ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Comparison of Text Extraction Techniques- A Review

Divya gera[1], Neelu Jain[2]

ME Scholar, Dept of E & C, PEC University of Technology, Chandigarh, India[1]

Associate Professor, Dept of E & C, PEC University of Technology, Chandigarh, India[2]

ABSTRACT: Text in images contain important contents for information indexing and retrieval, automatic annotation and structuring of images. Hence text extraction is the crucial stage of analyzing the images. The steps involved in text extraction algorithms are detection, localization, binarization, extraction, enhancement, and recognition of text from the image. Text extraction is a very challenging task due to the variations in text size, font, style, orientation and alignment as well as complex background. Several text extraction techniques based on edge detection, connected component analysis, morphological operators, wavelet transform, texture features, neural network etc. have been developed. This paper provides a review of the various techniques suggested by researchers and their comparative analysis in terms of precision rate, recall rate, detection rate etc.

KEYWORDS: Discrete wavelet transform, Connected Component, Edge, Support vector machine, Discrete cosine transform.

I. INTRODUCTION

The image content is classified into two categories: perceptual content and semantic content [1]. Perceptual contents include colors, shapes, textures, intensities, and their temporal changes while semantic contents include objects, events, and their relations. Text content contains high level of semantic information as compared to visual information. Therefore text extraction from images is very significant in content analysis. It has many useful applications such as automatic bank check processing [2], vehicle license plate recognition [3], document analysis and page segmentation [4], signboard detection and translation [5], content based image indexing, assistance to visually impaired persons, text translation system for foreigners etc.

Text appearing in images is classified into three categories: document text, caption text, and scene text [6]. In contrast to caption text, scene text can have any orientation and may be distorted by the perspective projection therefore it is more difficult to detect scene text.

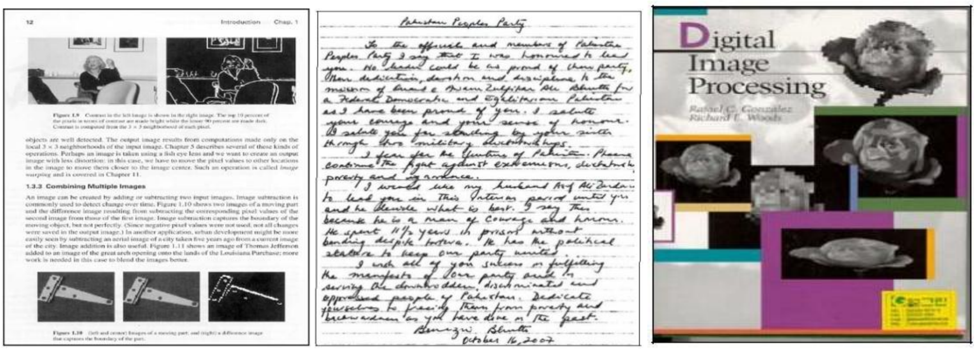
- Document text: A document image (**Fig. 1**) usually contains text and few graphic components. It is acquired by scanning journal, printed document, handwritten historical document, and book cover etc.
- Caption text: It is also known as *overlay text* or *artificial text* (**Fig. 2**). It is artificially superimposed on the image at the time of editing, like subtitles and it usually describes the subject of the image content.
- Scene text: It occurs naturally as a part of the scene image and contain important semantic information such as advertisements, names of streets, institutes, shops, road signs, traffic information, board signs, nameplates, food containers, street signs, bill boards, banners, and text on vehicle etc (**Fig. 3**).

A. *Properties of Text in Images:*

Texts usually have different appearance due to changes in font, size, style, orientation, alignment, texture, color, contrast, and background. These changes will make the problem of automatic text extraction complicated and difficult. Text in images exhibit variations due to the difference in the following properties [7]:

- Size: The size of text may vary a lot.
- Alignment: Scene text may be aligned in any direction and have geometric distortions while caption text usually aligned horizontally and sometimes may appear as non-planar text.
- Color: The characters tend to have same or similar color but low contrast between text and background makes text extraction difficult.

