# Intelligent Text Extraction System for Complex Images

Rosy K. Philip, Gopu Darsan

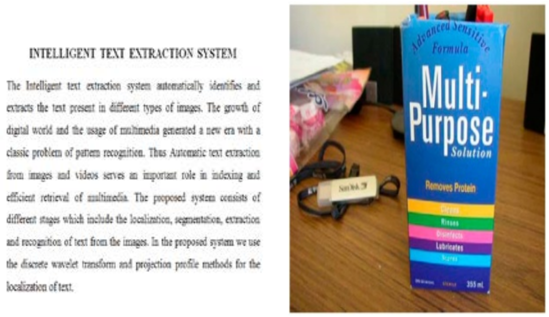
**Abstract—** The Intelligent text extraction system automatically identifies and extracts the text present in different types of images. The growth of digital world Detection and extraction of text regions in an image are well known problems in the area of image processing. The growth of digital world and the usage of multimedia generated a new era with a classic problem of pattern recognition. Thus Automatic text extraction from images and videos serves an important role in indexing and efficient retrieval of multimedia. The existing techniques such as region based , texture based techniques for the text extraction are not able to compact with all the applications of text extraction. The proposed Intelligent Text Extraction system automatically identifies and extracts the text present in different types of images. The system consists of different stages which include the localization, segmentation, extraction and recognition of text from the images. In the proposed system we use the discrete wavelet transform for the localization of text. The morphological operations are used which enhances the identification of correct text portions. The text part is segmented and is recognized using an efficient system. The advantage of the system is that the extracted text is shown in the .txt file. The proposed system also allows the modification of the recognized text from the image. This method shows better efficiency , precision and recall compared to the existing techniques. This shows the possibility of using this technique in more new and advanced applications.

**Index Terms**— DWT, image processing, morphological operations, Otsu, segmentation, text extraction, text recognition.

## I INTRODUCTION

T The rapid advancement in the technology and multimedia has digitalized the world. The availability of cameras and other systems contributes large number of images to the world. Ranging from cameras embedded in mobile phones to professional ones, Surveillance cameras to broadcast videos, every day images to satellite images, all these contributes to increase in multimedia data. Most of the images may contain text as part of it, which gives some information about that image. Therefore identification of these texts has relevance in many applications. This shows the importance of the text extraction system in lot of applications. It was stated that in recent years there was a drastic increase in multimedia libraries and the amount of data is growing exponentially with time. It is also known that there are number of television stations that broadcast everyday and the widespread of affordable digital cameras and inexpensive memory devices, the multimedia data is increasing every second.

The text present within an image enables applications such as keyword-based image search, text-based image indexing and automatic video logging. Extraction of these texts from images is a very difficult task due to variations in character fonts, styles, sizes and text directions, and presence of complex backgrounds and variable light conditions. Generally, the images can be categorized into three based on its type: document images, scene images and born-digital images. Document images are the image of the document which includes pdf, notes etc. The text present in document image is the document text. Figure 1(a) shows the document image. Born-digital images are the images generated by computer software. The text in these images is called as caption text or overlay text. Some researchers specify overlay text as superimposed text or artificial text. An example for overlay text is the text shown by the news channels etc.



INTELLIGENT TEXT EXTRACTION SYSTEM

The Intelligent text extraction system automatically identifies and extracts the text present in different types of images. The growth of digital world and the usage of multimedia generated a new era with a classic problem of pattern recognition. Thus Automatic text extraction from images and videos serves an important role in indexing and efficient retrieval of multimedia. The proposed system consists of different stages which include the localization, segmentation, extraction and recognition of text from the images. In the proposed system we use the discrete wavelet transform and projection profile methods for the localization of text.

(a) (b)

Figure1 (a) document image (b) scene text image

Scene images are the images that contain the text, such as the advertising boards, banners, which is captured naturally when the scene images are taken with the camera as shown in figure1(b), therefore scene text gets included in the background of image as a part of the scene taken. Compared with document images and born-digital images, the scene images, have more complex foreground/ background, low resolution, com-

- *Rosy K. Philip is currently pursuing Mtech in Computer Science and Engineering under University of Kerala, India, E-mail: rosykodiyattu@mail.com*
- *Gopu Darsan is currently working as assistant proffessor in Computer Science and Engineering department at Sree Buddha College of Engineering under University of Kerala,India. E-mail: gops601@mail.com*

pression loss, and severe edge softness. This makes the detection of text from scene text more difficult. Therefore automatic extraction of texts from images or video is a challenging task and research under this field is still under progress. Text present in images exhibit many variations based on the following properties:

(a)Size: The text can have variable size from small to large.

