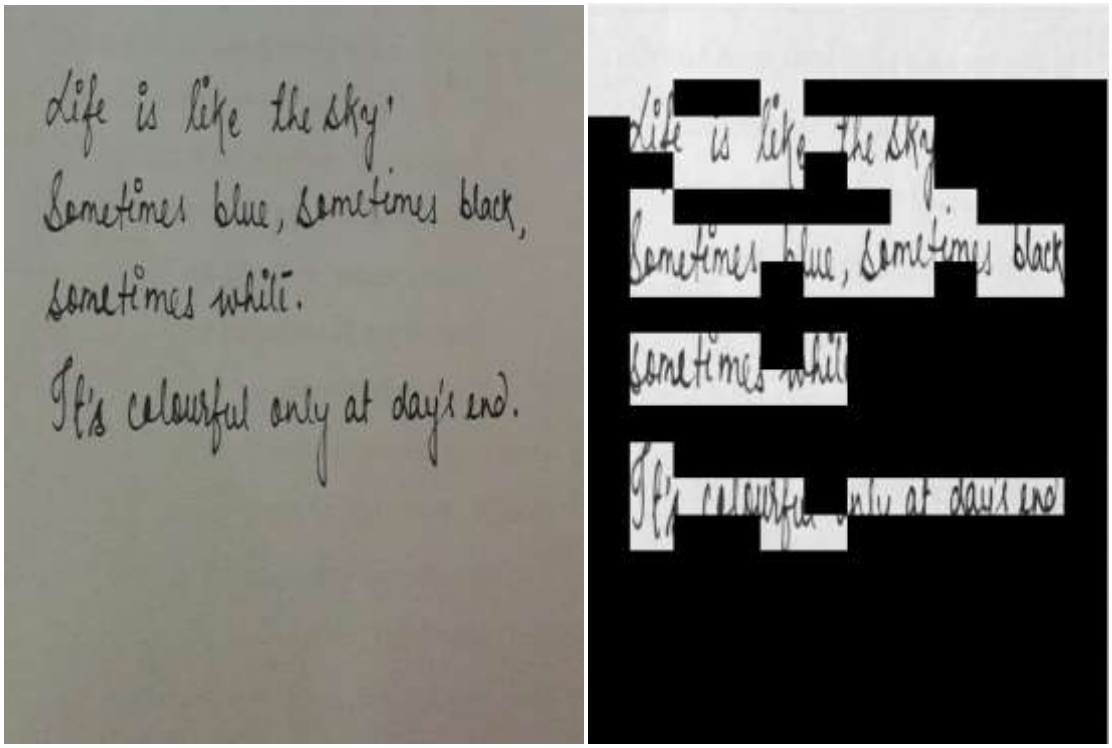

Fig.7: English handwritten image (Original image, Detection of text image)

