RESEARCHERID
THOMSON REUTERS

[Kumari, 4(10): October 2017]
DOI- 10.5281/zenodo.1005109

ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

WEBER'S LOCAL DESCRIPTOR AND DIFFERENTIAL EXCITATION DIFFERENCE BASED TEXT DETECTION TECHNIQUE: AN INTEGRATED APPROACH

M. Sharmila Kumari*

*Professor, Department of Computer Science and Engineering, P A College of Engineering, Mangalore

ABSTRACT

In this paper, we have proposed a robust and an efficient text detector in a video frame. The proposed model is based on Weber's excitation principle and difference of excitation information is used to highlight the textual data in a video frame. In order to exhibit the performance of the proposed model, we have conducted experimentation on the standard ICDAR-2003 dataset and our own video database. Experimental comparison and objective comparison is also provided with the recently proposed models.

Keywords*: Weber Descriptor, Local Descriptor, Text Detection*

I. INTRODUCTION

Several models have been proposed by many researchers to detect text from video frames which are broadly classified as bottom-up models [1, 14]; top-down models [12, 20] and machine learning based models [3, 19]. In the bottom-up based approaches, probable character regions are identified in a frame and then groups character regions into words by using geometrical constraints such as the size of the region, height and width ratio. In the second category, character attributes such as edge orientation, texture values, spatial information is used to localize texts. The machine learning based models use statistical measures [7, 13], multi-scale analysis [3] of video frame to detect and localize the textual information in video frames and argued that these categories of models exhibit better performance.

On the other hand, we have also observed that the connected-component based methods [2] are simple but not robust because they are based on geometrical properties of components and texture based methods may be unsuitable for small fonts and poor contrast text [11, 22]. In contrast to the preceding two approaches, edge and gradient based methods [5, 6, 16] are fast and efficient but give more false positives when the complex background is present in the video frame. However the major problem relies in deciding the suitable threshold values to classify between text and non text pixels. A method based on uniform colours in L* a* b* space is also proposed in [21] to locate uniform coloured text in video frames. Obviously, this method fails when text in video contains multiple colours in a text line or in a word. In this context, Shivakumar et al [15] proposed a robust technique based on the gradient difference, but it is observed that the model is parametric in nature and fails to detect text accurately when there is staircase effect in the video frame. Hence in this paper, we propose a new robust differential excitation difference technique for detecting text in video images based on Webers local descriptor model that enhances the text like features and suppresses image data information. It is observed that the high positive and negative differential excitation values exist nearer to text pixel or on text pixels when compared with non text pixel. This observation motivated us to propose an integrated excitation difference model with WLD for text detection in video images.

The rest of the paper is organized as follows. The proposed model is presented in Section II. Experimental results are given in Section III. Conclusion and future works are given in Section IV.

ISSN 2348 – 8034
Impact Factor- 4.022

II. PROPOSED MODEL

The objective of our work presented in this paper is to construct an automatic text detector which is independent with respect to the orientation and the colour of characters and also robust to noise and aliasing artefacts. We propose a new method of text detection in video frames that is based on local information of each pixel. The motivation behind this work is based on the work of Jie Chen et al., [8] where WLD is exploited for the purpose of face detection in a scene image. The proposed model has two stages namely computation of differential excitation information associated with each pixel of a video frame and excitation difference computation for text detection. In the first phase, given a video frame, differential excitation value is computed for each pixel based on Weber's law to highlight the textual portions in an image. During the second phase, row and column-wise differential excitation difference is performed on each pixel to highlight the text data from the non-textual data.

A. Computation of Differential Excitation Information

Ernst Weber observed that the ratio of the increment threshold to the background intensity is a constant and it can be expressed as: **ΔI/I = k** where **ΔI** represents the increment threshold and **I** represent the initial stimulus intensity and **k** signifies that the proportion on the left side of the equation remains constant despite of variations in the **I** term. The fraction **ΔI/I** is known as the Weber fraction.

Given a video frame, the change of the pixel under processing is computed which is nothing but the intensity differences between its neighbours and the pixel under processing. Due to this operation, the salient variation within an image is magnified more specifically in the text region which is similar to simulate human beings perception of patterns. To be specific, a differential excitation of a current pixel is computed as illustrated in Fig. 1, where **Ic** denotes the intensity of the current pixel; $\mathbf{I_m}$ (m=0, 1, …*n*-1) denote the intensities of *n* neighbors of $\mathbf{I_c}$ (n= 8 here).