- Edge: Most caption and scene texts are designed to be easily read, hence resulting in strong edges at the boundaries of text and background.
- Compression: Many images are recorded, transferred, and processed in compressed format. Thus, a faster text extraction system can be achieved if one can extract text without decompression.
- Distortion: Due to changes in camera angles, some text may carry perspective distortions that affect extraction performance.



(a) (b) (c)

Fig. 1 Document Images (a) Gray-scale (b) Handwritten (c) Multi-color

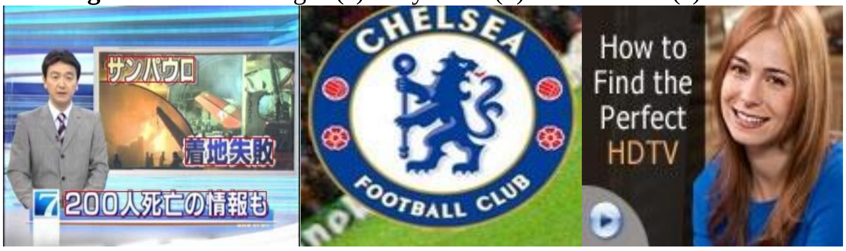


Fig. 2 Caption text images

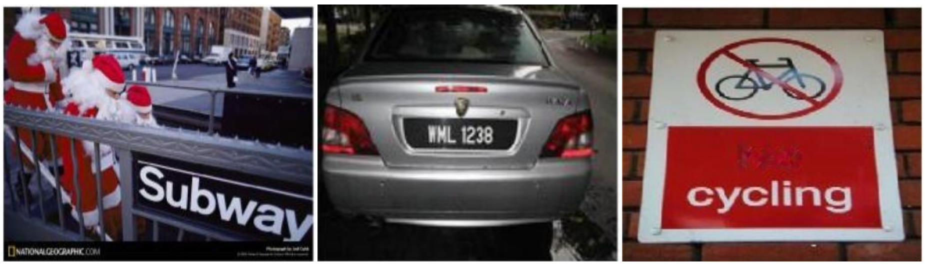


Fig. 3 Scene text images

B. *Process of Text Extraction:*

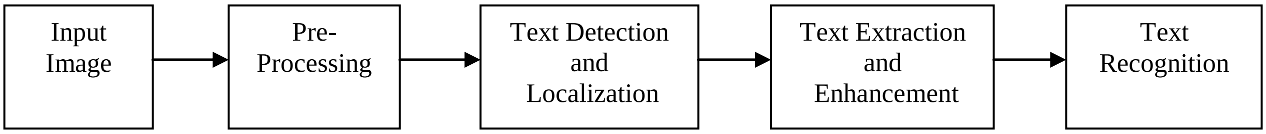

Input Image
Pre-Processing
Text Detection and Localization
Text Extraction and Enhancement
Text Recognition

Fig. 4 Block Diagram of Text Extraction

The input image may be gray scale or color, compressed on uncompressed format. Text detection refers to the determination of the presence of text in the image while text localization is the process of determining the location of text and generating bounding boxes around it. After that, text is extracted i.e. segmented from the background.

Enhancement of the extracted text is required as the text region usually has low-resolution and is prone to noise. Thereafter, the extracted text can be recognized using OCR. The block diagram of text extraction is shown in **Fig. 4**.

II. TEXT EXTRACTION TECHNIQUES

The various techniques of text extraction are as follow:

A. *Region based Method:*

Region-based method uses the properties of the color or gray scale in the text region or their differences to the corresponding properties of the background. They are based on the fact that there is very little variation of color within text and this color is sufficiently distinct from text's immediate background [20]. Text can be obtained by thresholding the image at intensity level in between the text color and that of its immediate background.

