An Efficient Algorithm for Text Localization and Extraction in Complex Video Text Images

Anubhav Kumar [1], *Member, IEEE* ,Neeta Awasthi [2]
[1] Assistant Professor, [2] Director
[1 2] Raj Kumar Goel Institute of Technology for women
Ghaziabad, (U.P.), India
Email: [1] rajput.anubhav@gmail.com

***Abstract* — This paper gives an efficient algorithm for text localization and extraction for detection of both graphics and scene text in video images. The Text size is a vital design parameter whose dimension should be properly elected to make the method more robust and insensitive to various font shapes and sizes, styles, color/intensity, orientations, languages, text directions, background and effects of illumination, reflections, shadows, perspective distortion, and the density of image backgrounds. Basically, it works in four steps: Edge generation using Line edge detection mask, Text localization using projection profiles based method, Text segmentation and Text recognition. The result of this paper demonstrates the capability of the proposed technique there by conducting experiments on the video images containing on a large group of complex video text images. The paper proves to be robust for various background complexities and text appearances. The method used in the paper has a high rate as good extraction result. This paper gives a method which has a high rate of good extraction results in complex video images and experimental results. The proposed technique gives better result than existing methods in terms of detection rate for large video image database with very few false alarms ,reliable recall rate and precision rate.**

***Keywords*— Text Extraction; Video Images; Text Localization; Text Segmentation; Text Recognition.**

I. Introduction and related works

Texts in multimedia, natural scenes, video frames provide us with much useful information. However, due to the complex background in video frames, it is difficulties in locating and segmenting text in images. Text segmentation and extraction has been a well-investigated field over the last decade .Basically video text has two types: scene text which exists in the real-world objects like indoor, outdoor, movie video text images and the superimposed text those are added in a video like News video frames. Normally video frames have surrounded complex background. The previous work approach in this field based upon the utilized features can be classified into three groups: connected component-based [3], edge-based [1,4, 5, 6] and texture-based [13] .

Shivakumara et al. [2] proposed a simple and novel method for detecting both graphic text and scene text in video images and introduces candidate text block selection to extract the text portion of the image in accordance with the results and rules based upon filters and edges. Gllavata et al. [3] proposed a method to localize and extract text from color images automatically. The component to be used after the transformation of color image into grayscale is the Y component. Earlier A Kumar et al. [4] proposed an edge based text extraction in images and video frames. This algorithm depends upon the line edge operator and the projection profile method is used to localize the text region better. A focus of attention based system for text region localization has been proposed by Liu at el [6]. For detecting text regions in an images the intensity profiles and spatial variance are used. Zhong Ji et al.[7] proposed a novel overlay and scene text detection method based on SVM and hybrid features, that is a wavelet transform coefficients features, gray-level co-occurrence matrix and oriented edge intensity ratio features, are employed to distinguish text from the background.

A fast and effective text detection approach is proposed by Xiaojun Li et al. [8] . The devised features based on the stroke maps in four directions are able to better represent the intrinsic characteristic of the text. Wonjun Kim et al.[9] proposed an overlay text detection and extraction from complex videos and compute the density of transition pixels and the consistency of texture around the transition pixels to distinguish the overlay text regions from other candidate regions. Shi Jianyong et al.[10] proposed an automatically detect and localize text regions in video frames, and transform the text regions into a binary image with white characters in a pure black background, which is a direct input to the OCR engine for recognition. Wu et al. [11] proposed a non-region-based approach, which uses edges in the text regions to form strokes, and then strokes are aggregated to form chips.

Gao and Tang [12] proposed to use horizontal and vertical projections of edges to localize text strings; however, it can only handle captions and cannot deal with complex text layouts. Some methods combine these two paradigms. Q. Ye et al.[13] Proposed , a novel coarse-to-fine algorithm that is able to locate text lines even under complex background using multiscale wavelet features and an SVM classifier is used to identify true text from the candidates based on the selected features. K. Kim et al. [14] present a novel texture-based method for detecting texts in images. A support vector machine (SVM) is used to analyze the textural properties of

978-81-920249-7-4/13/$31.00 ©2013 IEEE

texts. No external texture feature extraction module is used. Q. Liu at el. [15] propose stroke filter for text localization in video images. The novelty of stroke filter is that it discovers the intrinsic characteristics of text by the analyzing the relationship between the local regions.

Our proposed method for image text extraction system (shown in Fig.3) extracts a text region from an image. The rest of the paper is organized as follows. Section II is a literature of proposed algorithm and it describes the result and analysis in Section III. The conclusion process in Section IV.