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

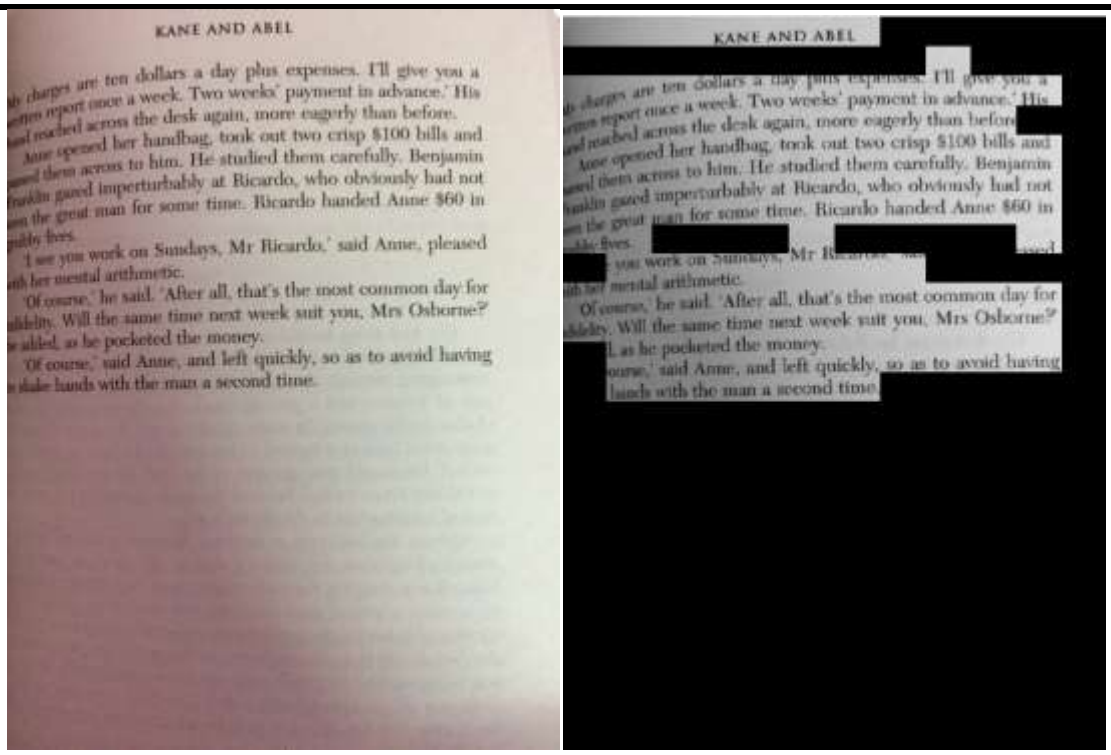

Fig.8: Skewed Document Image (Original image, Detection of text image)

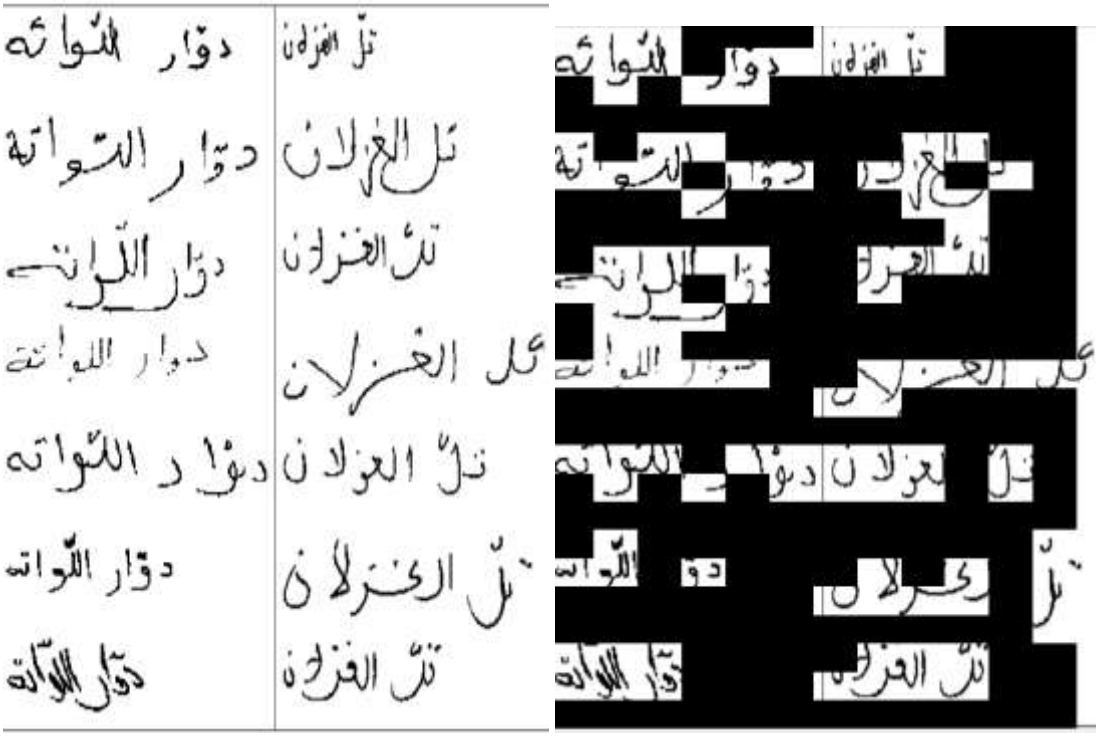

Fig.9: Non- English text images-Arabic language (Original image, Detection of text image)