I_0	I_1	I_2
I_7	I_c	I_3
I_6	I_5	I_4

$$\{\xi(I_c) = \left\lfloor \left[\sum_{m=0}^{n-1} \left(\frac{I_i - I_c}{I_c} \right) \right] \right\rfloor$$

Figure 1 Computation of differential excitation of a pixel I_c

In order to compute the differential excitation of each pixel, we initially calculate the differences between its neighbours and the pixel under processing. Let $\mathbf{I_c}$ be the pixel under processing and let $\mathbf{I_m}$ be one of the neighbour pixel, then the intensity difference, say **ΔI** is $\mathbf{I_m}$-$\mathbf{I_c}$.
Using Weber's notion, we compute the ratios of the differences to the intensity of the current pixel to highlight whether the pixel is a text pixel or not.

$$\text{i.e., } f_{ratio} = \mathbf{\Delta I/I_c} \qquad \dots (1)$$

Subsequently, the neighbour effects on the pixel under processing is obtained using the sum of the difference ratios:

$$\text{i.e., } f_{sum} = \left[\sum_{m=0}^{n-1} \left(\frac{I_i - I_c}{I_c} \right) \right] \qquad \dots (2)$$

In order to improve the robustness of this local descriptor, we use absolute value of a function on f_{sum} to obtain the differential excitation value of the pixel under processing.

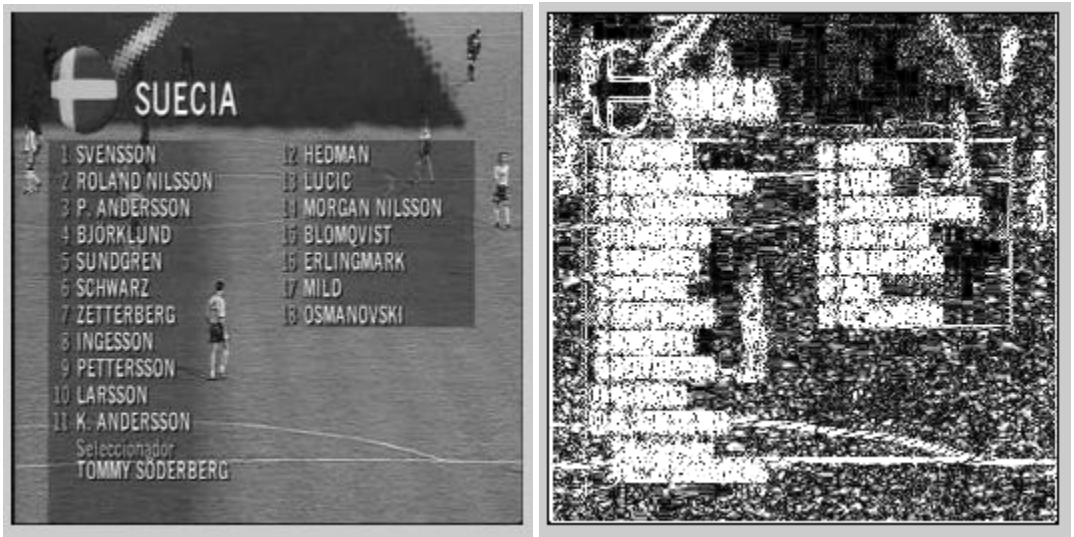
ISSN 2348 – 8034
Impact Factor- 4.022

$$\text{i.e.,}\quad \xi(I_c) = \left\| \left[\sum_{m=0}^{n-1} \left(\frac{I_i - I_c}{I_c} \right) \right] \right\| \qquad \ldots..(3)$$

Upon computing the excitation value of each pixel of a video frame, it is observed that the text pixels will have high excitation value when compared to image pixel. This made us to think of employing the horizontal and vertical masks to extract textual data from a video frame. The result of application of Weber's principle on one such video frame is shown in Fig. 2(b) for a video frame shown in Fig. 2(a). We have shown in Fig. 3 the horizontal and vertical differential excitation profile computed from scan line number 100 of the test image frame shown in Fig. 2(b). Note that the scan line cuts across the "1 SVENSSON" on the left of the image and the words "12 HED MAN" on the right. Large positive spikes on either side are due to text-to-background transitions and background-to-text transitions. Note that the magnitudes of the spikes for the text are significantly stronger than those of the image region.

