International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-9, Issue-3, June 2022
https://doi.org/10.55524/ijirem.2022.9.3.12
Article ID IRPV1040, Pages 89-92
www.ijirem.org

Text Extraction from Image

Ojas Kumar Barawal[1], and Dr Yojna Arora[2]

[1,2] Department of Computer Science & Engineering, Amity University, Haryana, India

Correspondence should be addressed to Ojas Kumar Barawal; ojaskumarbarawal@gmail.com

ABSTRACT-Text extraction is one of the key tasks in document image analysis. Automatic text extraction without characters recognition capabilities is to extract regions just contains text. The text extraction process includes detection, localization, segmentation and enhancement of the text from the given input image. In this paper we present a comparative study and performance evaluation of various text extraction techniques.

KEYWORDS- Text Extraction, Document Text Images, Heterogeneous Images, OCR, Text Detection

I. INTRODUCTION

Text extraction from an image is a challenging problem because of image contains text due to different size, style, orientation, alignment, low contrast, noise and has complex background structure. This extracted text contains only black text in white background, i.e. it can be recognized by any recognition system. Extracting text from an image or video includes in different applications like document processing, image indexing, video content summary, video retrieval, video understanding etc. The Figure1 shows the steps involved in the text extraction technique. Text detection refers to the determination of the presence of text in a given input image, done by exploiting the discriminate properties of text characters such as the vertical edge density, the texture or the edge orientation variance. Text regions should have high contrast than the background; otherwise it would not be easily readable. This is the basic idea behind text localization, which is referred as locating the text portion in the image. The located portion is extracted during the text extraction phase. The output from this phase is given to OCR in order to eliminate as many falsely identified text regions as possible [1], [2]

This paper mainly focuses on different methods for text extraction and presents a survey of techniques which includes region-based technique, edge-based technique, texture-based technique and morphological-based technique.

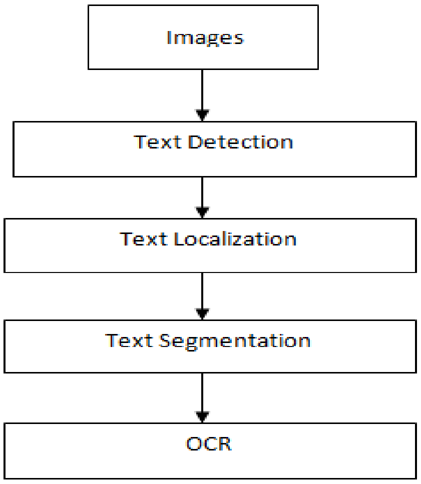


Figure 1: Text Extraction Block Diagram

II. TEXT EXTRACTION TECHNIQUES

Text extraction is one of the required stages prior to character recognition. The aim of text extraction is to separate each character so that it can be fed into the recognition stage. This paper discusses about different text extraction techniques such as region-based, edge-based, texture-based and morphological-based techniques. [3]

A. Region-Based Technique [4]

Region-based methods use the properties of the colour or gray-scale in a text region and their differences with the corresponding properties of the background. Regarding the image representation, region-based image representations provide a simplification of the image in terms of a reduced number of representative elements. In this representation, objects in the scene are obtained by the union of regions in an initial partition.

The bottom-up approach of Line Segmentation from handwritten text. Line segmentation is a process in which the consecutive lines are extracted or separated from each other from a text. For a line segmentation of handwritten text, first the picture is divided into small squares with height and width 10 pixels each. If 50% of the square box is filled up with black pixels then the total square is filled with black pixels. In this way graphically smooth image is found. Then, the height of each of components in the graphically smooth image is computed. Next a rectangular template is created with a specified width and height as maximum portable height. Depending on the height and the position information, these smoothed blocks are joined to get the individual lines. Next the lines are extracted with the help of upper and lower boundaries. Then these are placed one after another in a link list, i.e. the nodes of the link list are the lines. Thus an unconstrained handwritten script is line segmented.[5]

A method for identification of Text on colored book and journal covers. To reduce the amount of small variations in color, a clustering algorithm is applied in a preprocessing step. Two methods have been developed for extracting text hypotheses. One is based on a top-down analysis using successive splitting of image regions alternately in horizontal and vertical direction. Regions obtained under this top-down procedure are always of rectangular shapes and regions containing text include at least two colors. The other is a bottom-up region analysis detects homogeneous regions using growing algorithm.