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

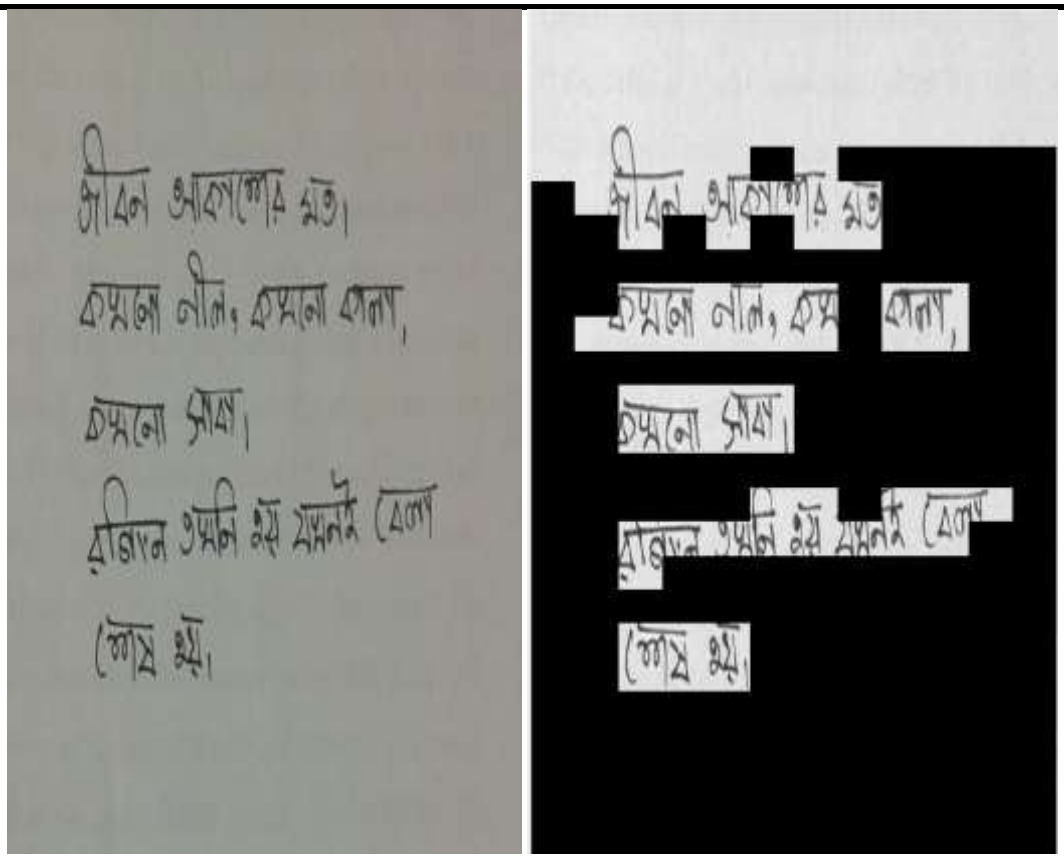

Fig.10: Non- English text images-Assamese language (Original image, Detection of text image)

