American Journal of Engineering Research (AJER) 2015
American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-4, Issue-11, pp-155-162
www.ajer.org
Research Paper Open Access

Text Detection in Video Using Haar Wavelet Transformation and Morphological Operator

Dinesh AnnajiKene[1], Dr. D. J. Pete[2]
Department of Electronics and Telecommunication Engineering.
DattaMeghe College of Engineering, Airoli, Navi Mumbai – 400708.
University of Mumbai. India

***ABSTRACT*:** *This paper presents simple and efficient method for text detection, extraction and localization from video or static images using Haar wavelet and Morphological operator. Haar wavelet transform have its coefficients either 1 or -1 , so that the operation speed of Haar wavelet transformation is fastest among all wavelets. The sub bands contain both text edges and non-text edges however the intensity of text edges is different that of the non-text edges. Instead of using Canny operator we used Sobal operator for edge detection because Sobal operator detect more edges than Canny operator when there is text information. Morphological operators are applied to edit or smoothing out the text region. Then detected text regions are further decomposed into character level. Then using some refinement the final text region are obtained.*

Keywords *: CC Based method, Texture based method, Edge based method, optical character recognition (OCR)*

I. INTRODUCTION

A brief history of Video Text detection:- The origin of video text detection research may be traced to document image analysis because of its core techniques in optical character recognition (OCR) . Patent on OCR for reading aids for the blind and for input to the telegraph was filed in the Nineteen century and working models were demonstrated by 1916 .however the field of document image analysis itself is only about 40 years old, OCR on specifically designed printed digits (OCR fonts) was first used in business in the 1950 s . The first postal; address reader and the social security administration machine to read typewritten earnings reports were installed in 1965. Devices to read typeset material and simple hand printed forms came in to their own in 1980 s, when the prices of OCR systems dropped by a factor of a 10 due to the advent of microprocessors bit mapped display and solid state scanners. The study of automatic text detection started about three decade ago. In the 1980-95, a lot of methods were proposed.[1]

Video content can be seen as two major categories, "Perceptual content" and "Semantic content". Perceptual content includes attributes such as colour intensity, shape, texture and their temporal changes, While semantic content includes object, events and their relations. Among them text in images / videos is of particular interest as 1) It is very useful for understanding the content of an image 2) It is very easy to extract compared to other semantic content and 3) It is useful for application such as text based image search and automatic video logging. The semantic gap refers to the different interpretation and meaning of the extracted low level feature (e.g. texture, colour, shape etc.) by the uses with respect to applications. Text in video can be used to fill this gap because in many videos there is considerable amount of text , Video text can thus help to improve the effectiveness of content based image and video retrieval system [1]

Compared with the text in a typical document the text in a video frame is in much smaller quality. However these texts in video often give crucial information about media contents. They usually appears in the form of names , location , products , brands , score of the match, date & time etc. which are helpful information to understand the video contents. The text in images and videos can be superimposed on arbitrary background or embedded on the surfaces of the object in the scene with varying fonts size, colour, alignment, movement and lighting condition. Hence text detection and extraction are extremely difficult.

A large number of approaches such as Connected Components based, Texture based and other methods have been proposed. Some of these methods have already obtained impressive performances. A majority of text detection and extraction approaches in the literature are developed based on scanned document nature, although most of them can be adopted for video images detection and extraction. Text in video presents unique challenges over that in document images, due to many undesirable properties of video for text detection and extraction problem. Fortunately text in video usually persists for at least several seconds to give human viewer sufficient time to read it. This temporal nature of video is very valuable and can be well utilize for text detection, extraction and recognition in video.

II. RELATED WORK

In order to show that the proposed method is effective and robust in terms of metrics and multi-oriented text lines detection, we have studied variety of related works. In the research book by PalaiahnakoteShivkumara, Chew Lim Tan, Wenyin Liu and Tong Lu," Video Text Detection" A Research book published by Springer-Verlag London 2014, shows the variety of methods, operators and models for image as well as video text detection. In that book he analyse the minute details of Text Detection in Video and also shows some application oriented methods.[1] P. Shivkumara, TrungQuePhan and Chew Lim Tan in their paper New Fourier Statistical Features in *RGB* Space for Video Text Detection proposed Fourier statistical method for color images , text to convert in gray and then again in color for detection and extraction [2].
As well PlaiahnakoteShivkumara, TrungQuyPhan et al. represented gradient vector flow and grouping based method. [3].

