ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

RESEARCH ON TEXT IDENTIFICATION, EXTRACTION AND RECOGNITION FROM NATURAL IMAGES

[1]Anjali Sharma, [2]Priyanka Rawat
[1,2] Department of Computer Science and Engineering,
[1,2] Lingaya's Vidyapeeth, Faridabad, India
[1]sanjali.gaur@gmail.com, [2]Priyankarawat3@yahoo.com

Abstract — Text extraction is that the way toward separating content from an image. The most work of an OCR is to frame editable papers from existing report papers or picture documents. Huge number of calculations is required to build up a Test extraction and essentially it works in two stages like character and word acknowledgment. Just if there should be an occurrence of a progressively modern methodology, an OCR likewise chips away at sentence location to safeguard a report's structure. Test extraction frameworks have set up a particular section place in design acknowledgment. OCR is furthermore well known among Android applications. Format is one among the premier open source libraries wont to actualize OCR in Android applications. It's discovered that in order to wash the picture record, the precision of the product is as high as 97.56%. It ought to be noticed that the exactness is estimated in light of the fact that the level of right characters and words. The existing system/the previous system of OCR on a grid infrastructure is just OCR without grid functionality. That is the existing system deals with the homogeneous character recognition without space or character recognition of single file with low Accuracy. Existing system convert only signal character detect. The benefit of proposed system that overcomes the drawback of the existing system is that it supports multiple functionalities such as character recognition with space and high accuracy and multi character convert in image to text. It also adds benefit by providing characters recognition. An accuracy of 99% means that 1 out of 100 characters is uncertain. While an accuracy of 99.9% means that 1 out of 1000 characters is uncertain. Measuring OCR accuracy is done by taking the output of an OCR run for an image and comparing it to the original version of the same text.

Index Terms — pre-processing, segmentation, feature extraction, pattern recognition

I. INTRODUCTION

The expanding headways inside the electronic data recovery and example acknowledgment has prompted the examination inside the advancement of ongoing applications which aren't just testing yet in addition with high computational capabilities [1 2 3]. One among the machine as of late affecting the instructional exercise explore and in this manner the business is that the Optical Character Recognition for example OCR. The OCR method has gotten one among the significant

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

part inside the scanners which are used in applications like language recognizable proof, identification verification, and so on it is regularly been an alluring idea of getting a machine which may supplant human functionality[4 5 6]. Be that as it may, when it includes design acknowledgment, a machine isn't as shrewd as human to just perceive the character or content, particularly from the photos. Here OCR comes into picture. There are fundamentally two sorts of OCR depending on the archive, regardless of whether it's written by hand or print[7 8]. First is that the Offline OCR which manages the prevalence of the content from the picture when filtered. What's more, second is that the Online OCR which perceives the character while it's being composed through detecting the development of the pen. It's generally utilized for transcribed content.

In the past two decades, researchers have proposed numerous methods for detecting texts in natural images or videos. There are mainly three types of methods:

1. Edge based Method: Edges are a solid capacity of the content power, game plan, direction, and so tense based strategy doesn't make a difference shading/focused on a high differentiation between the content and thusly the foundation [9 10 15 16]. is implanted, the three distinctive highlights of content in pictures for identifying content edges are quality, thickness and along these lines the direction fluctuation are utilized. Edge-based content extraction calculation might be a universally useful technique that can find rapidly and viably and separate the content from the two records and indoor/outside pictures. This technique isn't strong for enormous content taking care of it.

2. Texture based Method: This strategy uses the way that the content in the picture discrete surface properties that recognize them from the foundation [11 12 13 14]. Procedures Fast Fourier Transform (FFT) to the spatial difference, et al SVM classifier bolsters Gabor channels, wavelet are used the printed idea of the content can be seen inside the picture region. [17] This methord can perceive the content in complex foundation [18]. The main downside of this strategy is that the enormous computational exertion in grouping stage surface.

3. Morphological based Method: scientific morphology, a procedure upheld topology and geometry for picture investigation to be morphological element extraction procedure has been applied adequately for character acknowledgment and report investigation [19, 20]. It is utilized to extricate the basic qualities content difference of the picture is handled. These highlights are invariant to changes in geometric figures, for example, interpretation, revolution and scaling is likewise changed by the condition of the blaze or content shading, nor the qualities can generally be kept up. This technique utilizes a strong under various picture changes.

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

II. LITERATURE SURVEY

Table 1. shows performance of various approaches in text detection and extraction.