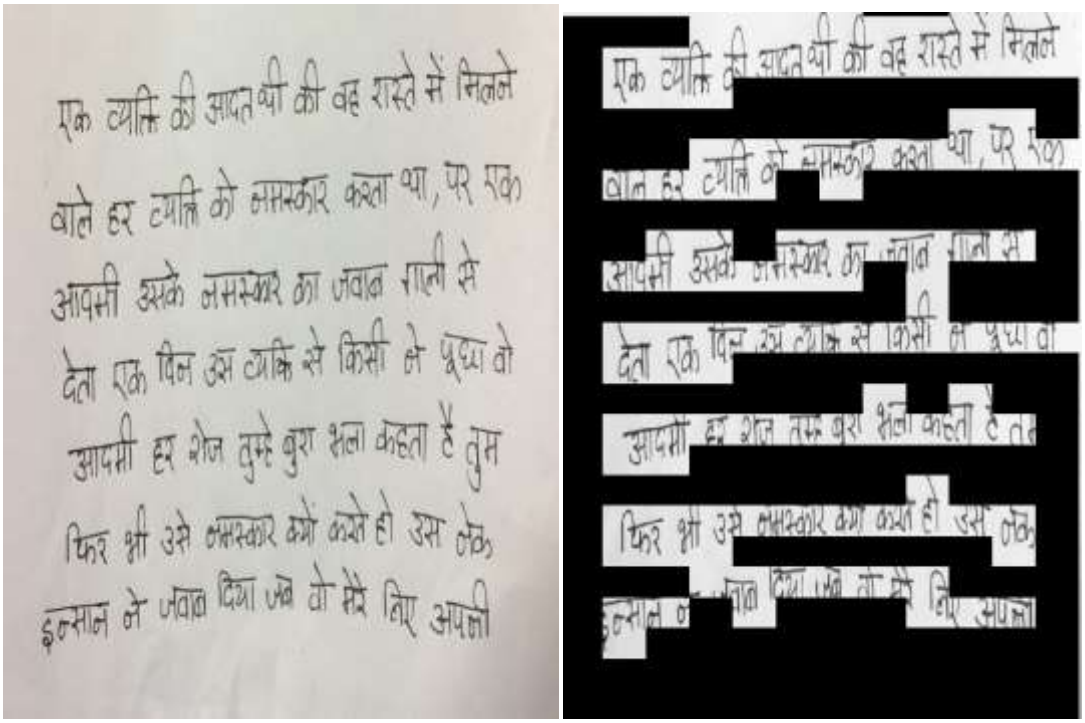

Fig.11: Non- English text images-Hindi language (Original image, Detection of text image)

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

VI. CONCLUSION AND FUTURE SCOPE

In this approach, we saw that via corner points on document images of any quality, orientation or handwritten, it could be very simple to obtain an accurate text extraction at low cost and without using a lot about parameters. To extract text from images we use a very simple approach based on FAST algorithm. Firstly, we divided the image into blocks and their density in each block was checked. The denser blocks were labeled as text blocks and the less dense were the image region or noise. Then we check the connectivity of the blocks to group the blocks so that the text part can be isolated from the image.

This method is very fast, less complex and versatile, it can be used to detect various languages, handwriting and even images with a lot of noise and blur. Even though it is a very simple program the precision of this method is closer or higher than 80%. Results show that with this simple method, precision is over 80% (most often around 85% in average). But, this technique fails for big fonts and for some specific pictures for which corners are responding too much. Despite of these problems it is fast (and can be parallelized) and less complex as compared to other OCR tools and could be further improved in the future. Finally, this method seems also to be very efficient in extracting more complex layouts such as paragraphs, and lines.

Future Scope:

***Font Independent OCR*:**

Development of OCR considering the multiple font style needs to be developed in the future. The corner point approach is very much useful for the font independent OCR, because, for font or character size, it finds the block and the blocks are analyzed to recognize the character.

OCR for all Indian Languages:

Development of OCR for languages other than English needs to be reserched on and developed in the future. The corner point approach is very much useful for the OCR of languages other than English, because, for font or character size, it finds the block and the blocks are analyzed to recognize the character. This further proves to be an efficient way to detect handwritten languages.