Beginning with the start region pixels within a 3x3 neighborhood are iteratively merged if they belong to the same cluster. The results of both methods are combined to robustly distinguish between text and non-text elements. Text elements are binarized using automatically extracted information about text colour. The binarized text regions can be used as input for a conventional OCR module. The proposed method is not restricted to cover pages, but can be applied to the extraction of text from other types of colour images as well.

B. Edge-Based Technique

Edge-based text extraction algorithm is a general-purpose method for text extraction. It quickly and effectively localizes and extracts the text from both document and images. Edges are considered as a very important portion of the perceptual information content in a document image, which represents the significant intensity variations, discontinuities in depth, surface orientation, change in material properties etc.[5], [6], [7]

Vertical edges are detected by using smooth filter and it is connected into text clusters for the purpose of text extraction in edge-based technique. an edge-based technique consists of four modules: multistage pulse code modulation(MPCM),text region detection (TRD) module, text box finding (TBF) module and optical character recognition (OCR).In the first module ,MPCM ,is used to locate potential text region in colour image. It convert image to coded image. In coded image each pixel encoded by a priority code ranging from 7 down to 0 in accordance with its priority and further produces a binary threshold image. The TRD module uses spatial filters to remove noisy regions and it also eliminate regions that are unlikely to contain text. Five filtering steps are included in this module: thresholding, elimination of isolated blocks, elimination of long vertical blocks, elimination of diagonally connected blocks and elimination of weakly connected vertical blocks. TBF module uses merge text regions and produces boxes that are likely to contain text. That is this module rectangularizes the text regions detected by TRD module and produce text boxes. The final OCR module eliminates the text boxes that produce no OCR output. The output of OCR module is a simple binary decision to determine whether a text box contains text. a multi-scale edge-based text extraction algorithm which can quickly and effectively localize and extract text from both documents and images.

The proposed method consists of three stages: candidate text region detection, text region localization and character extraction. The first stage aim to build a feature map by using three important properties of edges: edges strength, density and variance of orientations. The feature map is a gray-scale image with same size of the input image. Normally text embedded in an image appears in clusters that is, it's arranged compactly. Thus characteristics of clusters can be used to localize text regions. The purpose of character extraction stage is to extract accurate binary characters from the localize text regions so that we can use existing OCR directly for recognition.

C. Texture-Based Technique [8], [9]

Texture-based methods use the observation that texts in images have distinct textural properties that distinguish them from the background, to decide whether a pixel or block of pixels belong to text or not. Text feature extraction lies essentially on image pre-processing techniques, which is usually performed by linearly transforming or filtering the textured image followed by some energy measure or non-linear operator.

III. IMPLEMENTATION

During the implementation, initially the image will be uploaded and then the operations will be performed. The user interface for image uploading will look like as mentioned in the figure below.

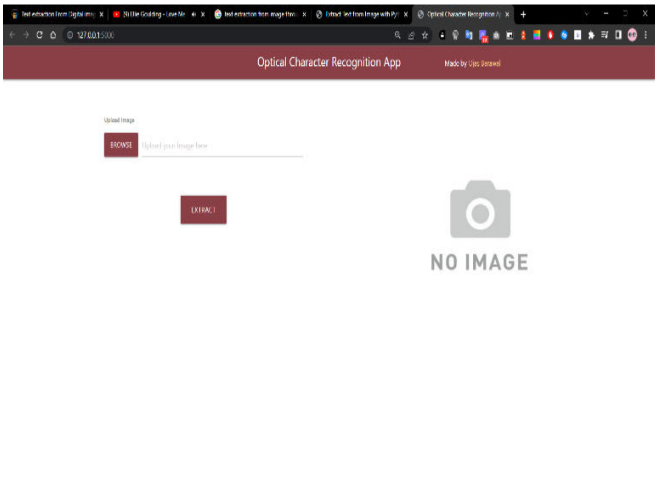
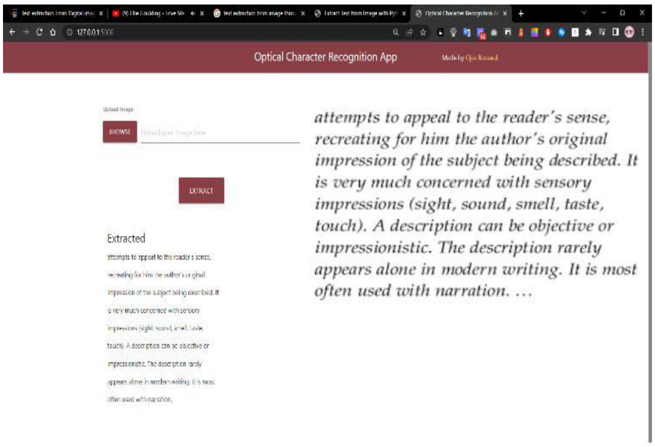