B. Computation of Differential Excitation Difference

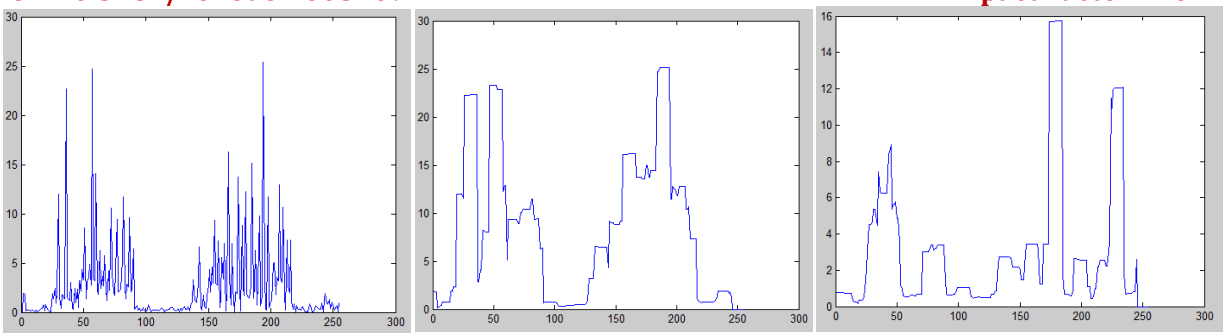
The Differential Excitation Difference (DED) is obtained for each pixel in ξ as the difference between the maximum and minimum excitation difference values within a horizontal mask of size 1 x *n* and vertical mask of size *n* x 1 centered at the pixel where *n* is a value that depends on the character's stroke width. The value of *n* will have an effect on the net result. Keeping in mind, the medium sized font, we have kept the value of *n* as 15. High positive and negative differential excitation values in text regions result from high intensity contrast between the text and background regions. Therefore, text regions will have both large positive and negative differences in a local region due to even distribution of character strokes. This results in locally large DED values. To detect such large values, we have defined an adaptive threshold which varies from frame to frame. Based on the computed threshold, a binarized frame is obtained which highlights the text block and subsequently using projection profile, text regions are extracted. The resultant horizontal differential excitation difference image and vertical differential excitation difference image due to the computation of DED for a video frame given in Fig. 2(a) is shown in Fig. 4(a) & (b) and binarized image is shown in Fig. 4(c) using the adaptive threshold.



(a) *(b)*

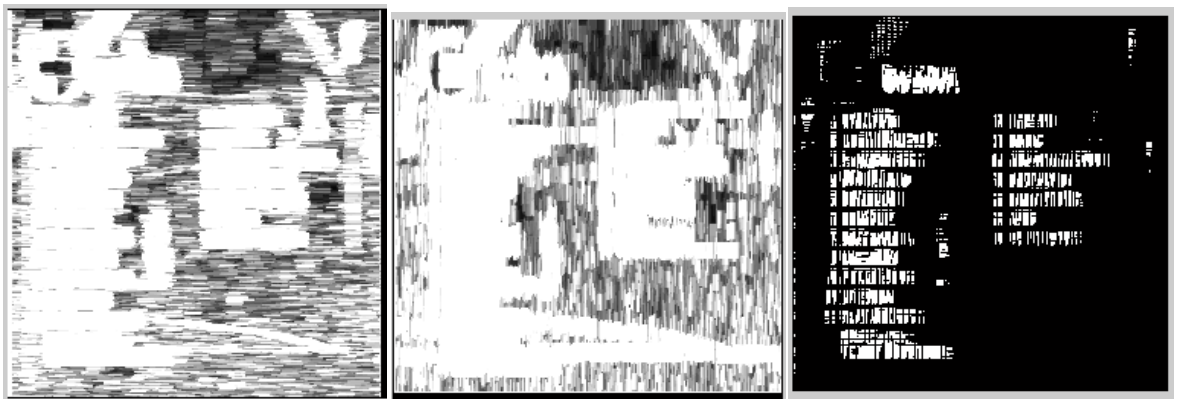
Fig. 2 (a) Original fray scale image; (b) Differential excitation Image

ISSN 2348 – 8034
Impact Factor- 4.022



(a) (b) (c)

Fig. 3(a) Scan line of an image; (b) Horizontal differential excitation difference of the scan line; (c) Vertical differential excitation difference of the scan line



(a) (b) (c)

Fig. 4. (a) Horizontal excitation difference image; (b) Horizontal excitation difference image; (c) Text blocks identified

More formally, the proposed model is described below.