Table 1. Various Text Extraction Techniques

Author, Year	Technique Used	Images	Parameters	Remarks
Raj et al. [14], 2014	CC based	Natural Scene Images (Devanagari text)	PR= 72.8%, RR=74.2 %	Fails for small slanted/curved text.
Anupama et al.[4], 2013	Morphology operators, Histogram Projection (X and Y histogram)	Handwritten Telugu document images.	DR=98.54%, Accuracy =98.29%	Fail in case of touching characters and over- lapping lines.
Azadboni et al. [6], 2012	FFT Domain Filtering , SVM Classification,K-means clustering	Scene text images	DR= 98.10%	Text characters having uniform colour.
Seeri et al. [15], 2012	Median filter, Sobel edge detector connected component labeling, order static filter.	Kannada text images	PR=84.21% RR=83.16% Accuracy = 75.77%	Fails to extract very small characters.

III. PROPOSED METHODOLOGY

There are five major stages in Text extraction. They are as follows:

1. Image Digitization
2. Image Pre-processing
3. Image Segmentation
4. Image Feature Extraction
5. Image Post-processing

1) **Digitization** – In this step of image processing digital image of document is captured that is image is represented into bits for further processing of text extraction. Here Input threshold image is a colored image which is converted into a gray scale image.

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

2) **Pre-processing** - during this stage, defects in the image are eliminated which could lead to poor recognition. This binary image is given as input. Character is damaged or smeared mashed with a filling and thinning process.

3) **Segmentation** - character image is segmented into sub-components, namely the segment. Segments can also be identified supported supported characters such as property or popularity, a component corresponding to a predetermined class. Segmentation in the context of the text isolates individual characters of the text.

4) **Extraction feature** - The target feature extraction is to capture the essential characteristics of symbols. Another important task relating to the classification feature extraction. Classification is that the process of identification of each character and assign it the correct character classes. Features extracted features diagonal, transitional features, zoning, directional, parabolic curve fitting features, crossing and features open end points, etc.

5**) Post-processing** - This includes clustering and fault detection. recognized text characters associated with the string, so we want to develop the first string. Error detection is also underway. If the word is not in the dictionary, errors are detected and corrected by changing the words to be the most similar word.

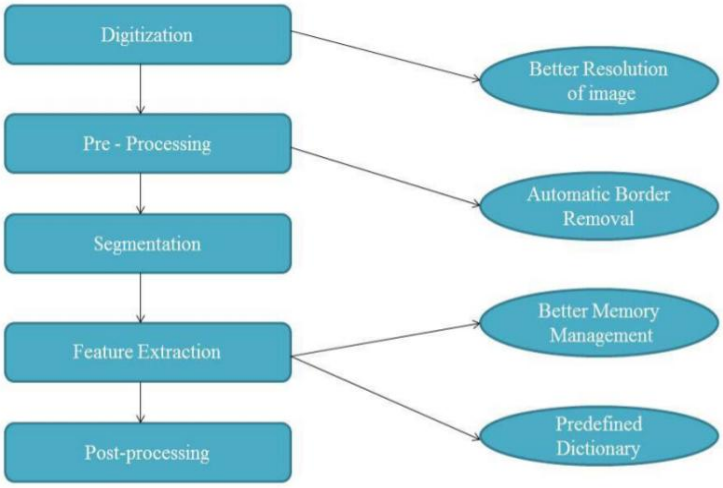

Figure 1. Flow chart

Digitization is just the conversion of handwriting or text documents in electronic form. Further image pre-processing pass, Pre-processing includes segmentation locations, noise reduction which consists of smoothing, thinning, fix broken, de-skewing, etc. images. The segmentation process with reference to the separation of the individual characters from the image. Once the characters are separated, the features of the individual characters are taken, which include the

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

diagonal, crossing, transition, direction, curve fitting, etc., and sent for post-processing. Post processing includes grouping character and fault detection.

IV. RESULTS

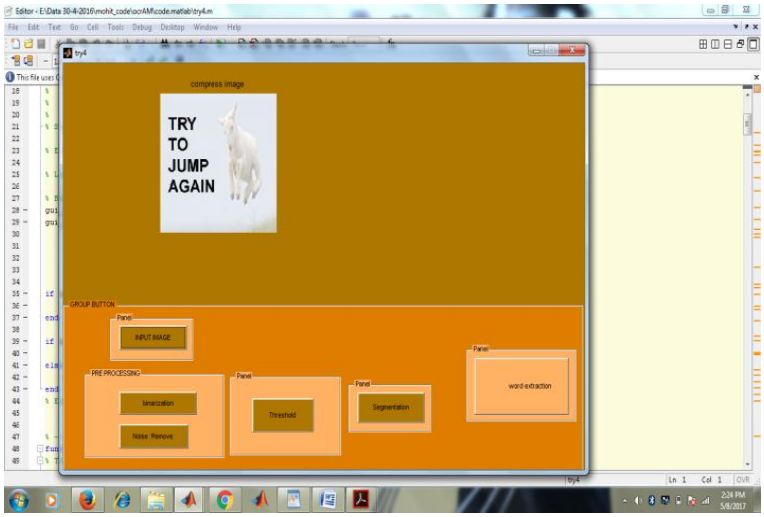

Figure 2. Browse Image in GUI (Graphical User Interface)

In this figure we can take input image in button click event made a press button which presents me to scrutinize via my working vault and select either a 'jpg' or 'bmp' photo.

