International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Text Information Extraction from Image

Zaiba Khanum, Syed Mustaqeem Uddin, Aimen Zuha Sharieff

Abstract-

Text data extraction is one of the critical tasks in document image analysis, as text data present in images contains valuable information for automatic annotation, indexing, and structuring of images. Automatic text extraction without character recognition capabilities is to extract regions that just contain text. The text extraction process includes detection, localization, segmentation, and enhancement of the text from the given input image. Text due to differences in size, style, orientation, alignment, low image contrast, and complex background make the problem of automatic text extraction extremely challenging. In this paper, we present various text extraction techniques and compare them.

Introduction

Images with text content can be utilized to extract a variety of information. The majority of data is currently available in the form of paper, images, or videos. While enhancing keyword-based searching, indexing, information retrieval, and automatic image captioning efficiency, text within images and videos delivers more relevant information about the visual material. Text information extraction (TIE) systems can detect, recognize, and locate text in visual data like images and videos. There are two forms of image content: perceptual content and semantic content. Perceptual contents include colors, textures, intensities, shapes, and their temporal variations. Semantic content includes objects, events, and their relationships. Textual material contains a great deal of semantic information.Text information extraction (TIE) from images and video has been proposed for a variety of applications, including page segmentation, address block location, license plate location, and content-based image/video indexing. Despite these thorough investigations, developing a general-purpose TIE system remains challenging. This is because there are so many distinct causes of variation when extracting text from a shaded or textured background, from low-contrast or intricate images, or from images with variations in font size, style, color, orientation, and alignment. As a result of these differences, the challenge of automatic TIE becomes more difficult.

Process of Text information Extraction-

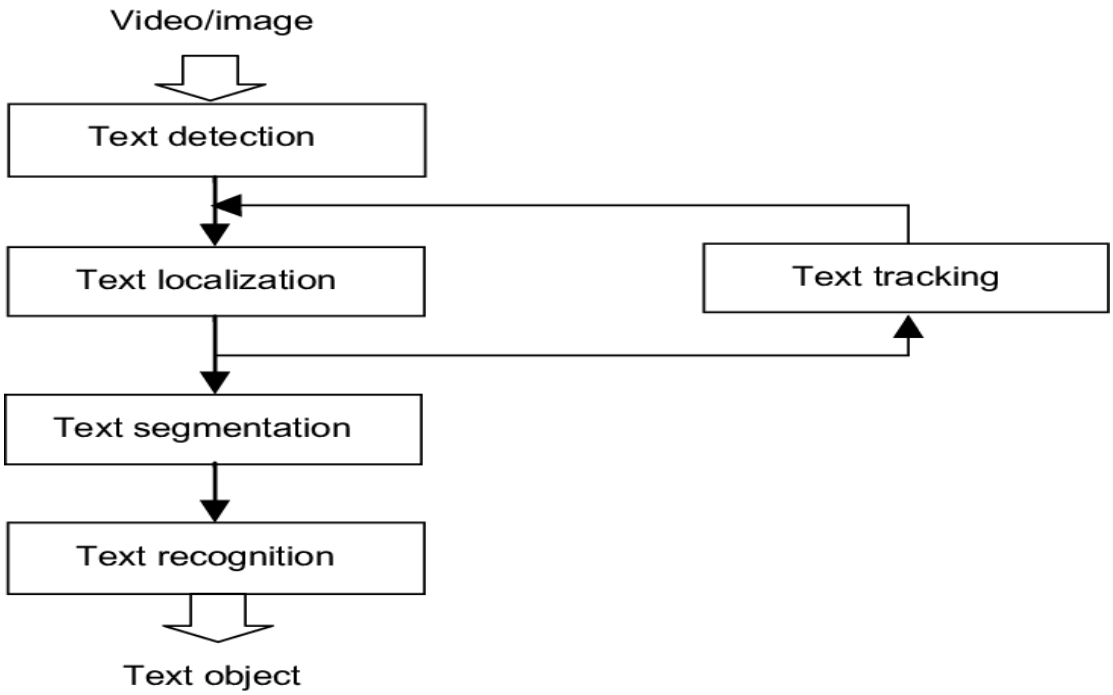


Fig - 1

Text Extraction Techniques -

Prior to character recognition, text extraction is one of the steps that must be completed. Text extraction is used to separate characters so that they can be supplied into the recognition stage. Different text extraction strategies are discussed in this paper such as region-based, edge-based, texture-based, and morphological-based techniques.

