# Text Extraction Techniques

Yash Gupta
Computer Science & Engineering
I.M.S. Ghaziabad, India

Shivani Sharma
Computer Science & Engineering
I.M.S. Ghaziabad, India

Tushina Bedwal
Computer Science & Engineering
I.M.S. Ghaziabad, India

## ABSTRACT
As the growth of technology is emerging, It is beneficial for us to take put some innovative efforts for pulling this computer science field at a higher level. Text extraction is one of the recent growing technique to be enhanced further. Text Extraction is the process of extracting, evaluating and analyzing images. Detection, Localization, Binarization, Extraction, Enhancement, and Recognition are some of the steps to be involved in the process of text extraction. In today's challenging world this technique is a very cumbersome task to be performed because it indulges various activities like changes in fonts,size,orientation,text.There are many text extraction techniques that are based on connected component analysis, edge detection, morphological operators, wavelet transform, neural network, texture features etc. have been developed. In this paper we are providing some of the study of the techniques and comparison between various techniques such as region based technique, texture based technique and hybrid technique.

## Keywords
Text Extraction, Detection, binarization. edge, connected component.

## 1. INTRODUCTION
Text extraction is the very crucial stage of evaluating the images. The steps involved are detection, localization, binarization, extraction, enhancement, and recognition of text from the image. The image content in the image is generally classified into two categories i.e. perceptual content and semantic content. Perceptual contents involve colors, textures, intensities, shapes, and their temporal changes. Semantic contents involve objects, events, and their relations between them. Text content include huge amount of semantic information. Today text extraction technique from images is very essential in content evaluation. This technique can be used in a variety of applications such as image searching, indexing, navigation, and human computer interaction. A text information extraction system receives a still image or a sequel of images in the form of input. These images can be in gray scale or may be colored , compressed or un-compressed,. The text extraction system can be subdivided into the following problems i.e. detection, localization, tracking, extraction and enhancement and recognition.

### 1.1 Properties of Text in Images
Texts usually have different appearance due to changes in font, size, style, orientation, alignment, texture, color, contrast, and background. Some of the characteristics are as follows:

1. Size: The size of text may vary a lot in different images.
2. Alignment: Text may be align in any of the direction and may contain some of the distortions and disturbances such as geometric distortion .
3. Color: The text characters may have same or similar color.
4. Edge: Most of the images have strong edges at the boundaries of text and background.
5. Compression: Some of the images are recorded, transferred, and processed in compressed format.

## 2. PROCESS OF TEXT EXTRACTION

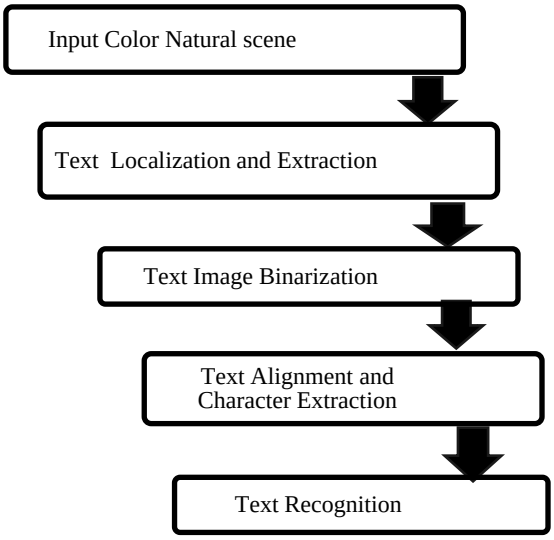


**Figure 1.**

## 3. TEXT CLASSIFICATION
Text is mainly classified into two sets i.e. Artificial Text. And Scene Text. Artificial text is also called the caption text that is inserted in any of the image or video. The text could be segmented, detected and extracted using various algorithms and techniques. The caption text is added into news channels, movies and videos. Caption texts may be the rotating text or any subtitle or moving text. The artificial text may be or not in a fixed position and shape and it may also have low resolution problem. Scene text is text is a type of text which can capture the image accidently. Any vehicle number plate,

street signboards, banners, traffic sign board can be the examples of scene text. Sometimes the scene text is difficult for extracting features because of its various activities such as its color, font, contrast, low and high resolution, orientation, alignment and shadowing effect.