Figure 2: User Interface

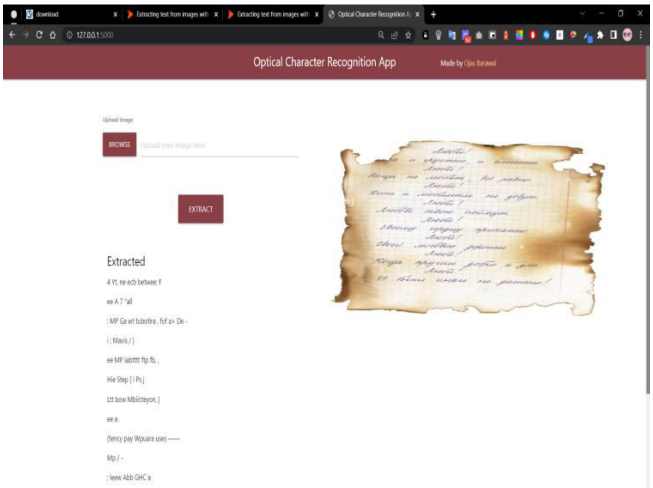


attempts to appeal to the reader's sense, recreating for him the author's original impression of the subject being described. It is very much concerned with sensory impressions (sight, sound, smell, taste, touch). A description can be objective or impressionistic. The description rarely appears alone in modern writing. It is most often used with narration. ...

Extracted

attempts to appeal to the reader's sense, recreating for him the author's original impression of the subject being described. It is very much concerned with sensory impressions (sight, sound, smell, taste, touch). A description can be objective or impressionistic. The description rarely appears alone in modern writing. It is most often used with narration.