1. Region-based-

Region-based methods make use of the colors or gray-scale attributes of a text region, as well as their variations from the background's corresponding properties. These methods are further divided into two types: connected component (CC)-based and edge-based approaches. These two techniques work from the bottom up, recognizing sub-structures such as CCs and edges, and then merging them to create text bounding boxes. In this concept, objects in the scene are created by linking regions in an initial partition.

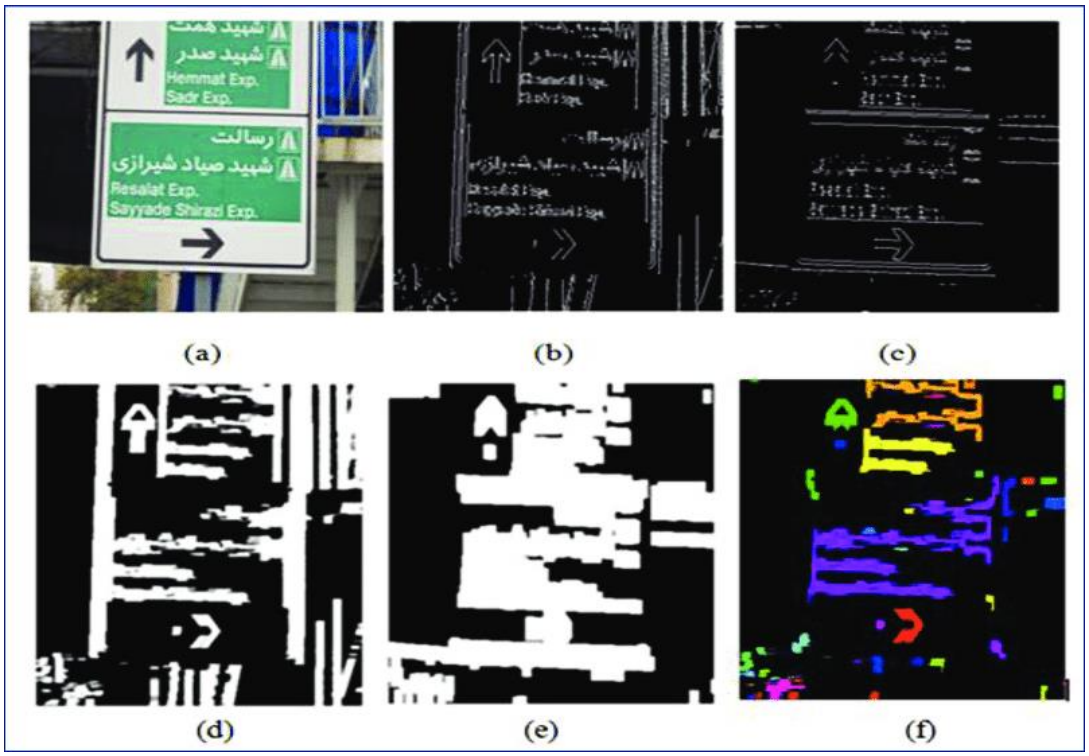


Fig- 1.1

It's a technique that uses a sliding window to analyze or detect text in any image, especially in a natural setting. The sliding window-based method is another name for this technology. This method is based on factors such as color, edge, shape, contour, and geometry. When compared to other procedures, the region-based method technique is quite slow.

2. Texture based-

The texture-based method extracts and evaluates text from a complicated image by utilizing many types of texture and their features. Wavelets, Fourier Transform and Gabor filters, DCT Transform Wavelet, and other approaches are utilized in this approach to extract textual information.

Texture-based methods employ the fact that text in photos has distinct textural features that distinguish it from the background to determine whether a pixel or a group of pixels belongs to text or not. Text feature extraction is based on image pre-processing techniques such as Gabor filters, Wavelet, FFT, spatial variance, and so on. These techniques can be used to detect the textural features of a text region in an image.