(b)Alignment: The characters in caption text may appear in clusters and sometimes they can also appear as non-planar texts. The scene texts possesses numerous perspective distortions .They may be aligned in any direction with geometric distortions

(c)Inter-character distance: Characters present in a text line have some uniform distance in between them.

(d)Color: In a simple image the characters in a text usually have the similar or same color. Connected component-based approach usually makes use of this property. But complex images or color documents usually contain text strings with two or more colors for effective visualization

(e)Edge: Most caption and scene texts are designed to be easily read, therefore strong edges are placed at the boundaries of text and background.

The remaining part of this paper is organized as follows. Section II describes the related works. Proposed system is explained in Section III and the experiments and results in section IV. Finally the conclusion is given in the Section V.

## II.RELATED WORKS

The different approaches related to text extraction includes region based methods, texture based methods. The region based methods are basically divided into two categories: edge based [1] and connected component based methods.

J.Gllavataet.al [3] proposed a connected component [2] based approach for the text extraction .It is based on color reduction technique and OCR is used for character recognition .It will only detect text with horizontal alignment. Low quality images will not be processed accurately. Zhong et al. [4] used a CC-based method, which uses color reduction. In that they quantize the color space by analyzing the color histogram generated in the RGB color space. This is done mainly based on the assumption that the text regions usually cluster together in this color space and they occupy a significant portion of an image. Each text component present will undergo filtering stage using a number of heuristics, such as area, spatial alignment and diameter. The performance of this system was evaluated with testing in CD images and other book cover images. Kim et al. [5] use transition map to detect overlay text. They proposed a method for overlay text detection and extraction from complex videos. The detection method is based on the observation regarding the existence of transient color between inserted text and its adjacent background. The transition map is generated first which is based on its logarithmical change of intensity and modified saturation. Then Linked mask is generated to create connected components for each candidate region and then each of these connected components is reshaped to have smooth boundaries. Researches based on Caption text detection are proposed in [9] .Xiaoqing Liu,et.al [6] proposed a method based on the properties of edges. This method is not sensitive to image color/intensity. Even though it can handle both printed and document images effectively. Since it mainly analyses texts in the form of blocks the small image regions and stroke are mis-identified as text in areas containing large characters.

The proposed system uses wavelet and different morphological operations in a different way compared to existing techniques for the identification of text part. So a better efficiency is obtained in the extraction of text. Additional features are also added in our system with the enhancement of technology.

## III.PROPOSED SYSTEM

The proposed method works based on the fact that the texts present in images have some unique features which include the properties of edges. The architecture of the proposed system is shown in Figure 2. The proposed system is mainly divided into three modules: Edge map generation module, Text area segmentation module, Text recognition module. The input is given as image to the system and the output obtained is in the form of a text file.

### 3.1. Edge map generation module

The input is passed to this module. The input image can be grayscale or color, compressed or uncompressed format. This module contains different steps in it. Firstly, the input image undergoes a preprocessing stage.

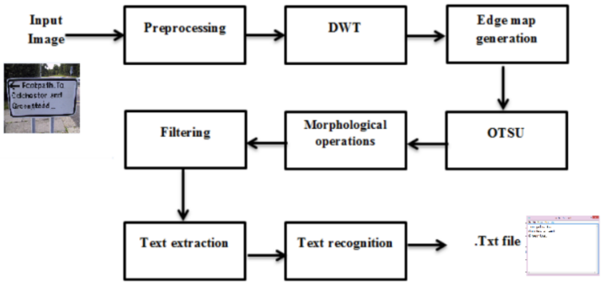


Figure 2: Architecture of the proposed system

### 3.1.1 Preprocessing

Edge strength and density are fundamental characteristics of that embedded text in a complex image. This serves as a vital feature for text extraction. In the proposed algorithm, the input taken is a color or a grayscale image. If the input is a color image it undergoes the preprocessing stage. In this stage , the color image is converted to grayscale image using the equation below.