Let I(x, y) be the given gray scale image and let ξ(x, y) be the differential excitation image obtained due to computation of differential excitation of every pixel of I(x, y). Let G_H and G_V be the DED images obtained by performing a convolution operation respectively with a horizontal and a vertical mask of size 1 x *n* and *n* x 1. We compute the minimum and the maximum DED values within a horizontal window, H_w and vertical window V_w over G(x,y) as follows.

$$H_{min}(x,y) = \min_{x_i, y_i \in H_w(x,y)} (\vartheta(x_i, y_i)) \quad \ldots (4)$$

$$H_{max}(x,y) = \max_{x_i, y_i \in H_w(x,y)} (\vartheta(x_i, y_i)) \quad \ldots (5)$$

$$V_{min}(x,y) = \min_{x_i, y_i \in V_w(x,y)} (\vartheta(x_i, y_i)) \quad \ldots (6)$$

$$V_{max}(x,y) = \max_{x_i, y_i \in V_w(x,y)} (\vartheta(x_i, y_i)) \quad \ldots (7)$$

Using Eqs. (4) and (5), we compute the horizontal excitation difference image, say H_{DED} and by using Eqs. (6) and (7), we compute the vertical excitation difference image, say V_{DED}. Then the pixel is classified as text pixel based on the following rule.

$$T(x,y) = \begin{cases} Text\ pixel & if\ G_H(x,y) > \mathcal{T}_H\ and\ G_V(x,y) > \mathcal{T}_V \\ Non\ text\ pixel, & Otherwise \end{cases}$$

ISSN 2348 – 8034
Impact Factor- 4.022

The threshold $\mathcal{T}_H$ is determined based on the average value of horizontal excitation difference computed as follows. First we compute the average horizontal excitation difference values as:

$$H_{AVG} = \frac{1}{pxq}\sum_{i=1}^{p}\sum_{j=1}^{q} G_H(x,y)$$

where p and q are the dimension of the horizontal excitation difference image. Next we count the number of significantly higher horizontal excitation difference values as:

NH_{HOR} = COUNT ($G_H(x, y)$) = H_{AVG}.

The sum of horizontal G_H is computed as

$$SH_{AVG} = \sum_{i=1}^{p}\sum_{j=1}^{q} G_H(x,y)$$

Finally the value of $\mathcal{T}_H$ is computed as follows:

$$\mathcal{T}_H = \frac{SH_{AVG}}{((pxq) - NH_{HOR})}$$

Similarly, the vertical excitation difference image based threshold $\mathcal{T}_V$ is determined using the vertical excitation difference image. The image due to horizontal and vertical excitation difference is shown in Fig. 4(a) & (b). The text detected frame based on threshold values: $\mathcal{T}_H$ *and* $\mathcal{T}_V$ is shown in Fig 4(c). This frame containing only the text highlighted region is used to extract texts from the original video frame for subsequent processing.

III. EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to corroborate the success of the proposed model. We have conducted experimentation on our own video database of images such as movies, news clips, sports and music videos, and scene images, document images (ICDAR-2003) available over internet. The video database contains video segments of varying size, different formats and also of different colours. All experiments are performed on a P-IV 2.99GHz Windows machine with 504 MB of RAM.

The text detection results on some sample images due to the proposed model are given in Fig. 5. We have also provided in Fig. 6, an experimental comparison with Hanif and Prevost algorithm [20] proposed for text detection. The results of the proposed model are able to locate text better when compared to Hanif and Prevost [20] algorithm.
To give an objective comparison of the proposed model, we have made a comparison with Liu et al., [4] and Shivakumara et al., [17] works. Similar to their evaluation criteria, we have used detection rate and false positive rate as decision parameters and metrics. To judge the correctness of the text blocks detected, we manually count the Actual Text Blocks (ATB) in the video frames. Also we manually label each of the detected blocks as either **truly detected text block (TDB)**: a detected block that contains text or a **falsely detected text block (FDB)**: a detected block that does not contain text. Based on the number of blocks in each of the categories mentioned, the following metrics are calculated to evaluate the performance of the approaches:

Detection rate (DR) = Number of TDB / Number of ATB.
False positive rate (FPR) = Number of FDB / Number of (TDB + FDB).

The performance of the proposed approach in comparison with the other existing approaches is summarized in Table 1.

ISSN 2348 – 8034
Impact Factor- 4.022

Fig. 5. (a) Document images (ICDAR-2003) containing text and non-text data

(b) Document images containing only text data highlighted.