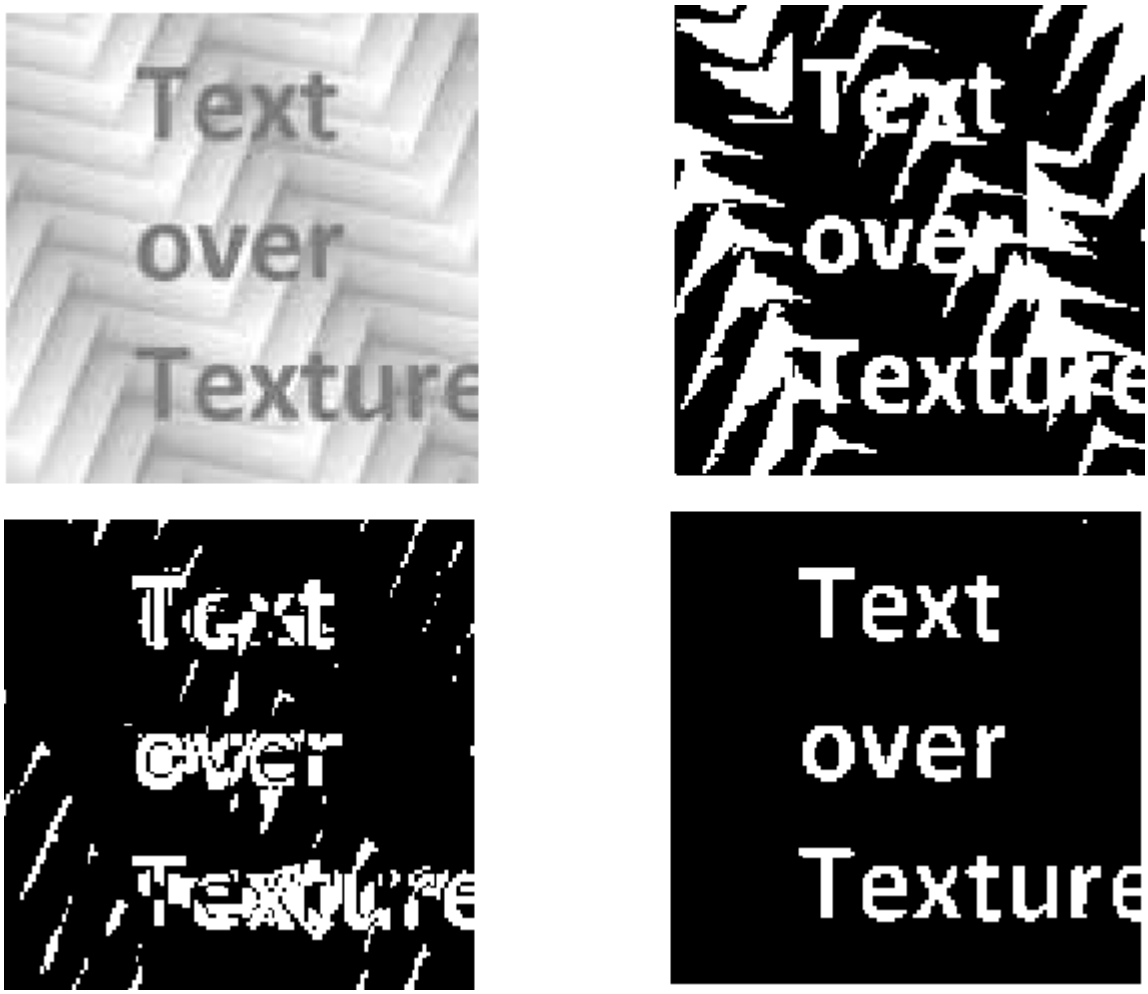


Fig- 1.2

3. Morphological based-

Mathematical morphology is a method of image analysis that is based on topological and geometrical principles. In a variety of applications, it provides sophisticated tools for extracting geometrical structures and expressing shapes. Character recognition and document analysis are two fields where this technology has been frequently used.

