Extraction of text from an image

[1]Sanskruti Ghodke, [2]Milind Rane
[1,2]Electronics Department, Vishwakarma Institute of Technology

Abstract-Sometimes text which appears in the images is important for fully understanding the images. Now days, use of digital images and digital videos has increased tremendously. Although there are many methods have been proposed over the past years for the text extraction from natural scene images, the text detection and extraction from born-digital images are still a challenge. In this paper, I describe various existing methods and key ideas. I will try to find a new way which can comprehensively utilize existing methods to detect and extract text from digital image.

Keywords-*Image Pre-processing; Binarization; Localization; Character Segmentation; Character extraction*

I. INTRODUCTION

Now a day, most of the data is offered either on paper or within the sort of pictures or videos. Massive data is kept in pictures. Thus, there is a demand for a system to extract text from any general backgrounds. Text Extraction and recognition in pictures has become a possible application in several fields like Image categorization, Robotics, Intelligent transport systems etc. as an example capturing vehicle plate data through a video camera and extracting identification number in traffic signals. However, variations of text thanks to variations in size, style, orientation, and alignment, also as low image distinction and sophisticated background build the matter of automatic text extraction very difficult. For example, within the familiar Project at Carnegie Philanthropist University, text occurrences in pictures and videos are one vital supply of knowledge to supply full-content search and discovery of their T digital library of newscasts and documentaries [1]. Content-based image categorization refers to the method of attaching labels to photographs supported their content

1) Perceptual content and
2) Semantic content

Perceptual content includes attributes such as Color, intensity, shape, texture, and their temporal changes. A number of studies on the use of relatively low level perceptual content for image have already been report Semantic content means objects, events, and their relations. Studies on semantic image content in the form of text, face, vehicle, and human action have also attracted some recent interest.

II. RELATED WORK

Handwriting recognition has become wide and also the vital space of analysis within the field of image process and pattern recognition. With the growing procedure power character recognition methodologies are improved and increasing its demand in numerous applications. It's a tough task to develop a sensible system of written character recognition with high accuracy of recognition. Within the existing systems the accuracy of recognizing the text depends vastly on the standard of the input document. Optical character recognition (OCR) is typicallyspoken as AN off-line characterrecognition method to mean that the system scans and acknowledges static pictures of the characters [2].

III. EXPERIMENTATION

A. Pre Process Preprocessing

Preprocessing is that the primary and also the major step of OCR code. At this stage bound operations are performed on the scanned image i.e. de-skew, changing a picture Text Extraction from pictures 981 from color to black and white, identifies columns, paragraphs, captions as completely different blocks and normalization. .

B. Segmentation

The aim of image segmentation is to supply label to every picture element in a picture such pixels with identical label share bound visual characteristics. Image segmentation is usually accustomed find objects and bounds (lines, curves etc.) in pictures. The tactic of segmentation employed in this is often edge detection

C. Feature extraction

The aim of feature extraction is to capture the essential characteristics of the symbols, and it's been accepted that this is often one amongst the most important issues of pattern recognition. During this the approach is to extract bound options that characterize the symbols, however leaves out the unimportant attribute.

The Selection of the suitable feature extracting technique is maybe one amongst the foremost vital factors in achieving high recognition performance [3].

D. Classification and recognition

The categoryifying and distinctive of every character and assignment to that the right character class is named classification. During this stage the choice creating of a recognition system uses all the options extracted within the earlier stage [4].

E. Post processing

It's the ultimate step of recognition system being mentioned. It prints the corresponding characters that were recognized within the structured text type that is completed bythecalculationofequivalent code price exploitation recognition index of the check samples [5].

II. BLOCK DIAGRAM

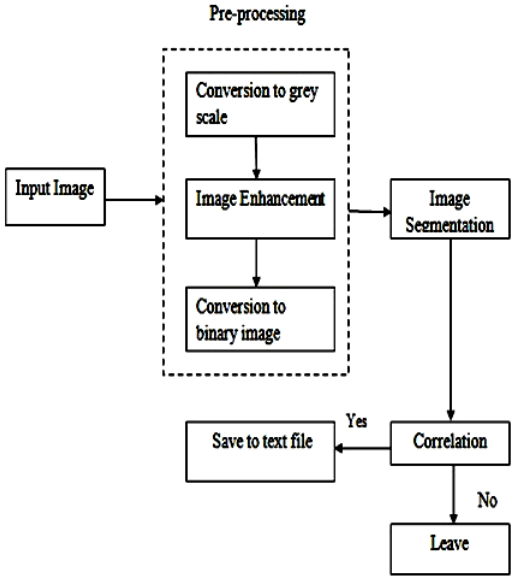

Fig .1: Block diagram of extraction of text from an image.(Extraction of text from an image)

III. PERFORMANCE AND EXPERIMENTS

Matlab code:-

```
%Extracting text from image
clc
close all
clear all
Input=imread('C:\Users\Sanskruti\Desktop\sanskruti.jpg')
;
% Showing input image
figure
imshow(Input);
title('INPUT IMAGE')
%RGB to a Gray conversion
if size(Input,3)==3 % RGB image
 Input=rgb2gray(Input);
end
% Convert to binary image
threshold = graythresh(Input);
Input =~im2bw(Input,threshold);
figure
% Remove all object containing fewer than 40 pixels
Input = bwareaopen(Input,40);
pause(1);
% Labelling all the connected components
[L, R]=bwlabel(Input);
%Measure the properties of the Image regions and Plot
the bounding Box
props=regionprops(L,'BoundingBox');
imshow(~Input);
hold on
for n=1:size(props,1)
rectangle('Position',props(n).BoundingBox,'EdgeColor','r'
,'LineWidth',1)
end
hold off
pause (0.5);
%Letter segmenatation & Objects extraction
for n=1:R
  [r,c] = find(L==n);
  n1=Input(min(r):max(r),min(c):max(c));
  figure
  imshow(n1);
  pause(0.5)
end
```