P. Shivkumara, T.Q. Phan and C.L. Tan [3] in their paper A Laplacian Approach to Multi –Oriented Text Detection in Video represented a Laplacian method for multi-oriented text detection [4]. In which it shows that the performance of these methods is low because scene text poses many challenges compared to graphics text. Multi-oriented text has only been partially addressed where the algorithm is limited to caption text and a few selected directions. Recently, addressed this multi-oriented issue in based on Laplacian and skeletonization methods. However, this method still has room for improvement in the following respects: Recall, Precision and F-Measure. [4]

Chung-Wei Liang et.al. used DWT based text localization method in his paper," DWT Based Text Localization". In which he used Haar wavelet transformation and Morphological operator along with AND operator.[5]

Neha Gupta et. Al. also used Haar Wavelet and Morphological operator in their paper," Localization of Text in Complex Images using Haarwavlet transformation". For the experimental result they used complex background images as input.[6]

M. Naguet. Al. used Neural network along with wavelet and Morphological operator in their research paper, "Morphological Text Localization using Wavelet and Neural network". Their paper implemented Neural networks for recognition of text character from isolated text images and make it editable. [7]

Narsimha Murthy K N et. Al. in his research paper," A Novel Method for Efficient Text Extraction from Real Time Images with Diversified Background using Haar Discrete Wavelet Transform and K-Means Clustering" shows that they used real time images for data input and get efficient text output using K-means clustering[8].

III. PROPOSED SYSTEM

In order to show that the proposed method is effective and robust in terms of metrics and multi-oriented text lines detection, we have considered a variety of datasets which include images and videos of sports news, low contrast, different fonts and size, different orientation etc.

Methodology

Many efforts have been made earlier to address the problems of text area detection, text segmentation and text recognition. Current text detection approaches can be classified into three categories:

The first category is connected component-based method, which can locate text quickly but have difficulties when text is embedded in complex background or touches other graphical objects.

The second category is texture-based, which is hard to find accurate boundaries of text areas and usually yields many false alarms in "text-like" background texture areas.

The third category is edge-based method. Generally, analyzing the projection profiles of edge intensity maps can decompose text regions and can efficiently predict the text data from a given video image clip.

Text region usually have a special texture because they consist of identical character components. These components contrast the background and have a periodic horizontal intensity variation due to the horizontal alignment of many characters. As a result, text regions can be segmented using texture feature.

3.1 Video Preprocessing

Automatic video text detection is extremely difficult so we need video preprocessing to reduce the complexity of the succeeding steps probability consisting of video text detection, localization, extraction, tracking, reconstruction and recognition.

3.2 Image Segmentation

In video image segmentation, after inputting an original videos frames image segmentation techniques are employed to extract all the pixels belonging to text. The aim of segmentation is to divide the pixels of each frame from a video clip in to two classes of regions, which do not contain text and region which contain text. A lot of techniques have been developed to perceive images, which in general consist of three layers (as shown in figure) of so called Image Engineering (I. E.), processing (low layer), analysis (middle layer) and understanding (high level). The objective of segmentation is to partition an image in to separated region which ideally composed to interested object and accordingly obtain a more meaningful representation of image. There are some general rules to be followed.

1. They should be uniform and homogenous with respect to specific characteristics.
2. Their interior should be simple enough and without too many small holes.
3. Adjacent regions should have significantly different values with respect to the characteristics on which they are uniform.
4. Boundaries of every segment should be simple, not ragged and must be spatially accurate.

Each image can be segmented in to multiple separated regions using two strategies namely, Bottom-Up & Top-Down segmentation.