This method is used to extract text-related properties from processed photos, however, some image alterations such as translation, rotation, and scale still exist. This approach performs admirably under a variety of image manipulations. On the basis of their many kinds of parameters such as precision rate, recall rate, accuracy, and so on, each of the methodologies described above has its own advantages and limits.

Fig- 1.3

For numerous reasons, a phase of mathematical morphology follows the binarization step: To decrease noise, correct classification errors utilizing information from each pixel's immediate surroundings, and connect characters to make entire sentences.

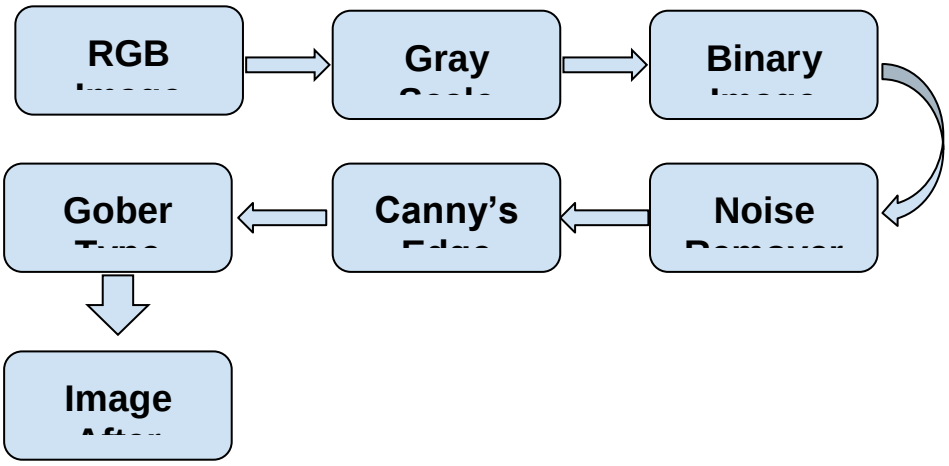
4. Edge-Based Techniques-

The edge-based text extraction algorithm is a versatile text extraction method. It locates and extracts text from both documents and photos fast and effectively. The emphasis is on the "strong contrast between the text and the background." Edges are a critical component of the perceptual information content of a

document image, as they reflect major intensity variations, depth discontinuities, surface orientation, and changes in material qualities, among other things. In the edge-based technique, vertical edges are recognized using a smooth filter and then joined into text clusters for text extraction.

4.1. Proposed System-

Edge-based approaches focus on the 'strong contrast between the text and the background' among the various textual qualities of an image. The edges of the text boundary are located and fused, and then non-text sections are filtered out using multiple heuristics. The edge-based text extraction technique is a general-purpose method for extracting, evaluating, and localizing text from both documents and images more quickly and effectively. ("This approach isn't as well-suited to dealing with enormous amounts of text").



In this case, we'll use the RGB method. The RGB components of a color input image are merged in the following way to produce an intensity image Y: R, G, and B are the red, green, and blue components, respectively, in

Y = 0.299 R + 0.587 G + 0.114 B.

In the gray-level image, several text portions that are prominent in the color image are difficult to recognize. The edges are recognized using a morphological gradient operator after the color conversion. The binary edge image is created by thresholding the resulting edge image. Each potential region in the intensity image that is less sensitive to lighting conditions and reflections is subjected to adaptive thresholding.

Dilation groups edges that are spatially close together to produce candidate regions, while erosion removes

minor components.Non-text components are filtered out using size, thickness, aspect ratio, and gray-level homogeneity, but if we utilize the Canny operation and enhance the edge of text using scale information, this method appears to be noise-resistant. To reduce computational complexity, only one edge point in a small window is used in the estimation of scale and orientation.

This scale information is then used to enhance the text's edges. Morphological dilatation is used to connect the edges into clusters, and heuristic knowledge can be used to filter out non-text clusters. Then, for producing input features, two groups of Gober-type asymmetric filters can be used. The edge data is then scaled up to make it more useful. As a result, structures that do not have the prescribed scales are either eliminated or blurred.