This method is not robust to complex background. This method is further divided into two sub-approaches: connected component (CC) and edge based.

i.) *CC based Method*:

CC-based methods use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image [9-12]. A geometrical analysis is required to merge the text components using the spatial arrangement of those components so as to filter out non-text components and the boundaries of the text regions are marked. This method locate locates text quickly but fails for complex background.

ii.) Edge based Method:

Edges are a reliable feature of text regardless of color/intensity, layout, orientations, etc. Edge based method is focused on high contrast between the text and the background [5,13-15]. The three distinguishing characteristics of text embedded in images that can be used for detecting text are edge strength, density and the orientation variance. Edge-based text extraction algorithm is a general-purpose method, which can quickly and effectively localize and extract the text from both document and indoor/ outdoor images. This method is not robust for handling large size text.

B. *Texture based Method:*

This method uses the fact that text in images have discrete textural properties that distinguish them from the background. The techniques based on Gabor filters, Wavelet, Fast fourier transform (FFT), spatial variance, etc are used to detect the textual properties of the text region in the image [16-19]. This method is able to detect the text in the complex background. The only drawback of this method is large computational complexity in texture classification stage.

C. *Morphological based Method:*

Mathematical morphology is a topological and geometrical based method for image analysis [16,17,20]. Morphological feature extraction techniques have been efficiently applied to character recognition and document analysis. It is used to extract important text contrast features from the processed images. These features are invariant against various geometrical image changes like translation, rotation, and scaling. Even after the lightning condition or text color is changed, the feature still can be maintained. This method works robustly under different image alterations.

III.PERFORMANCE ANALYSIS

A. Various parameters are used to analyze the performance of text extraction techniques and given as follow:

$$(i)\ \textit{Detection rate}\ (DR) = \frac{\textit{correct}\ \text{detected}\ \textit{text}}{\textit{ground truth text}}$$

$$(ii)\ \text{Precision}\ \textit{rate}\ (PR) = \frac{\textit{correct}\ \text{detected}}{\textit{correct}\ \text{detected} + \textit{false positive}}$$

$$(iii)\ \text{Recall } rate\ (RR) = \frac{correct\ \text{detected}}{correct\ \text{detected} + false\ negative}$$

$$(iv)\ False\ alarm\ rate\ (FAR) = \frac{no.\,of\ text\ blocks\ falsely\ \text{detected}}{total\ no.\,of\ text\ blocks}$$

B. *Comparative Analysis of Related Work:*

Many researches have been done on various text extraction techniques such as region based (CC based and edge based), texture based, morphological based or combination of these technique (i.e. hybrid approach). Researchers have used different type of images for their experimentation. The detailed analysis of text extraction techniques is shown in **Table 1**.

Table 1 Comparison of Various Text Extraction Techniques

Author, year	Technique Used	Images	Parameters	Remarks

Yao et al.[9], 2007	CC and Support Vector Machine (SVM)	Complex background images	PR=64%
			RR=60%
				Pixels of each character assumed to have similar color.

Lai et al. [13], 2008	Edge detection and K-means clustering	Signboard Images		Efficient for uneven illumination.

Zhang et al. [16], 2008	Discrete Wavelet Transform (DWT), k-means clustering, morphology Operations	Background images with different languages, fonts and sizes	DR= 94.5%,
			FAR= 13.6%
				Text character Color independent.

Song et al. [21], 2008	Histogram Projection and color based K-means clustering	Chinese text	PR=77.05%
			RR=75.63%
				K=3 gives best performance.

Dinh et al. [5], 2008	Edge detection and Histogram Projection	Signboard Texts		Low complexity algorithm.

Fan et al.[22], 2009	Stroke features and connected component	Caption text images	PR=95.2%
			RR= 94.5%
				Color information is not fully used.

Audithan et al.[17], 2009	Haar DWT, Morphological Dilation operator, logical AND operator, Dynamic thresholding	Document images	DR =94.8 %	Independent of contrast.
Angadi et al.[18], 2010	Discrete Cosine Transform and texture features extraction	Natural scene images	DR=96.6%	Inefficient when background in the image is more complex like trees, vehicles.

Anoual et al.[14], 2010	Edge detection, texture features, connected component analysis	Complex background images	PR=95%
			RR=89%
				Robust and effective.

Kumar et al [10], 2010	CC Analysis	ICDAR 2003 scene images	PR=90%
			RR=89%
				Capable of Multilingual Text extraction.