$Y = 0.299R + 0.587G + 0.114B$

The RGB image is converted to Hue-Saturation-Value (HSV) color space. The Y of the above equation refers to value component of the Hue-Saturation- Value (HSV) color space. Thus RGB color image gets converted to the gray scale image. We can also process the image by converting it to YCbCr color space and taking the Y, the luminance part for processing. On applying a median filtering to this gray scale image the noise can be filtered out. The best known order statistics filter is the median filter. It replaces the value of a referred pixel by another value which is calculated by taking the median of the gray levels in the neighborhood of that pixel. Median filtering has excellent noise reduction capabilities with less blurring compared to linear smoothing filters of similar size.it is very effective in case of bipolar and unipolar impulsive noise. After this filtering step is done , most of noise present in the image gets removed while the edges in the image will be still preserved. The Image Y is then processed with the 2-D discrete wavelet transform.

### 3.1.2 Discrete Wavelet Transform

In this system, the proposed algorithm uses the Haar discrete wavelet trans- form. The Haar DWT provides a well powerful tool which models most of the characteristics of the image. The textured images are mostly well characterized by their edges. On applying Discrete Wavelet Transform (DWT), it is decomposed into components of frequency domain [8]. On applying the DWT the input image is decomposed to four sub-bands or components. i.e one average component and three detail components. To obtain the components it has to deal with row and column direction separately. First High Pass Filter (H.P.F) G and Low Pass Filter (L.P.F) H are exploited for each row data and then are down sampled by 2 to get high and low frequency components of the row . Next the high and low pass filters are applied again for each high and low frequency components of the column and then down sampled by 2 . By way of the above processing, the four-subband images are generated: HH1, HL1, LH1 andLL1.

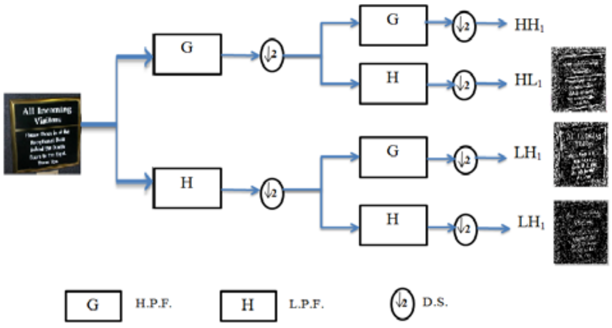


Figure 3: Block diagram of 2D DWT

The detail component sub-bands containing the vertical details LH1, horizontal details HL1 and diagonal details LL1 are used to detect text edges present in the original image. The process on applying DWT to image is shown in Figure 3. The D.S. represents the down sampling of the image by 2.

Since filtering is done before applying DWT, the effect of noise in the components can be reduced. Now edge detection method is applied at each component. The wavelet function and the scaling function of haar wavelets are defined below.

$$\psi(t) = \begin{cases} 1, 0 \le t < \frac{1}{2} \\ -1, \frac{1}{2} \le t < 1 \\ 0, otherwise \end{cases}$$

$$\phi(t) = \begin{cases} 1, 0 \le t < 1 \\ 0, otherwise \end{cases}$$

The Haar is simpler than other wavelets which reduces the complexity of the algorithm. The advantage of using Haar wavelet is that it is the only wavelet that allows perfect localization in the transform domain. Its coeffcients are either 1 or -1 and are real symmetric and orthogonal. On applying this, the portions with higher edge strength in identical directions can be found out. Thereafter the threshold is calculated. This helps to filter out the unfeasible edges. The second derivative of intensity is employed in the measurement of edge strength as it provides improved detection of intensity places which leads to a characterization of text in images that usually leads to a characterization of text in images. The traditional edge detection filters may also be able to provide the similar result but it will not be able to detect three kinds of edges at a time. Therefore, the processing time for the traditional edge detection filters works slower than 2-D DWT.