## 4. TEXT IMAGE CLASSIFICATION

Based on image text it is classified into three categories such as document images, caption text images and scene text images. The document image may involves various text and rare graphics components[1]. These are inherited by scanning some types of journals, any printed document or any degraded document images Caption text refers to that text that can be artificially super imposable on any video/image at the time of editing only. This text only explains the subject of the image or video text. . Caption text is particularly called as overlay text or cut line text. The application of caption text is in sports video for the creation of sports highlights. Scene text consists of particularly an important semantic information such as advertisements that include name of the street, institutions, shop, road sign. Some researcher also called scene text as graphic text.

## 5. TECHNIQUES

There are many types of text extraction algorithms and techniques which are used in various research field by many scientists. These techniques are proving very beneficial in different kinds of areas. Some of the techniques are region based, texture based , hybrid technique, connected component based method and edge based method which are discussed below:-

### 5.1 Region Based Method

Region based method is a type of method that uses a sliding window to analyze or detect a text from any kind of image particularly a natural scene. This technique is also called as sliding window based method. This approach rely on some criteria such as color, edge, shape, contour and geometry features. The speed of region based method technique is very slow as compared to other techniques.

### 5.2 Texture Based Method

The texture based method uses various kinds of texture and its properties to extract and evaluate a text from a complex image. Various types of methods are used for this approach to extract textual information like Wavelets, Fourier Transform and Gabor filters, DCT Transform Wavelet etc.

### 5.3 Hybrid Technique

The hybrid technique is a technique of combination of both techniques, i.e. region based and texture based approach. In this approach, firstly region based approach is used to detect a text or character. then by the process of texture based method all the features are extracted from the text region The major disadvantage of the approaches that single method is not suitable for all the natural scene images due to its various features as size, color, font variation .

### 5.4 Connected Component Method

Connected Component based methods use a bottom-up approach or technique in which small components of a image into larger components in any image .This process will run until all regions are identified in a particular image.

All of the three techniques used have their own advantages and limitations on the basis of their different kinds of parameters as precision rate, recall rate, accuracy etc.

### 5.5 Edge Based Method

Edges are a reliable feature of any text regardless of its color, intensity, layout, orientations, etc. Edge based method is a method used to develop a high level contrast between the text and its background. The main essential features or characteristics of the text that is embedded in images are its edge strength, density and the orientation variance. Edge-based text extraction technique is a general-purpose approach that can more quickly and effectively localize ,extract and evaluate the text from both document and image. This approach is not as just robust for handling large amount of text.

### 5.6 Morphological Based Method

Mathematical morphology or morphological based method is a type of method that is used for image analysis and evaluation based on some topological and geometrical method. These technique has been widely used in some of the applications that are character recognition and document analysis[2]. This method is used to extract text related features from the processed images but there are still some changes regarding images like translation, rotation, and scaling. Under different kinds of image alterations this technique works very robustly.

All the techniques explained above have their own advantages and limitations on the basis of their different kinds of parameters as precision rate, recall rate, accuracy etc.

## 6. WHOLE PROCESS

The whole process of text extraction can be explained as in brief. Firstly the natural image is taken off from any natural scene. Then text is detected with a particular technique. After that localization process is performed[3].All the related feature about a particular image is gathered. Scene text extraction and enhancement methods are applied with the help of agorithms.And finally the extracted text is collected from the image.

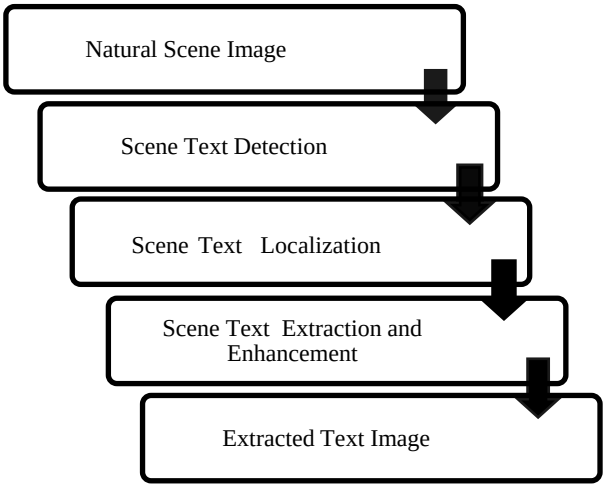


**Figure 2.**

## 7. COMPARISION BETWEEN DIFFERENT TECHNIQUES

**Table:1**

| AUTHOR | YEAR | TECHNIQUE USED | MEASURE |
|---|---|---|---|
| Anhar risnumawan | 2014 | Edge detection | 69% |
| Xu- Cheng Yin etal | 2014 | Connected Component | 76% |
| Xu- Cheng Yin | 2013 | Hierarchal clustering method | 47.56% |
| Yi- Feng Pan et al. | 2013 | Hybrid approach | 92.5% |
| Wahyono et al. | 2014 | Canny edge detector | 68% |