II. PROPOSED ALGORITHEM

The specific flow chart of the method proposed method is shown in figure 3. Basically it works in three steps: firstly edge generation using Line edge detection mask applied. After this, text localization using projection profiles based has been done. At last text segmentation and text recognition has been applied one the Loaclized images. These steps are elaborated asunder:

A. Edge map Generation –

Intensity image binarized image usually both have edge maps containing real edge pixel of rich contrasts region. Intensity image contains edge maps having scattered edge pixels as noise in an edge map and binarized image contains an edge map has another type of noise caused by binarization process. Enriched edge pixels of text are obtained by merging the two edge maps and get an edge map which contains minimum edge pixels of background. The proposed method allows better detection of intensity peaks that generally leads to a characterization of in text also it uses the magnitude of the second derivative of intensity as a measurement of edge strength. The distribution is relatively concentrated so that the richer the edge of the text near as well as smaller the distance of the text edges in the same region. In this step firstly line detection mask (1) is created to detect edges at Gx and Gy orientations converting video images to intensity based edge image using line edge operator shown in figure.2 .Further more directional edge maps that are used to represent the vertical and horizontal directions edge density and edge strength up to now absolute value and complex magnitude has been determined and average map will be drawn next.

Obtain the minimum and the maximum gradient values in w over G(x,y) as follows

$$Min\ (x,y) = \min_{x_t y_t \in W(x,y)} (G(x_t, y_t)) \qquad (1)$$

$$Max\ (\ x, y\ ) = \max_{x_t y_t \in W(x,y)} (G(x_t, y_t)) \qquad (\ 2)$$

A global threshold (T) is calculated on the average value of the gradient difference as follows. First we compute the average gradient values. Using equation (1) and (2), compute GD(x,y) a follows

$$GD(x,y) = Max\ (x,y) - Min\ (x,y) \qquad (3)$$

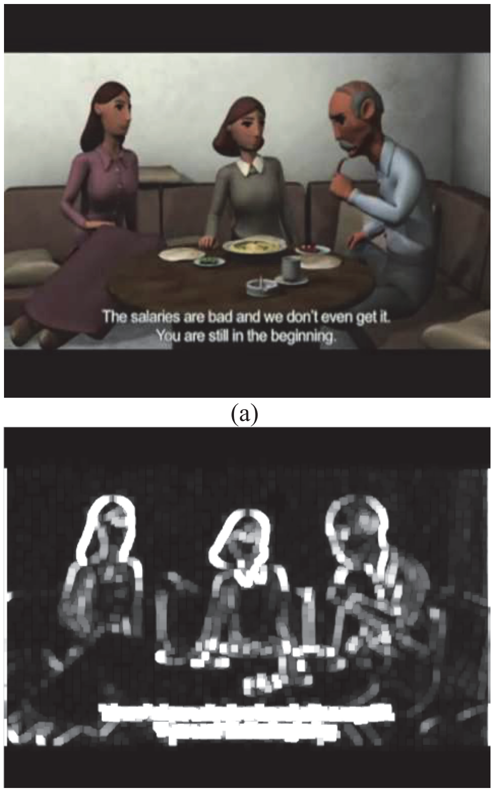
Then a pixel is classified as follows

$$(x,y) = \begin{Bmatrix} Text\ Pixel\ , & if(GD(x,y)>T) \\ Non\ Text\ pixel\ , & Otherwise \end{Bmatrix} \qquad (4)$$

$$\begin{bmatrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{bmatrix} \qquad \begin{bmatrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{bmatrix}$$

Gx Gy

Figure 1. Horizontal (Gx) and Vertical (Gy) edge line detector operator


The salaries are bad and we don't even get it.
You are still in the beginning.

(a)

(b)