### 3.1.3 Text region detection

The process of identifying text regions can be split into two subproblems: detection and localization. In the detection step, general regions of the frame are classified as text or non-text. The size and shape of these regions differ in different algorithm. For example, some algorithms classify 8x8 pixel blocks, while others classify individual scan lines. In the localization step, the results of detection are grouped together to form one or more text instances. This is usually represented as a bounding box around each text instance. For the detection of the text region, the three detailed sub components which is obtained on applying Haar DWT is used . Sobel edge detection algorithm is applied to each detaied component. Thus hc,vc,dc are images obtained after edge detection in horizontal ,vertical and diagonal component. Therefore hc contains all the edges in horizontal direction,vc contains all the edges in vertical direction and the dc contains all the edges in diagonal direction . Now an edge map is generated using hc, vc and dc. It is generated by using weighted OR operation. The equation used for the generation of edge map is shown below

I = (40 * hc + 70 * vc + 30 * dc)

The I is the image obtained after appling the edgemap. hc, vc and dc indicates the horizontal, vertical and diagonal sub component respectively after sobel edge detection algorithm is applied. The edge map of an example image is shown in Figure 4 .This algorithm uses the sobel edge detector as it is more effcient to locate the strong edge present. Now the image I shows edges with all the possible text regions.

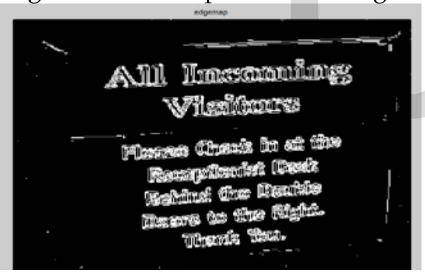


Figure 4 :Edgemap of the image obtained

### 3.2 Text area segmentation module

After edge map is obtained , the binirization of the image is done. The binary form of the edge map is obtained by the thresholding operation. For binarization of the image the Otsu algorithm is used. The thresholding operation removes the non text regions identified so far. The steps involved in computing Otsu threshold are

- Reshape the image to 1 dimensional.
- Compute histogram and values at each intensity level
- Initialize a matrix with values from 0 to 255
- Step through all possible thresholds maximum intensity
- Compute the mean , weight and the variance for the foreground and background.

Varience: $\sigma_b{}^2(t)=\sum_{i=1}^{n} p_i.(x_i-\mu)^2$

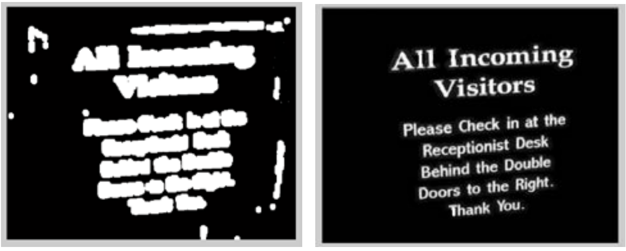
Weight: $\omega(t)=\sum_{i=0}^{t} p_i.$

Mean : $\mu(t)=[\sum_{i=0}^{t} p\ (i)x(i)]/\omega_1$

- Calculate weight of the foreground * variance of the foreground + weight of the background* variance of the background.
- Find the minimum value.

The minimum value calculated is taken as the threshold for the binarization process. The localization process of the text involves further enhancement the text regions by eliminating non-text regions from the image. One of the main property the text exhibits is that all characters present will appear close to each other in the image. Therefore it can form a cluster. By taking into account this property we use morphological operations. The possible text pixels can be clustered together by using the dilation operation, thus eliminating pixels that are far from the candidate text regions are possible. Dilation can be defined as an operation that expands or enhances the region of interest, by the usage of the structural element of the required shape and size. Large structuring element is used in the dilation process so that regions which lie close to each other can be enhanced. To localize the text part clearly we use the morphological operations. The morphological operations include the erosion, dilation with line and disk as the structuring element is done. The segmentation of the identified text portions are done in this segmentation phase .For that the connected components are labeled and the connectivity used here is 8. The set of properties, shape and measurements of the connected components are computed. The area and bounding box shape measurements as per the requirement are only considered. The area of the connected component is considered as scalar as it represents the number of pixels in that region.