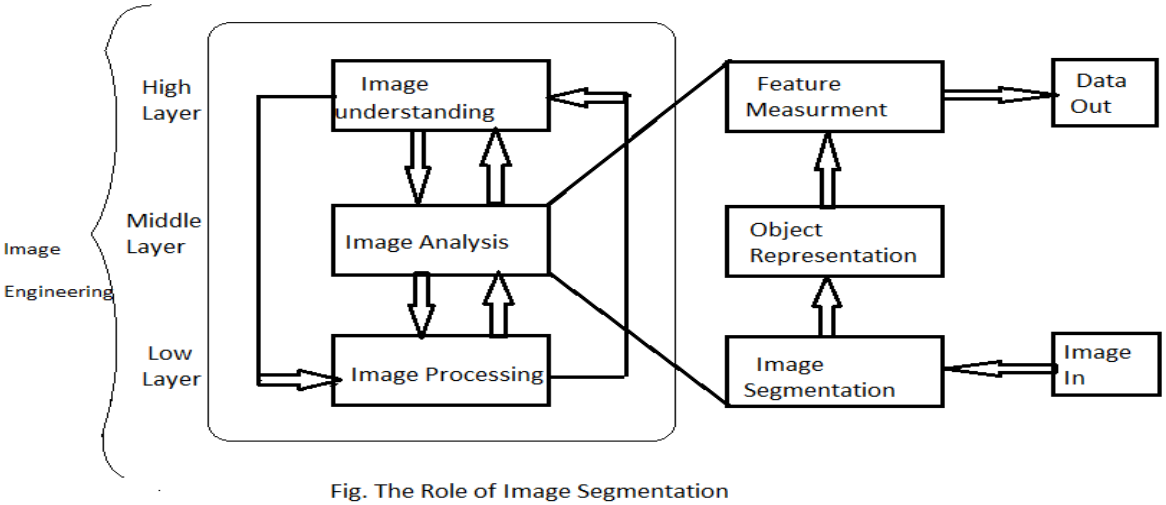


Fig. The Role of Image Segmentation

3.3 Pre-processing Operators

The aim of preprocessing operations for video text detection is an improvement of the image data by suppressing undesired degradation and simultaneously enhancing specific text relevant features. There are several preprocessing operators that are either enhance the image feature or suppress the information which is not relevant to video text detection.

a) *Image Cropping and Local operator* :
These are useful for brightness adjustment & contrast enhancement. Commonly used local operators are Multiplication &Addition, Another commonly used operator is Linear Blend operator, which is designed to cross-dissolve two video frames.

b) Neighborhood operator :

These operators are used to remove noise and sharpness the image detail. The neighborhood operators are

1. Linear Filter Operator
   i) Average Filter
   ii) Median Filter
2. Derivative operator
   i) Robert operator
   ii) Sobel operator
   iii) Laplacian operator

c) Morphological operator :

Mathematical Morphology is a tool for extracting image component that are useful in the representation and descriptive of region shape, such as boundaries, skeleton and the convex hull. It is defined fundamental Morphological operation such as Dilation, Erosion, Majority, Opening and Closing to perform such operations. We often need, first convert colorful video frame in to a binary image which has only two elements of either a real foreground or a background complexity with respect to the others.

d) *Color Based Preprocessing* **:**

A color image is built of several color channels, each of them indicating the value of the corresponding channel. For example, an RGB image consist of three primary color channels of Red , Green and Blue, An HIS model has three intuitive color characteristics as Hue, Saturation and Intensity. These color representation are designed for specific devices and have no accurate definition for a human observer. Therefore color transformation is required to convert the representation of color to gray scale representation.

If the input image is a gray-level image, the image is processed directly starting at discrete wavelet transform. If the input image is colored, then its RGB components are combined to give an intensity image. The pictures are often in the Red-Green-Blue color space. Intensity image Y is given by:

$$Y= 0.299R+0.587G+0.114B \quad (1)$$

Image Y is then processed with 2-D discrete wavelet transform. The Y is actually value component of the Hue-Saturation-Intensity (HSI) color space. In stage the color image RGB is converted in to HSI color space, then value component is extracted from HSI color space using above expression.

The RGB to HSI transformation is

$$H= \arctan(\beta/\alpha) \quad (2)$$

$$S= \sqrt{\alpha 2 + \beta 2} \quad (3)$$

$$I = (R+G+B)/3 \quad (4)$$

Where $\alpha = R-1/2\,(G + B)$, $\beta = \sqrt{3}/2\,(G - B)$

The noise of image is reduced by using a medium filtering that is applied on the gray scale image. After this filtering a major part of noise will be removed while the edges in the image are still preserved.