Figure 2. (a) Original image (b) Average edge line detector image

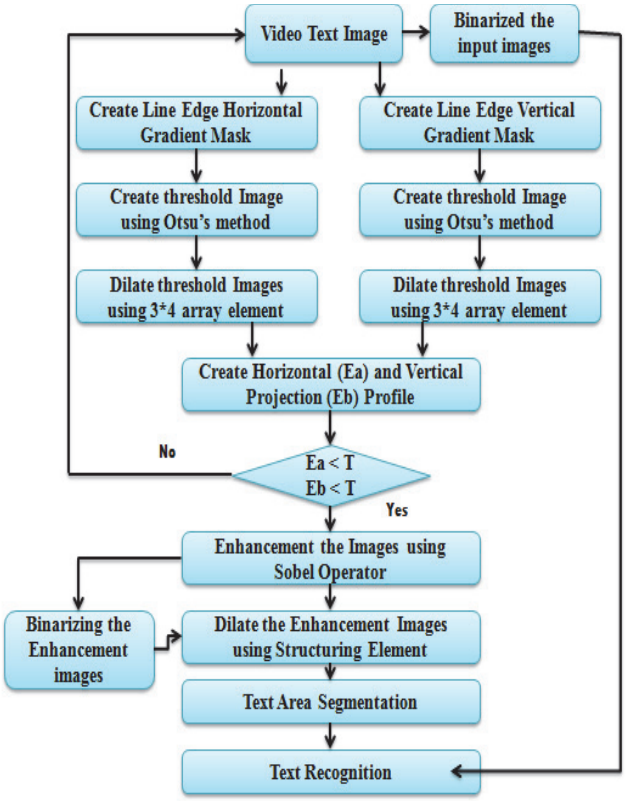

Video Text Image
Binarized the input images
Create Line Edge Horizontal Gradient Mask
Create Line Edge Vertical Gradient Mask
Create threshold Image using Otsu's method
Create threshold Image using Otsu's method
Dilate threshold Images using 3*4 array element
Dilate threshold Images using 3*4 array element
Create Horizontal (Ea) and Vertical Projection (Eb) Profile
No
Ea < T
Eb < T
Yes
Enhancement the Images using Sobel Operator
Binarizing the Enhancement images
Dilate the Enhancement Images using Structuring Element
Text Area Segmentation
Text Recognition

Figure 3. Flow Chart of the proposed algorithm

B. Text Localization –

Mean vertical edge is given by image with some short edge remained unconnected in the previous process. Erosion is used to eliminate these edges.

$$(A \Theta B)(x) = \{x \epsilon X, x + b \;\epsilon\; A : b \;\epsilon\; B\} \qquad (5)$$

Where set of A is image , *B* is the disk structuring element and -b is the scalar multiple of the vector b by -1. This procedure helps to decrease the number of candidate regions for the next step and build the search process faster. Even after the elimination process these are a number of connected text and non connected text regions and the same process is repeated to search for text region from this image. So, narrow based on the geometric features of text such as height and width has been used in this process. Certainly, text in a label is organized in a strip of character in a straight line, tilted or curved nature with length higher than, frequently several times of its height. Depending of prescribed features, properties of the image in the entire region have been detected. If the measured height and width is fond greater than the specified measurement; then it is eliminated from the text. Text strings are indicated by horizontal rectangular axes with high density and projection find such high density area. Many edges dense blocks are fond containing multiple texts.

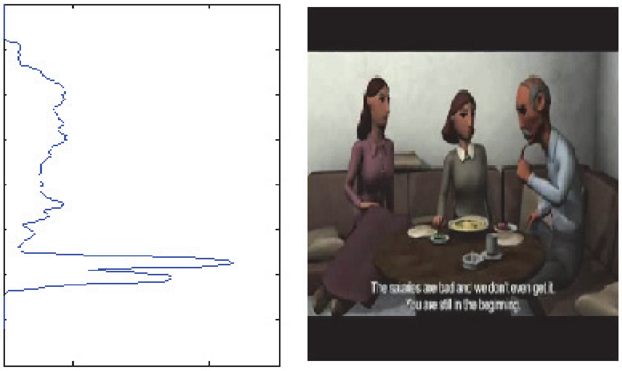


Figure 4. Horizontal projection profiles along with Video image.

Projection profile separates the blocks into single text line .The vector sum of the pixel intension over each column and row given the horizontal and vertical projection profile. The horizontal (Ea) and vertical projection (Eb) of the processed edge map is found. Threshold (T) is termed as the average of the minimum and the maximum value of the vertical projection. Thereafter the row having sum of pixel intensities above the threshold are taken. According to the filter out edges incompatible with the text size: the edge maps of image (Aa) and its binarized image (Ab) are generated, and then the two edge maps are merged into one using AND operation emerging method (7).

$$A = Aa \; \textbf{\textit{AND}} \; Ab \qquad (7)$$

Where AND denotes a AND operation for each pair of pixels in both binary images . In the second step, for evaluating the distribution of edge pixels in both horizontal and vertical direction projection profile method is used[10], thereafter thresholding methods applied to identify adjacent pixel row and column containing text pixels and these rows and columns compose the text region.

```
If   (Ea) < (T), Text regions in horizontal direction…
    Else if     (Eb) < (T), Text regions in Vertical direction…
        Else Non text regions
```

In the third step, to check the text region contains real text or not. If the region contains real text, then the region is sent to the next extraction step, else if there is a false alarm, the current process would be terminated. In this step only gives the horizontal projection of only those rows whose value is less than the value of average threshold. Some regions prove out to be too large or too small to be instances of text using projection profile. Therefore edges exceeding the maximum height and width are removed. Thus localization of the text regions of the image is achieved. The corresponding regions in the original image are taken and the text localized image is shown in figure 5.