Hassanzadeh et al.[20], 2011	Morphological operator, Decision classifier	Logo detection in document images	PR=95.6%
			Accuracy=86.9 %
				A novel and fast method for logo detection.

Zaravi et al. [8],	DWT, Dynamic thresholding,	Colored book and	DR=91.20%	Robust to noise.

2011	Region of Interest (ROI)	journal cover sheets		
Zhang et al. [11], 2012	Edge Enhancement and CC	Web images and caption text images	DR=92.4%	Not sensitive to various types of background noises.

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.

Azadboni et al. [19], 2012	FFT Domain Filtering , SVM Classification, K-means clustering	Scene text images	DR= 98.10%	Text characters having uniform colour.
Anupama et al.[23], 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.
Raj et al. [12], 2014	CC based	Natural Scene Images (Devanagari text)	PR= 72.8%, RR=74.2 %	Fails for small slanted/curved text.

IV. CONCLUSION

In this paper, various techniques such as region based, edge based, connected component (CC) based, texture based, morphological based etc. have been discussed and a detailed comparison of these techniques on the basis of various parameters such as precision rate, recall rate, accuracy etc. has been done. Every approach has its own benefits and restrictions. Even though there are many numbers of algorithms, there is no single unified approach that fits for all the applications due to variation in font, size, alignment, complex background of text etc. It is concluded that texture based method can detect and localize text accurately even when images are noisy, complex background and low resolution.

REFERENCES

1. [1] 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.
2. [2] 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.
3. [3] 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.
4. [4] A.K. Jain, Y. Zhong, *Page Segmentation using Texture Analysis*, Pattern Recognition, Vol. 29, No. 5, Elsevier, 1996, pp. 743–770.
5. [5] 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.
6. [6] 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.
7. [7] D. Ghai, N. Jain, *Comparison of Various Text Extraction Techniques for Images- A Review*, International Journal of Graphics & Image Processing, Vol. 3, No. 3, 2013, pp. 210-218.
8. [8] D. Zaravi, H. Rostami, A. Malahzaheh, S.S Mortazavi, *Text Extraction using Wavelet Thresholding and New Projection Profile*, World Academy of Science, Engineering and Technology, Vol. 5, 2011, pp. 528-531.
9. [9] J.L. Yao, Y.Q. Wang, L.B. Weng, Y.P. Yang, *Locating Text Based on Connected Component And SVM*, International Conference On Wavelet Analysis And Pattern Recognition, Vol. 3, 2007, pp. 1418 - 1423.
10. [10]M. Kumar, Y.C. Kim and G.S. Lee, *Text Detection using Multilayer Separation in Real Scene Images,* 10th IEEE International Conference on Computer and Information Technology, 2010, pp. 1413-1417.
11. [11] Y. Zhang, C. Wang, B. Xiao, C. Shi, *A New Text Extraction Method Incorporating Local Information*, International Conference on Frontiers in Handwriting Recognition, 2012, pp. 252-255.
12. [12]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.
13. [13]A.N. Lai, G.S. Lee, *Binarization by Local K-means Clustering for Korean Text Extraction, IEEE Symposium on Signal Processing and Information Technology*, 2008, pp. 117-122.

14. [14]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.
15. [15]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.
16. [16]X.W. Zhang, X.B. Zheng, Z.J. Weng, *Text Extraction Algorithm Under Background Image Using Wavelet Transforms,* IEEE Proceedings of International Conference On Wavelet Analysis And Pattern Recognition, Vol. 1, 2008, pp. 200-204.
17. [17]S. Audithan, RM. Chandrasekaran, *Document Text Extraction from Document Images using Haar Discrete Wavelet Transform*, European Journal of Scientific Research, Vol. 36, No. 4, 2009, pp. 502-512.
18. [18]S. A. Angadi, M. M. Kodabagi, *Text Region Extraction from Low Resolution Natural Scene Images using Texture Features*, IEEE 2nd International Advance Computing Conference, 2010, pp. 121-128.
19. [19]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.
20. [20]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.
21. [21]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.
22. [22]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.
23. [23]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.