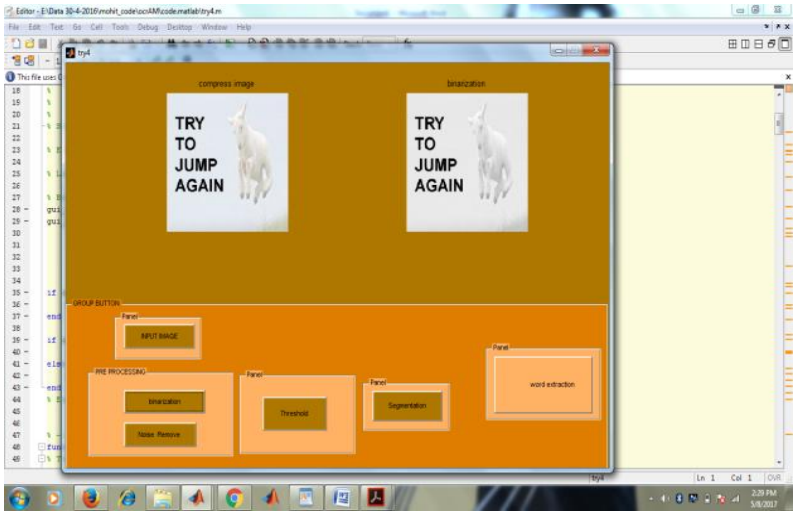

Figure 3. Preprocessing binarization & noise Removal images

In this figure The broken or smeared characters are smoothed by filling and thinning processes. Filling eliminates small gaps, holes, breaks in the character. Thinning reduces the width of the line. Also de-skewing of the image takes place. Pre-processing – In this stage, the defects in the image are eliminated which may cause poor recognition.

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

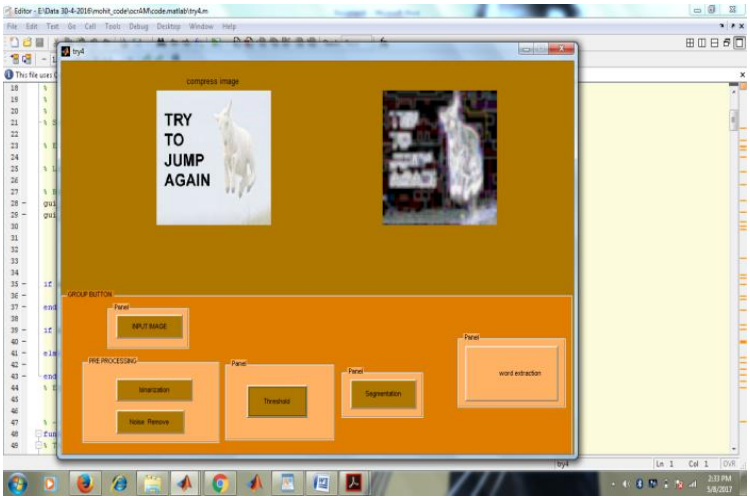

Figure 4. Images Thresholding

Thresholding – Character image is Threshold into its sub-components i.e. Threshold. The Threshold may identified based on character like properties or based on the recognition, components that match the predefined class. Threshold in context of text is isolating the individual character from the text.

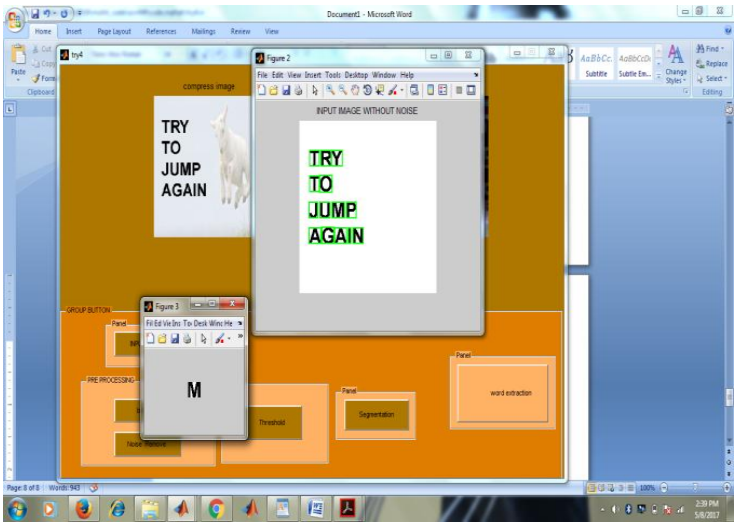

Figure 5. Segmentation and feature extraction

segmentation and feature extraction– The objective of feature extraction is to capture essential characteristics of symbols. Another important task associated with feature extraction is classification. Classification is the process of identifying each character and assigning to it correct character class. The features extracted are diagonal features, transition features, zoning, directional, parabolic curve fitting features, intersection and open end point features, etc.