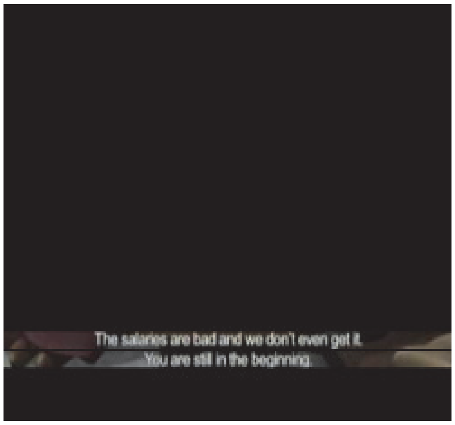

The salaries are bad and we don't even get it.
You are still in the beginning.

Figure 5. Text Area Localized image.

C. Text Segmentation –

In this Step, Localized video image is converted to intensity based edge map using a Sobel edge operator. It is followed by a morphological process applied to edge based intensity image. Strong and weak sub bonds are diluted in these steps .The regions of foreground pixel are enlarged by dilation operates done by adding pixels to the boundaries of the object in a binary image (8). The holes within the region become smaller while foreground pixels grow in size. So in this step segmented text area image are found shown in figure 6.

$$(A \oplus B)(x) = \{x\epsilon X, x = a + b: a \,\epsilon\, A : b \,\epsilon\, B\} \qquad (8)$$

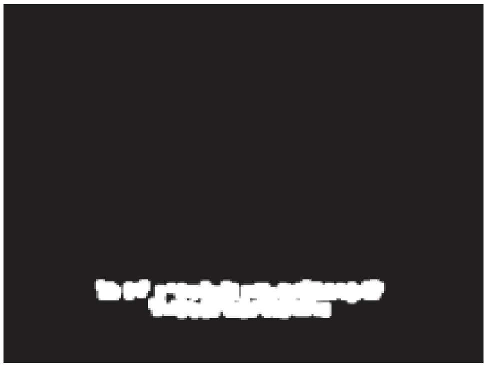


Figure 6. Text Area Segmentation image.

D. Text Recognition –

The segmented images are used for the extraction process in this step. The extracted text to be parsed and recognized by the Common OCR systems. The segmented text image is multiply the segmented text image with input binary image and text extraction result has been obtained in the form of white with pure black background pixels. The text extraction results are shown in figure 7.

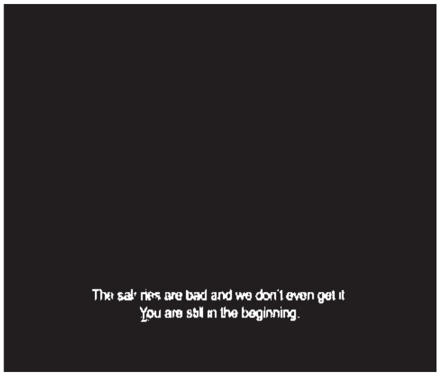
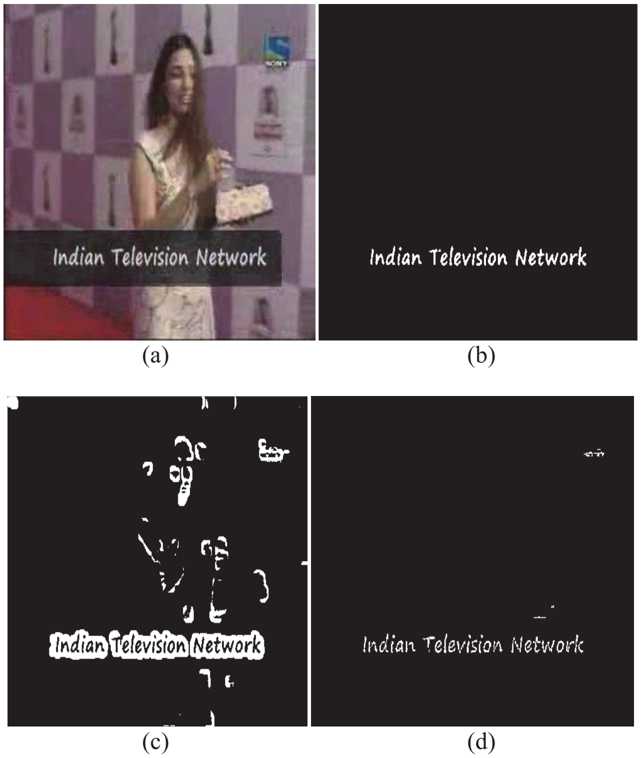