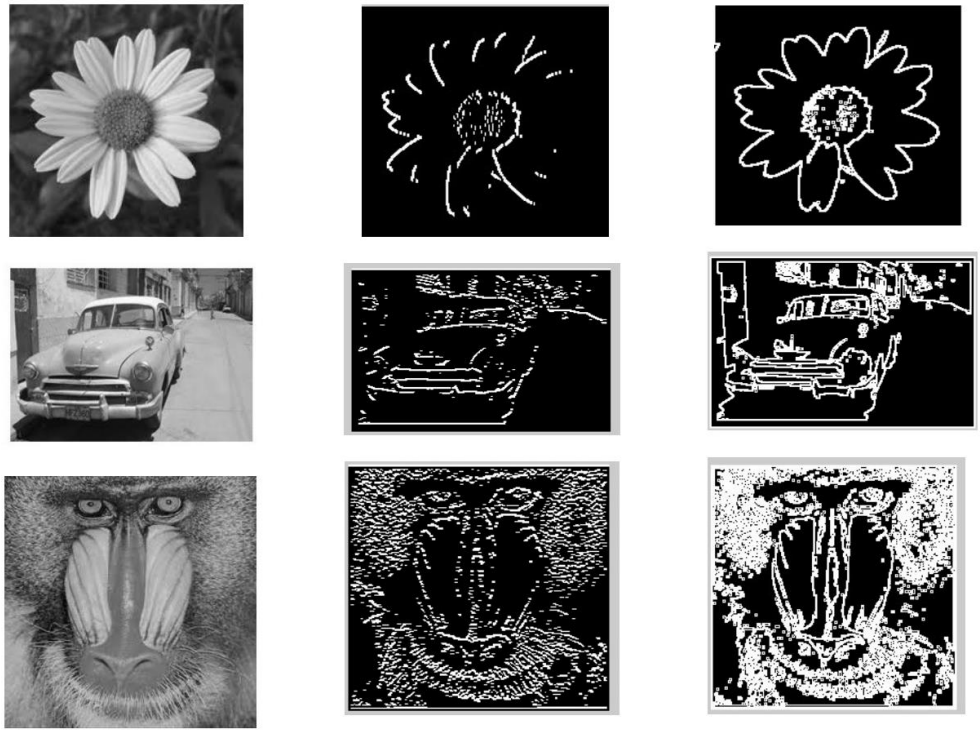


Fig - 1.4

Conclusion-

In this article, a single-scale edge-based segmentation technique is proposed for automatically detecting and extracting text from outdoor scene images and object label images. This method is independent by font sizes, styles, color/intensity, text orientations, or alignment. According to the results of the experiments, the proposed method for extracting text regions from complicated images is successful and efficient.

Reference-

1. Kumary R Soumya, Ancy. T.V, Athulya Chacko, "TEXT EXTRACTION FROM IMAGES: A SURVEY", Volume 3, No.2, February 2014.
2. Keechul Jung, Kwang In Kim, Anil K. Jain, "Text Information Extraction in Images and Video: A Survey"
3. Kiran Adnan, Rehan Akbar, "An analytical study of information extraction from unstructured and multidimensional big data", Article number: 91 (2019).
4. Figure 1 - https://www.researchgate.net/profile/Muhammad-Javed-78/publication/272644687/figure/fig4/AS:668774260826121@1536459617698/Architecture-of-text-extraction-and-recognition-process.png
5. Figure 1.1 - https://www.researchgate.net/figure/Candidate-text-region-extraction-a-original-image-b-vertical-edges-detected-in_fig2_257719721
6. Figure 1.2 - https://d3i71xaburhd42.cloudfront.net/9628be8c612203eab17d5c6e3bdd39b91d5c71c9/2-Figure1-1.png
7. Figure 1.3 - https://www.semanticscholar.org/paper/Text-Extraction-and-Character-Recognition-form-and-Uddin-Sultana/c99e93519819a308ae6142c207ea441bc69961a7
8. Figure 1.4 - https://miro.medium.com/max/1400/1*3hDhkzqJStXTp37lS_M5aA.png