## 8. APPLICATION AREAS

There are several applications of text information extraction system .Some of them are vehicle license plate extraction, visual search system[4]. These applications have been have been developed. Some other are product recognition, landmark recognition, translators for tourists, information retrieval systems in indoor and outdoor environments. There are numerous applications of a text information extraction system, including document analysis, vehicle license plate extraction, technical paper analysis, and object-oriented data compression. Some of the following applications related to this field are:

1. Wearable or portable computers: With the rapid development and advancement of computer hardware technology, wearable computers are now on a boom.
2. Content-based video coding or document coding: The MPEG-4 standard supports object-based encoding. When the text based regions which are segmented from other regions in any image, this can provide more higher compression rates and more better image quality.
3. License container plate recognition: There is still a development needed in the phase of vehicle license plate and container plate recognition.
4. Texts in WWW images: The extraction of text from WWW images can provide relevant information on the Internet.
5. Video content analysis: Extracted text format can be prove very beneficial in genre recognition.
6. Industrial automation: Regarding this concern, part identification can be accomplished and approved by using the text information.
7. Text-based image indexing: This indexing consists of many automatic text-based video structuring methods.

## 9. CONCLUSION

In this paper, various techniques on text extraction have been discussed thoroughly such as region based, edge based, connected component (CC) based, texture based, morphological based method. This paper gives a detailed description of image and image text classification. Each approach used in text extraction have its own importance but all the techniques varies according to their precision rate, recall rate, accuracy etc. Every approach has some benefits and limitations based on their parameters.

But still there needs an improvement in this field because text extraction from the images is still difficult as some of the images have low contrast or complex background. Furthermore efforts can be made so as to improve the reliability and accuracy of the system.

## 10. REFERENCES

[1]. Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan and Chew Lim Tan, "A Robust Arbitrary Text Detection System For Natural Scene Images", Expert System with Application 41(2014) 8027-8048.

[2].Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao, "Robust Text Detection in Natural Scene Images", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 36, no. 5, May 2014.

[3]. H.K. Kim, Efficient Automatic Text Location Method and Content-Based Indexing and Structuring Of Video Database, Journal of Visual Communication and Image Representation vol. 7, no. 4 ,1996, pp. 336–344.

[4] C. Y. Suen, L. Lam, D. Guillevic, N. W. Strathy, M. Cheriet, J. N. Said, and R. Fan, Bank Check Processing System, International Journal of Imaging Systems and Technology, vol. 7, No. 4 1996, pp. 392–403.

[5] D.S. Kim, S.I. Chien, Automatic Car License Plate Extraction using Modified Generalized Symmetry Transform and Image Warping, Proceedings of International Symposium on Industrial Electronics, Vol. 3, 2001, pp. 2022–2027.

[6] A.K. Jain, Y. Zhong, Page Segmentation using Texture Analysis, Pattern Recognition, Vol. 29, No. 5, Elsevier, 1996, pp. 743–770.

[7] T.N. Dinh, J. Park and G.S. Lee, Low-Complexity Text Extraction in Korean Signboards for Mobile Applications, IEEE International Conference on Computer and Information Technology, 2008, pp. 333-337.

[8] Q. Ye, Q. Huang, W. Gao, D. Zhao, Fast and Robust Text Detection in Images and Video Frames, Image and Vision Computing, Vol. 23, No. 6, Elsevier, 2005, pp. 565–576.

[9]. 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.

[10]. Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang, S. Tang, A Novel Image Text Extraction Method Based on K-means Clustering, Seventh IEEE/ACIS International Conference on Computer and Information Science, 2008, pp. 185-190.

[11]. W. Fan, J. Sun, Y. Katsuyama, Y. Hotta, S. Naoi, Text Detection in Images Based on Grayscale Decomposition and Stroke Extraction, Chinese Conference on Pattern Recognition, IEEE, 2009, pp. 1-4.

[12]. N. Anupama, C. Rupa, E.S. Reddy, Character Segmentation for Telugu Image Document using Multiple Histogram Projections, Global Journal of Computer Science and Technology, Vol. 13, 2013, pp. 11-16.