ISSN 2348 – 8034
Impact Factor- 4.022

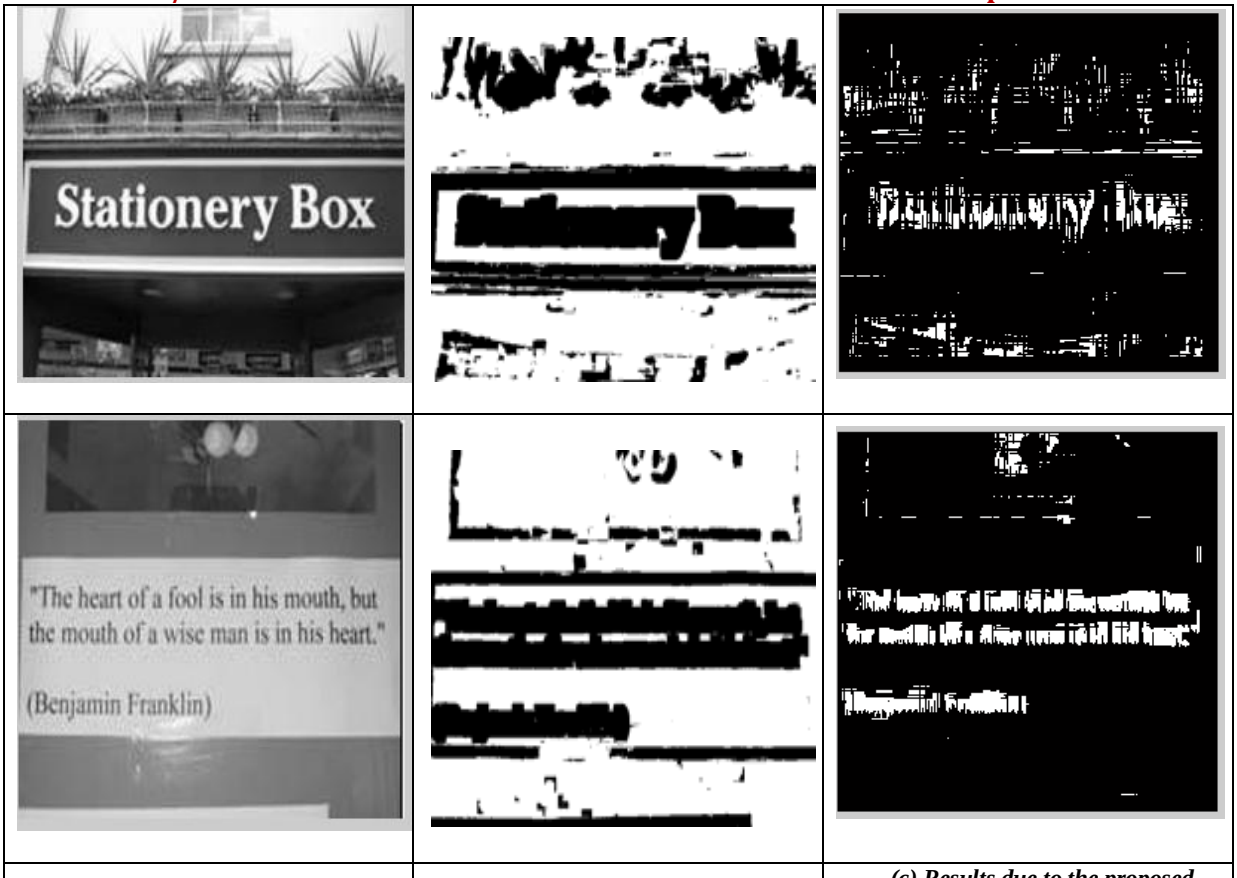


Fig.6. (a) Original document image | *(b) Result due to [20].* | *(c) Results due to the proposed model.*

Table 1. Results of the proposed model and other existing methods

Method	DR	FPR
Liu et al., [4]	79.90	17.90
Shivakumara et al., [17]	89.50	10.60
Proposed Model	92.24	18.33

IV. CONCLUSION

A new model for text detection from video frames is proposed in this work. The concept of Weber's theory is used in the proposed model and excitation image analysis is performed to locate the text region in a video frame. The proposed model works better in the case of more textured images. Although the proposed model is parametric in nature, it is quite simple to implement and the results are on par with the existing models. In our future work, we would like to make the proposed model parameter independent. Experiments on the images taken from ICDAR 2003 robust reading and text locating database and our video database reveals that the proposed model can be utilized for text based video indexing and retrieval applications. An experimental comparison is provided with Hanif and Prevost algorithm [20] and accuracy of the proposed model is compared with Liu et al., [4] and Shivakumara et al., [17] to exhibit the performance of the proposed model.