Bounding box is placed to the connected component identified and it must be the smallest containing the required text region[11]. The width, height and the upper left corner position are identified. A new value can be computed by multiplying height and width of bounding box. The ratio of this new value and area is taken. If the ratio is less than 1.5, then the regions so obtained are considered as text regions. The resultant image got after dilation operation may even consist of some non-text regions or any noise which are needed to be eliminated. To eliminate noise blobs present in the image an area filtering is done. After that only those regions in the final image whose area is greater than or equal to 1/15 of the maximum area region identified are retained. The Figure 5(a) shows the image after morphological operations and Figure 5(b) gives the image after text extraction.



(a) (b)

Figure:5(a)image after morphological operations(b)output image after segmentation

### 3.3. Text Recognition module

The image got from the previous stage is considered as input in this phase. Text recognition is done by using the tesseract OCR. The binary image with polygonal text regions defined is given as input to the tesseract. The processing of the system follows a traditional pipeline. The first step is analysis of connected component. It outlines the components stored. This is a computationally expensive design decision, but it has a significant advantage. At this stage, outlines are gathered together, by nesting, into Blobs. The Blobs thus obtained are organized into text lines. Then the lines and the regions identified are analyzed for getting fixed pitch or proportional text. Now the obtained text lines are divided into words in accordance with the character spacing. Fixed pitch text is then chopped immediately by using the character cells. Proportional text after that is broken down into words by using definite spaces. Recognition process here undergoes a two-pass process. In the first pass, an attempt is made to recognize each word. Every word that is satisfactory recognized is then passed to an adaptive classifier as a training data. The adaptive classifier then will more accurately recognize text in the page. With the completion of first pass , a second pass is then executed over the page .In this the words that were not recognized in the previous pass well enough are recognized again. A final phase resolves spaces, and locates the small text by checking the alternative hypotheses for the x-height to it . After recognition the UTF-8 code of the character are returned. This can be easily converted corresponding characters and are displayed as output in the form of a text file.

## IV.EXPERIMENT AND RESULTS

The system is experimented with large number of images and the results obtained are described here.

Table 1 Comparisons with existing methods

| Method | Precision rate (%) | Recall rate (%) |
|---|---|---|
| Proposed method | 99.65 | 99.8 |
| Samarabandhu et. al[6] | 91.8 | 96.6 |
| J. yang et. al[11] | 84.90 | 90.0 |
| K.C.kim et al[10] | 63.7 | 82.8 |
| J. Gllavata et.al[3] | 83.9 | 88.7 |

It is clear from the Table 1 that the proposed method have better precision rate and recall rate compared to other existing techniques. The method proposed by Samarabandhu et.al shows 91.8 % of precision and 96.6% of recall whereas the technique given by j.yang et. al have lesser rate of about 84.90% and 90.0 % precision and recall. But our method shows 99.65% of precision rate and 99.8% of recall. The other existing methods proposed by Kim et.al and J.Gllavata exhibits even smaller rate as given in the Table 1.

The accuracy of the algorithm in the proposed method is computed by counting the number of correctly located characters, which is taken as the ground truth. The precision and recall rate are calculated by the equation

$$precisionrate = \frac{correctlylocated}{correctlylocated + falsenegative} * 100\%$$

$$recallrate = \frac{correctlylocated}{correctlylocated - falsepositive} * 100\%$$

Comparisons with some existing methods are shown in table 1 which shows a clear improvement over existing methods. The performances of other methods shown are cited from the published works. The proposed system is implemented using matlab[13]

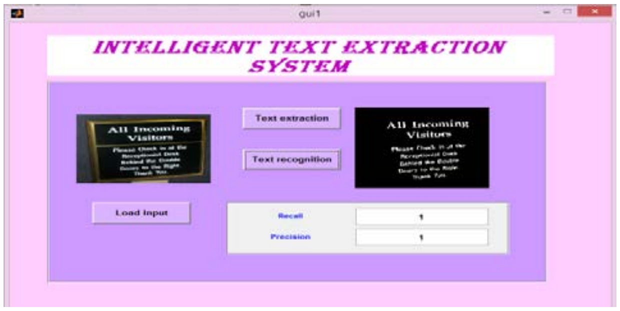
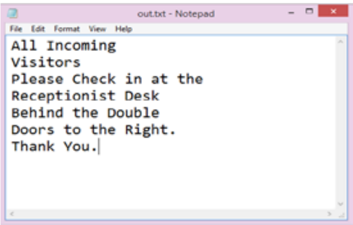


Figure 6:The GUI showing the input image and the image with the text portion

The screenshot of GUI of the intelligent text extraction system is presented in figure 6



```
All Incoming
Visitors
Please Check in at the
Receptionist Desk
Behind the Double
Doors to the Right.
Thank You.
```

Figure 7: The output after text recognition

The output obtained was shown in a notepad as illustrated in figure 7.