Figure 3: Visible Image Upload



Extracted

4 Yt, ne ecb betwee; F

ee A 7 'all

; MP Ga wt tubotira , fof a> De -

i : Mavis / |

ee MP iabtttt ftp fb. ,

Hie Step [i Ps j

Ltt bow Mbiicteyon, |

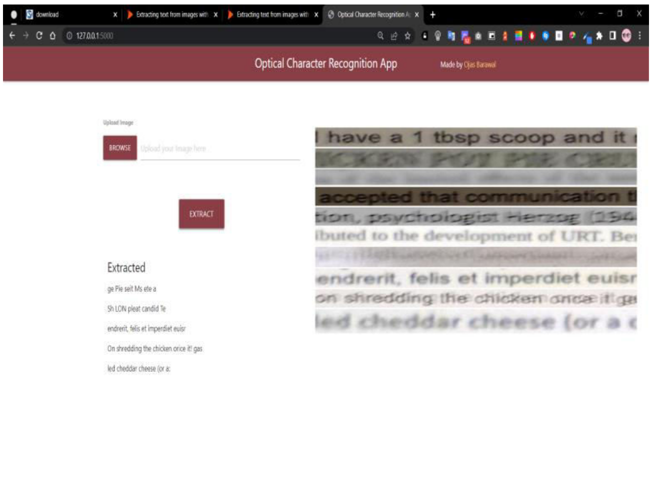
ee a

(tency pay Wpuara uses ——

Mp / -

; leew Abb GHC a

Figure 4: Burnt Document Image



Extracted

ge Pie seit Ms ete a

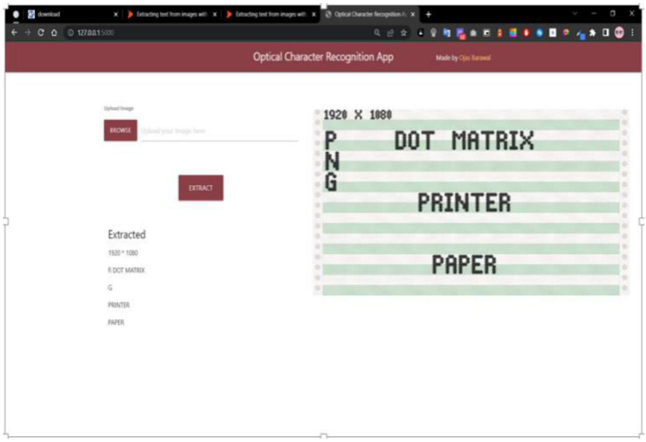
Sh LON pleat candid Te

endrerit, felis et imperdiet euisr

On shredding the chicken orice it! gas

led cheddar cheese (or a:

Figure 5: Blurred Ima



Extracted

1920 * 1080

fi DOT MATRIX

G

PRINTER

PAPER

Figure 6: Dot Matrix Printer Document

IV. COMPARISON AND PERFORMANCE EVALUATION

The performance evaluation of information retrieval can be done using precision and recall rate. The precision rate measures the percentage of correctly detected text boxes with in each image as opposed to detected boxes, where aspercentage of correctly detected text boxes that actually contain in text are measured by recall rate.

Precision rate = Number of correctly detected text boxes / Number of detected text boxes

Recall rate = Number of correctly detected text boxes / Number of text boxes

Table 1. Performance Comparison

S No.	Document Type	Accuracy
1	Clear Image	98.1%
2	Burnt Image	62.5%
3	Blurred Image	96%
4	Document Image from Dot Matric Printer	85.2&%

V. CONCLUSION

Automatic text detection and extraction from an image is an important research branch of content-based information retrieval and text based image indexing. Some of the applications fields of text extraction are mobile robot navigation, vehicle license detection and recognition, object identification, document retrieving, page segmentation etc. Based on the information collected from various techniques it is found that morphological and edge based techniques can quickly and effectively localize and extract text from images. The remaining methods, region and texture based techniques, show the poor performance compared to morphological and edge based technique. In this work, an interface is implemented for text extraction, where one can save the time and be productive. It will automatically detects and extracts text from images very efficiently using inbuilt functions of pytesseract and opencv.

EFERENCES

[1] Y. Zhan, W. Wang, W. Gao (2006), "A Robust Split-And-Merge Text Segmentation Approach For Images", International Conference On Pattern Recognition,06(2):pp 1002-1005.

[2] Thai V. Hoang , S. Tabbone(2010),"Text Extraction From Graphical Document Images Using Sparse Representation"in Proc. Das, pp 143–150.

[3] Sumathi, C. P., Santhanam, T., & Devi, G. G. (2012). A survey on various approaches of text extraction in images. International journal of computer science and engineering survey, 3(4), 27.

[4] Leon, M., Vilaplana, V., Gasull, A., & Marques, F. (2009, November). Caption text extraction for indexing purposes using a hierarchical region-based image model. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 1869-1872). IEEE.

[5] Takahashi, H., & Nakajima, M. (2005, July). Region graph based text extraction from outdoor images. In Third International Conference on Information Technology and Applications (ICITA'05) (Vol. 1, pp. 680-685). IEEE.

[6] Liu, X., &Samarabandu, J. (2006, July). Multiscale edge-based text extraction from complex images. In 2006 IEEE International Conference on Multimedia and Expo (pp. 1721-1724). IEEE.

[7] Liu, X., &Samarabandu, J. (2005, July). An edge-based text region extraction algorithm for indoor mobile robot navigation. In IEEE International Conference Mechatronics and Automation, 2005 (Vol. 2, pp. 701-706). IEEE.

[8] Jung, K., & Han, J. (2004). Hybrid approach to efficient text extraction in complex color images. Pattern Recognition Letters, 25(6), 679-699.

[9] Okun, O., &Pietikäinen, M. (2000). A survey of texture-based methods for document layout analysis. In Texture Analysis in Machine Vision (pp. 165-177).