ISSN 2348 – 8034
Impact Factor- 4.022

REFERENCES

1. *A. Wernicke and R Lienhart. Localizing and segmenting text in images and videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18(8), pp. 256-268, 2002.*
2. *A.K. Jain and B. Yu. "Automatic Text Location in Images and Video Frames". Pattern Recognition, Vol. 31(12), 1998, pp. 2055-2076.*
3. *A.K. Jain, Y Zhong and K. Karu. Locating text in complex color image. Pattern Recognition, pp. 1523-1536, 1995.*
4. *C. Liu, C Wang and R Dai. Text Detection in Images Based on Unsupervised Classification of Edge-based Features. IEEE ICDAR, pp. 610-614, 2005.*
5. *C. Liu, C. Wang and R. Dai. "Text Detection in Images Based on Unsupervised Classification of Edge-based Features". ICDAR 2005, pp. 610-614.*
6. *E. K. Wong and M. Chen. "A new robust algorithm for video text extraction". Pattern Recognition 36, 2003, pp. 1397-1406.*
7. *H Li and D Doermann. A video text detection system based on automated training. In Proceedings of the International Conference on Pattern Recognition, ICPR'00, 2000.*
8. *J. Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikainen, Xilin Chen, Wen Gao. WLD: A Robust Local Image Descriptor, IEEE Transactions on Pattern Analysis and Machine Intelligence 06 Aug. 2009.*
9. *J. Zang and R. Kasturi. "Extraction of Text Objects in Video Documents: Recent Progress". The Eighth IAPR Workshop on Document Analysis Systems (DAS2008), Nara, Japan, September 2008, pp 5-17.*
10. *K. Jung, K.I. Kim and A.K. Jain. "Text information extraction in images and video: a survey". Pattern Recognition, 37, 2004, pp. 977-997.*
11. *K. L Kim, K. Jung and J. H. Kim. "Texture-Based Approach for Text Detection in Images using Support Vector Machines and Continuous Adaptive Mean Shift Algorithm". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, December 2003, pp 1631-1639.*
12. *K. Sobottka and H Bunke. Identification of text on colored book and journal covers. In International Conference on Document Analysis and Recognition, pages 57-62, Bangalore, India, September 1999.*
13. *P. Clark and M Mirmehdi. Combining statistical measures to find image text regions. In Proceedings of the 15th International Conference on Pattern Recognition, pp. 450-453. IEEE Computer Society, September 2000.*
14. *P. Clark and M Mirmehdi. Finding text regions using localized measures. In Proceedings of the 11th British Machine Vision Conference, pages 675-684. BMVA Press, September 2000.*
15. *P. Shivakumara, T Q Phan and C L Tan, A Gradient Difference based Technique for Video Text Detection, International Conference on Document Analysis and Recognition, 2009, pp. 156 – 160.*
16. *P. Shivakumara, W Huang, C L Tan. An Efficient Edge Based Technique for Text Detection in Video Frames, The Eighth IAPR International Workshop on Document Analysis Systems, pp. 307-314, 2008.*
17. *P. Shivakumara, W. Huang and C. L. Tan." An Efficient Edge based Technique for Text Detection in Video Frames". The Eighth IAPR Workshop on Document Analysis Systems (DAS2008), Nara, Japan, September 2008, pp 307-314.*
18. *Q. Ye, Q. Huang, W. Gao and D. Zhao. "Fast and robust text detection in images and video frames". Image and Vision Computing 23, 2005, pp. 565-576.*
19. *R. Lienhart. Automatic text recognition in digital videos. In SPIE, Image and Video Processing IV, pp. 2666-2675, 1996.*
20. *S. M. Hanif and L Prevost. Text detection and localization in complex scene images using constrained AdaBoost algorithm, 10th International Conference on Document Analysis and Recognition, pp. 1-5, 2009.*
21. *V. Y. Marinano and R. Kasturi. "Locating Uniform-Colored Text in Video Frames". 15th ICPR, Volume 4, 2000, pp 539-542.*
22. *Y. Zhong, H. Zhang and A.K. Jain. "Automatic Caption Localization in Compressed Video". IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 4, 2000, pp. 385-392*