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

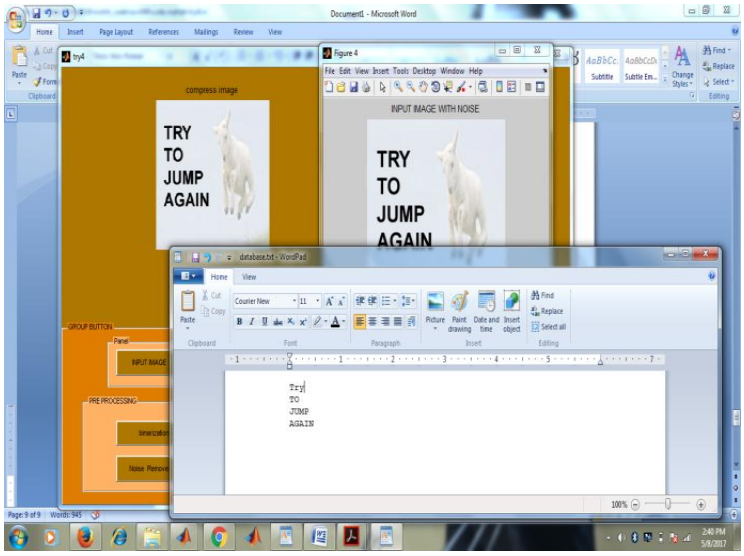

Figure 6. Final Output After all process we can get output in text format form image.

PRECISION AND RECALL

Precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP+FN}$$

No. Image	Image type	Precision	Recall
50	Simple text text image	98%	78%
50	Image text	97%	72%

Table 2. Precision and Recall

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

V. CONCLUSION

As a result, text-image analysis required to enable the text information extraction system which will be used for all kinds of images, including scanned images of documents, real scene image through a video camera, a picture of a text description. Studies show above that, most methods fall under one among techniqus above and also that there are limitations of every technique to supply better detection rate with fewer false alarms without constraints for extracting text regions in various sorts of images. But still, it requires a very powerful technique and customary for text segmentation, it's difficult to supply appropriate input for optical character recognition systems. Thus, the combined methods are proposed for automatic extraction of text content from a special image that's independent of the varied characteristics of the text.

VI. REFERENCE

[1] N. Venkatia Rao, 2DR. A.S.C.S. Sastry, 3a.S.N.Chakraviarthy, 4 Kalyanchakraviarthi P, Optical Character Recognition Algorithm Engineering, Journal of Theoretical and Applied Information Technology 20 Jan 2016 Vol.83. no.2.

[2] Noiman Islam, Zeeshian Islaim, Nazia Noor, A Suarvey on Optical Character Recognition System, Journal of Information & Communication Technology-JICT Vol. 10 Issue. 2 December 2016.

[3] Er. Neeteu Bhatiae, Optical Character Recognition Techniques: A Review, International Journal of Advanced Research in Computer Science and Software Engineering.

[4] Shalitn A. Chopra1, Atmit A. Ghadge2, Ontkar A. Padwal3, Karan S. Punjabi4, Prof. Gandhali S. Gurjar, Optical Character Recognition, International Journal of Advanced Research in Computer and Communications Engineering Vol. 3, Issue 1, January 2014.

[5] Amarjott Singh, Ketan Batcchuwar, and Akshay Bhtasin, A Survey of OCR applications, International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.

[6] H. Raj, R. Ghosh, Devanagari Text Extraction from Natural Scene Images, 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2014, pp. 513-517.

[7] S.V. Seeri, S. Giraddi and Prashant. B. M, A Novel Approach for Kannada Text Extraction, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, 2012, pp. 444-448.

[8] M.K. Azadboni, A. Behrad , Text Detection and Character Extraction in Color Images using FFT Domain Filtering and SVM Classification, 6th International Symposium on Telecommunications. IEEE, 2012, pp. 794-799.

ISSN: 2278-4632
Vol-10 Issue-6 No. 10 June 2020

[9] H. Anoual, D. Aboutajdine, S.E. Ensias, A.J. Enset, Features Extraction for Text Detection and Localization,5th International Symposium on I/V Communication and Mobile Network, IEEE, 2010, pp. 1-4.

[10] S.Hassanzadeh, H. Pourghassem, Fast Logo Detection Based on Morphological Features in Document Image, 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, 2011, pp. 283-286.