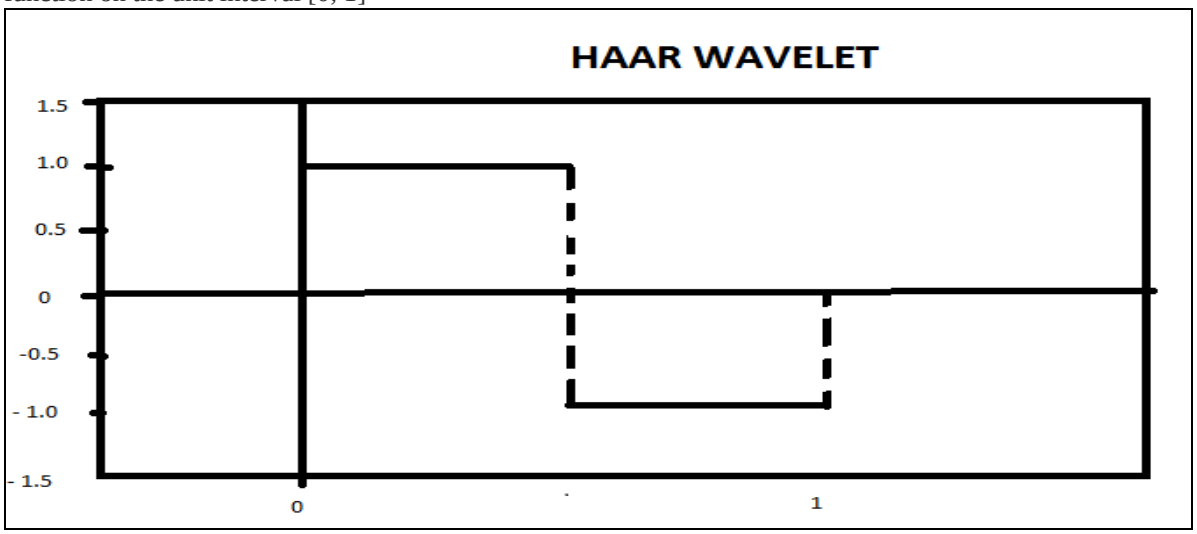
IV. WAVELET TRANSFORMATION

Digital image is represented as a two-dimensional array of coefficients, each coefficient representing the intensity level at that coordinate. Most natural images have smooth color variations, with the fine details being represented as sharp edges in between the smooth variations. Technically, the smooth variations in color can be termed as low frequency variations, and the sharp variations as high frequency variations.

The low frequency components (smooth variations) constitute the base of an image, and the high frequency components (the edges which give the details) add upon them to refine the image, thereby giving a detailed image. Hence, the smooth variations are more important than the details. Separating the smooth variations and details of the image can be performed in many ways. One way is the decomposition of the image using the discrete wavelet transform. Digital image compression is based on the ideas of sub-band decomposition or discrete wavelet transforms. Wavelets, which refer to a set of basic functions, are defined recursively from a set of scaling coefficients and scaling functions. The DWT is defined using these scaling functions and can be used to analyze digital images with superior performance than classical short-time Fourier-based techniques, such as the DCT. Shivkumar et.al. used 2-D Haar wavelet decomposition to directly detect video text based on three high frequency sub-band images of LH, HL and HH .after computing features for pixels, K means algorithm is applied to classify the feature vectors in to two cluster of background and text candidates.

V. HAAR WAVELET

The Haar Wavelet is a sequence of rescaled "Square Shapes " functions which together form a wavelet family or basis. Wavelet analysis is similar to Fourier analysis in that it allows a target function over an interval to be represented in terms of an orthogonal function basis. The Haar sequence is now recognized as the first known wavelet basis and extensively used as a teaching example. The Haar sequence was proposed in 1909 by Alfred Haar. Haar used these function to give an example of an orthogonal system for the space of square integrable function on the unit interval [0, 1]



The Haar wavelet's mother wavelet function **ψ (t)** can be described as

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

Its scaling function **ϕ (t)** can be described as

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

A. HaarTransform :

The Haar transform is the simplest of the wavelet transform. This transform cross multiplies a function against the Haar wavelet with various shift and stretches like the Fourier transform cross multiplies a function against a sine wave with two phases and many stretches.

The Haar transform is derived from the Haar matrix. An example of a 4 x 4 Haar transformation matrix is shown below.

$$H4 = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -\sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{2} & \sqrt{2} \end{pmatrix}$$

Property

The Haar transform has following properties

1) No need for multiplications. It require only addition and there are many element with zero value in the Haar matrix, so the composition time is short , it is faster than Walsh transform whose matrix is composed of + 1 and – 1.
2) Input and output length are the same . however the length should be a power of 2
   i.e. $N = 2^K$, $K \in N$
3) It can be used to analysis the localized feature of signals , Due to the orthogonal property of the Haar function the frequency components of input signal can be analyzed.