Cursive Characters OCR:

There is heavy demand for an OCR system which recognizes handwritten cursive scripts. This avoids keyboard typing and font coding for the image. This method helps in detecting handwritten characters with a precision of about 90%.

Language Converter through OCR:

Once we detect languages we can can develop a converter to convert sentences from one language to another through a conversion and translation scheme.

Speech recognition from OCR:

Speech recognition is one of the most important application today. The recognized Printed or Handwritten OCR could be recorded and speech output could be generated. This would help the blind to send and receive information.

Speech to text converter through OCR:

Speech recognition is one of the most important application today. The recognized speech could be recorded and output of text could be generated.

REFERENCES

[1] Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal ,"Survey of Methods for Character Recognition", *International Journal of Engineering and Innovative Technology (IJEIT),* Volume 1, Issue 5, May 2012, ISSN: 2277-3754.

[2] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar,"Optical Character Recognition", *International Journal of Advanced Research in Computer and Communication Engineering* ,Vol. 3, Issue 1, January 2014,pp. 4956-4958, ISSN (Online) : 2278-1021,ISSN (Print): 2319-5940.

[3] Sarika Pansare, Dhanshree Joshi," A Survey on Optical Character Recognition Techniques", *International Journal of Science and Research (IJSR),* Volume 3 Issue 12, December 2014, pp. 1247-1249, ISSN (Online): 2319-7064.

[4] Sukhpreet Singh," Optical Character Recognition Techniques: A Survey", *Journal of Emerging Trends in Computing and Information Sciences,* Vol. 4, No. 6 June 2013, pp. 545-550, ISSN 2079-8407.

[5] Deepika Ghai, Neelu Jain , " Text Extraction from Document Images- A Review", *International Journal of Computer Applications (0975 – 8887)* , Volume 84 – No 3, December 2013 , pp. 40- 48.

[6] Keechul Junga, Kwang In Kim, Anil K. Jain, "Text information extraction in images and video: a survey", Pattern Recognition, 37, pp. 977-997, 2004.

[7] Line Eikvil," Optical Character Recognition", Norsk Regnesentral, P.B. 114 Blindern, N-0314, December 1993.

https://dx.doi.org/10.22161/ijaers.4.3.43 *ISSN: 2349-6495(P) | 2456-1908(O)*

[8] Vikas Yadav, Nicolas Ragot," TEXT EXTRACTION IN DOCUMENT IMAGES: HIGHLIGHT ON USING CORNER POINTS", in *2016 12th IAPR Workshop on Document Analysis Systems*, pp. 281-286.

[9] Viswanathan, Deepak Geetha. "Features from Accelerated Segment Test (FAST)." (2009), pp. 1-5.

[10] Nauman Saleem, Hassam Muazzam, H.M.Tahir , Umar Farooq ," AUTOMATIC LICENSE PLATE RECOGNITION USING EXTRACTED FEATURES" in *4th International Symposium on Computational and Business Intelligence*,September 5-7, 2016, Olten, Switzerland, pp. 221-225.

[11] Mr. Rohit Verma, Dr. Jahid Ali," A Comparative Study of Various Types of Image Noise and Efficient Noise Removal Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 10, October 2013, ISSN: 2277 128X,pp. 617-622.

[12] Yao Wang," Image Filtering: Noise Removal, Sharpening, Deblurring", *EE 3414 Multimedia Communication Systems*, Polytechnic University, Brooklyn, NY11201.,

[13] Ajay Kumar Boyat and Brijendra Kumar Joshi," A Review Paper: Noise Models In Digital Image Processing*", Signal & Image Processing : An International Journal (SIPIJ)* , Vol.6, No.2, April 2015, pp. 63- 75.

[14] Q. Yuan, C. L. Tan," Text Extraction from Gray Scale Document Images Using Edge Information" , Washington, Sept. 10-13 (2001) , pp. 302-306.

[15] http://www.imageprocessingplace.com/downloads_V3/root_downloads/tutorials/contour_tracing_Abeer_George_Ghuneim/connectivity.html