ADVANTAGES

The system is not sensitive to color or intensity. It is robust with respect to font size, style, perspective, reflection and uneven illumination. The analysis of text in blocks makes the system computationally efficient. Identifies text regions from textures like regions including wall , door , window patterns etc . Higher recall and precision rate.Output in the form of text file.Text extracted from the image can be modified and can be used in many applications.

## V .CONCLUSION AND FUTURE WORK

In this paper, a relatively simple, fast and effective method for text detection and extraction are proposed. The method uses DWT for the efficient working of the algorithm. This process requires less processing time which is mainly essential for real time applications and shows high precision rate. Most of methods fail when the characters are not aligned well or when the characters are too small. Those methods also result in some of missing characters when the characters have very poor contrast with respect to the background. But the proposed method is not sensitive to color or intensity of image, and also the uneven illumination and reflection effects. This can be used in large variety of application fields such as vehicle license plate detection, object identification, identification of various parts in industrial automation, mobile robot navigation which helps to detect text based land marks, analysis of technical papers with the help of maps, charts, and electric circuits etc. This algorithm is good at handling both scene text images and documents images effectively. Even though there are large numbers of algorithms in this area, it is observed that there is no single unified approach or algorithm that fits for all the applications. Further the work can be enhanced by using inpainting along with this system.Thus the text part can also be removed without affecting the image.

## REFERENCES

[1]Xin Zhang, Fuchun Sun, Lei Gu “A Combined Algorithm for Video Text Extraction” *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*,2010.

[2]J. Ohya, A. Shio, and S. Akamatsu, “Recognizing Characters in Scene Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994, 214-224.

[3]. Gllavata, R. Ewerth, and B. Freisleben, “A robust algorithm for text detection in images” *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, pp.611– 616, ISPA, 2003.

[4] Zhong, Yu., Karu, K., and Jain, A.K.” Locating text in complex color images*” Proceedings of the Third International Conference on Document Analysis and Recognition*1995, 1, 14-16: 146-149.

[5]Wonjun Kim,Changick Kim, “A New Approach for Overlay Text Detection and Extraction From Complex Video Scene,” *IEEE Transactions on Image Processing*, V.18 , No.2, pp. 401 – 411, 2009.

[6] Xiaoqing Liu, Jagath Samarabandu “Multiscale Edge-Based Text Extraction from Complex Images,” *International Conference on Multimedia and Expo*, pp.1721-1724, 2006

[7] Xiao-Wei Zhang, Xiong-Bo Zheng, Zhi-Juan Weng, “Text Localization Algorithm Under Background Image Using Wavelet Transforms”, *Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition,* Hong Kong, pp.30-31, Aug. 2008.

[8]N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Transactions on Systems, man and Cybernet,* 1979.

[9] Debapratim Sarkar, Raghunath Ghosh ,”A Bottom-Up Approach of Line Segmentation from Handwritten Text”2009.

[10] K.C. Kim, H.R. Byun, Y.J. Song, Y.W. Choi, S.Y. Chi,K.K. Kim and Y.K Chung, Scene Text Extraction in Natural Scene Images using Hierarchical FeatureCombining and verification, *Proceedings of the 17International Conference on Pattern Recognition (ICPR'04),IEEE.*

[11] J. Yang, J. Gao, Y. Zhang, X. Chen and A. Waibel, “AnAutomatic Sign Recognition and Translation System”,*Proceedings of the Workshop on Perceptive User Interfaces(PUI'01), 2001*, pp. 1-8.

[12] A.K Jain,”Fundamentals of Digital Image Processing”,Englewood cliff, NJ: Prentice Hall, 1989,

[13] R C.Gonzalez,”*Digital Image Processing Using MATLAB*