IV. FUTURE SCOPE

There are numerous applications of a text information extraction system, including Using the document analysis, vehicle license plate extraction, technical paper analysis, and object oriented data compression. In the following, we briefly describe some of these applications. Wearable or portable computers: with the rapid development of computer hardware technology, wearable computers are now a reality. A TIE system involving a hand-held device and camera was presented as an application of a wearable vision system. Translation camera can detect text in a scene image and translate Japanese text into English after performing character recognition. Content-based video coding or document coding: The MPEG-4 standard supports object based encoding. When text regions are segmented from other regions in an image, this can provide higher compression rates and better image quality. As a result, they can achieve a higher quality rendering of documents containing text, pictures, and graphics. Text-based image indexing: This involves automatic text-based video structuring methods using caption data. License/container plate recognition: There has already been a lot of work done on vehicle license plate and container plate recognition. Although container and vehicle license plates share many characteristics with scene text, many assumptions have been made regarding the image acquisition process (camera and vehicle position and direction, illumination, character types, and colour) and geometric attributes of the text. Texts in WWW images: The extraction of text from WWW images can provide relevant information on the Internet. Video content analysis: Extracted text regions or the output of character recognition can be useful in genre recognition. The size, position, frequency, text alignment, and OCR results can all be used for this. Industrial automation: Part identification can be accomplished by using the text information.

V. RESULTS

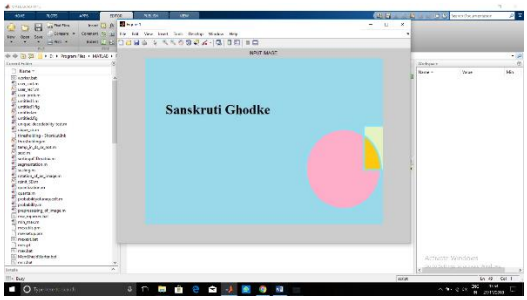

Fig .2:. Result 1 of extraction of text from an image on matlab.(Extraction of text from an image)

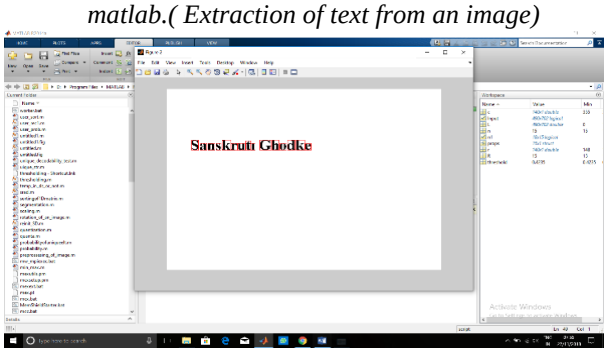

Fig .3:. Result 2 of extraction of text from an image on matlab.(Extraction of text from an image)

VI. CONCLUSION

Earlier optical character recognition can be used for activities like increasing telegraphy and making reading devices for all the blind folks [6]. Throughout 1914 a mortal named Emanuel Reuben Lucius Goldberg had developed a tool that scan characters and converts them into normal telegraph code. Throughout that point, Edmund Fournier d'Albe was developing AN optophone, a hand-held scanner that once touched across a written page, created tones that helped in recognizing specific characters. However it did not scan non optical characters that totally different researches befell. The event befell and (Intelligent Character Recognition) was introduced by M. Sheppard in 1951[7]. Intelligent character recognition (ICR) is a sophisticated optical character recognition (OCR) or rather a lot of specific handwriting recognition system that enables fonts and totally different varieties of handwriting to be learned by a laptop throughout process to boost accuracy and recognition levels. Most ICR computer code features a self-learning system remarked as a neural network, whose job is to mechanically update the popularity info for the contemporary handwriting patterns, thereby extending the quality of scanning devices for the aim of document process, from the written character recognition (a operate of OCR) to hand-written matter recognition, as this method is concerned within the recognition of hand writing[8], typically the accuracy levels might not be excellent however are able to do 97%+ accuracy rates in reading the written content in structured forms. Principally for achieving these high recognition rates many scan engines are used at intervals the computer code and everyone hass given elective pick rights to work out verity reading of characters. In the numeric fields, engines that are designed to scan numbers take preference, whereas in alpha fields, engines are designed to scan hand written letters that have higher elective rights. once these are utilized in conjunction with a tailor-made interface hub, the hand writing will be mechanically be inhabited into a back workplace system avoiding backbreaking manual keying will and may |and might} be a lot of correct than ancient human information entry Intelligent word recognition (IWR) can acknowledge and extract not solely written. -handwritten info, however cursive handwriting still.

REFERENCES

[1]. R. Lienhart and A. Wernicke, Localizing and Segmenting Text in Images and Videos, Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 4, April 2002.
Toolbox" International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1, February, 2013

[2]. Vijay Laxmi Sahu, Babita Kubde, "Offline Handwritten Character Recognition Techniques using Neural Network: A Review" , International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064.

[3]. J. Pradeep, E. Srinivasan, S. Himavathi, " Diagonal based feature extraction for handwritten alphabets recognition system using neural network", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[4]. N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 31, no. 2, (2001), pp. 216 – 233.

[5]. Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre, " Character Recognition Using Matlab's Neural Network.

[6]. Nafiz Arica and Fatos T. Yarman-Vural, " An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 31, no. 2, May 2001