VI. MORPHOLOGICAL OPERATION

Mathematical morphology is a tool for extracting image components that are useful in the representation and descriptive of region shape, such as boundaries, skeletons and the convex hull. It is defined two fundamental morphological operations, dilation and erosion, in terms of the union or intersection of an image with a translated shape called as structuring element. To perform such operations we often need first convert a colorful image which has only two elements of either a real foreground or a background complementary with respect to the other. Morphological operations often take a binary image and structuring element as input and combine them using a set operator (Insertion, Union, Inclusion, Complement). They process object in the input image based on characteristics of its shape which are encoded in the structuring element.

Usually the structuring element is sized 3 X 3 and has its origin at the center pixel. It is shifted over the image and at each pixel of the image its elements are compared with the set of the underlying pixels. If the two sets of elements match the condition defined by the set operator, the pixels underneath the origin of the structuring element is set to a pre-defined value (0 or 1). A Morphological operator is therefore defined by its structuring element and the applied set operator.

VII. MOTION ANALYSIS

As a preprocessing technique motion vector analysis from temporal frames is believed helpful for video text detection. It produces time dependent information about the motion of both the foreground object and their background to provide hints for identifying text from consecutive video frames. More ever it potentially play a key role in detecting and tracking video text from multiple directions or with occlusions.

The major goals of motion analysis are to refine the text detection in terms of removing false alarms in individual frames interpolating the location of accidently missed text character and words in individual frames and temporarily localizing the beginning and end of eachword as well as its spatial localization within each frame. Their method proposed for motion analysis is based on the comparison of region in successive frames and includes five typical steps.

1. *Text Tracking*: this step is responsible for tracking the already detected text along with the frames that constitute each text sequence targeting the formation of temporally related chains of characters. Each

character chain here represents the same characters during its existence in a collection of similar regions occurring in several continuous frames. Every time a character region is detected for the first time, a position is stored and a signature is computed for that character region by using the features of luminance, size and shape. Each frame contributes with one region classified as a character chain.

2. *Text Integration* :- this step groups the character chains in order to form word based on the special and temporal analysis of each character chain. Three temporal elements are adopted a) Temporal coexistence b) Duration and c) Motion. The chain not include in words at this phase are considered as noise and are discarded.
3. *Character Recovery* ;- Video temporal redundancy is explored to complete the words with missing character as least for some frames e.g. due to noise or too textured background. In order to complete the words they are extended to the size of their biggest chain of character and the characters missing in the chains are recovered by means of temporal interpolation with motion composition. Thus by using temporal redundancy, the text detection for each frame is improved by completing the words with missing character for some frames
4. *Elimination of overlapped words*:- It is important to improve the performance for scene text usually, overlapping of words occurs when false words are detected e.g. due to shadow or three dimensional text. Every time two or more words overlap more precisely their bounding boxes overlap at least for one frame, the words with smaller area are discarded.
5. *Text Rotation* :- This steps perform the rotation of the text to the horizontal position in order to be recognized OCR system.

VIII. RESULTS

The results of the work of first stage Localizer are computed in the following manner:

1. Created dataset with various Text images and videos
2. Colour images and frames of videos are converted in to Gray scale
3. Texture of text area analysed by Haar wavelet transformation
4. Localized the text region or segments using Connected Component method (Bottom Up approach) and Sobal operator.
5. Used Morphological operator to edit the localized text region.
6. Resize the image .
7. Adjust Gray colour at extracting place with neighbourscolour.
8. Label the connected component region.
9. Apply the bounding boxes.
10. Convert the image again back from gray to RGB scale by comparing colours of the starting image.
11. Display the result.

The results of the work of second stage Localizer are computed in the following manner:

1. After 1st localizer output, the detected text region further decomposed in character level using Connected Component and Morphological operator.
2. Applied the bounding box.
3. Display the result.

The results of the work of Third stage Localizer are computed in the following manner.

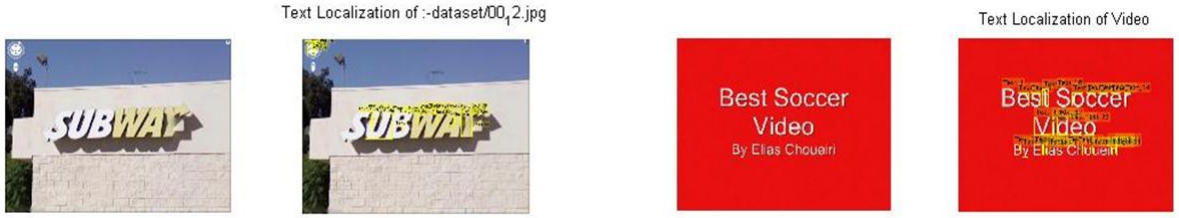
1. Eliminate the lines of bonding box.
2. Resize the text region using Connected Component major axis lenth& minor axis length, pixel index list.
3. Used again Morphological operator to edit the detected text .
4. Label the text region.
5. Display the result

Result of Localizer 1

Result of image 1.1Result of Video frame 1.1

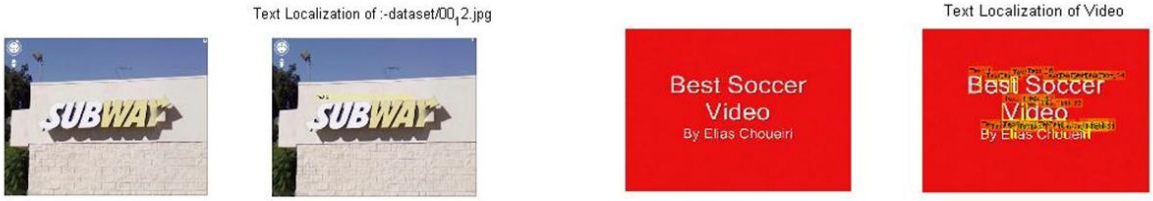
Result of Localizer 2

Result of image 2.1 Result of Video frame 2.1



Result of Localizer 3

Result of image 3.1Result of Video frame 3.1



IX. CONCLUSION & FUTURE SCOPE

In this work, we present a relatively simple and effective algorithm for text detection and extraction in videos by applying DWT to the images. As it requires less processing time which is essential for real time application. Mostly all the previous methods fail when the character are not aligned well or when the characters are too small. They also results in some missing character when the character have very poor contrast with respect to the background. But this algorithm is not sensitive to image color or intensity, uneven illumination and reflection effects. This algorithm can be used in large variety of application fields such as vehicle license plate detection to detect number plate of vehicle [5], mobile robot navigation to text based land marks, object identification of various parts in industrial automation, analysis of technical papers with the help of charts, maps, and electrical circuits etc. this algorithm can handle both scene text images and printed documents. In future work we plan to exploit the colour homogeneity of text in video, temporal text detection from frame to frame, Multi-frame integration for image enhancement.

X. REFERENCES

[1] PalaiahnakoteShivkumara, Chew Lim Tan, Wenyin Liu and Tong Lu," Video Text Detection" *A Research book published by Springer- Verlag London* 2014.

[2] P. Shivakumara, TrungQuyPhan and Chew Lim Tan, "New Fourier-Statistical Features in RGB Space for Video Text Detection", *IEEE Transactions on CSVT*, VOL. 20, NO. 11. pp 1520-1532., November 2010

[3] PalaiahnakoteShivkumara and TrungQuyPhan,"Gradient Vector Flow and Grouping-based Method for Arbitrarily Oriented Scene Text Detection in Video Images". *IEEE Transaction on circuits and systems for video technology*, Vol.23, No.10, pp1729-1738.October 2013

[4] P. Shivakumara, T, Q. Phan and C. L. Tan, "A Laplacian Approach to Multi-Oriented Text Detection in Video", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.33, No.2, pp 412-419. February 2011

[5] Chung-Wei Liang and Po-Yueh Chen, "DWT Based Text Localization'', *International Journel of Applied Science and Engineering*, pp.105-116,2004

[6] Neha Gupta and V. K. Banga," Localization of Text in Complex Images using Haarwavlet transformation" *International Journal of Innovative Technology and Exploring Engineering(IJITEE)*, Vol-1, Issue-6, pp 111-115, November 2012

[7] M. Nagu,B. Raja Rao,"Morphological Text Localization using Wavelet and Neural network". *International Journal of Computer Science and Information Technology(IJCSIT)*, Vol-2(2), Issue-6, pp 891-896, 2011

[8] Narsimha Murthy K N et. al.," A Novel Method for Efficient Text Extraction from Real Time Images with Diversified Background using Haar Discrete Wavelet Transform and K-Means Clustering" *International Journal of Computer Science issues (IJCSI)*, Vol-8, Issue-5, pp 235-245, 2011