The sal ries are bad and we don't even get it
You are stil in the beginning.

Figure 7. Text extraction result.

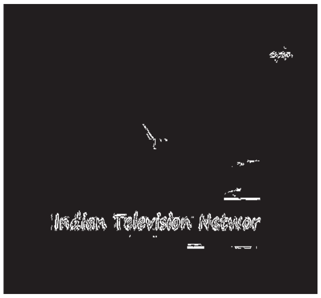
III. RESULT AND ANALYSIS

The method introduced in this paper was tested upon of color video text images. The experiments were performed by using a PC (Laptop: Intel Core (i5) 4 2.5GHz) with MATLAB. This method was evaluated with that of typical Xiaoping Liu et al. [6] and Gllavata *et al*. [3] with my previous algorithm [4] independently in term of effectiveness, quality and robustness respectively by measuring the computing time and by manually estimating the rates of recall rate and the precision rate. Here, 72 experimental video Text frame images are taken from the internet randomly which are having 3501 total text words. Performance results for the eight videos are given in Table I.


Indian Television Network
Indian Television Network
Indian Television Network
Indian Television Network

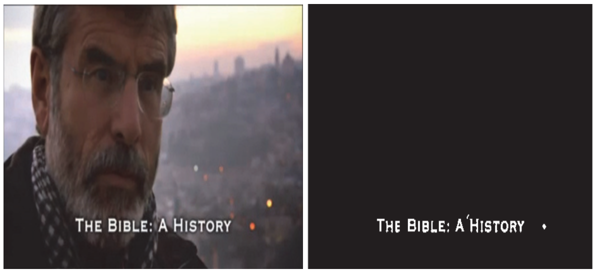
(a) (b)

(c) (d)



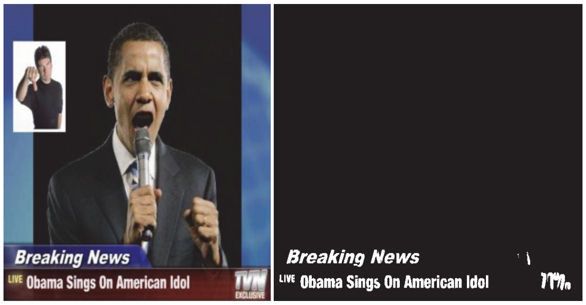
(e)

Figure 8. (a) Original book cover Image (b) Output from proposed method (c) Output from Liu et al. [6] method (d) Output from Kumar et al. [4], (e) Output from Gllavata et al. [3] method



(a) (b)

Figure 9. (a) Original video Images , (b) Output from proposed method


Breaking News
LIVE Obama Sings On American Idol
TVN EXCLUSIVE
Breaking News
LIVE Obama Sings On American Idol

(a) (b)

Figure 10. (a) Original News images Images , (b) Output from proposed method

TABLE I. RESULT OF THE DETECTED WORDS IN PROPOSED AND OTHER EXISTING METHOD

Image type		**Detected words**			
	Total words	**Proposed Method**	**Liu method [5]**	**Kumar et el.[3]**	**Gllavata *et al.*[2]**
News Images	**1506**	**1415**	**1115**	**1345**	**1313**
TV Images	**1133**	**1122**	**815**	**1024**	**1001**
Video images	**862**	**851**	**521**	**813**	**648**

TABLE II. RESULT OF THE FALSE POSITIVE IN PROPOSED AND OTHER EXISTING METHOD

Image type	**False Positive**			
	Proposed Method	**Liu method [6]**	**Kumar et el.[4]**	**Gllavata *et al.*[3]**
News Images	**56**	**264**	**46**	**72**
TV Images	**83**	**443**	**42**	**35**
Video images	**21**	**470**	**73**	**33**

TABLE III. PERFORMANCE OF THE RECALL RATE IN PROPOSED AND OTHER EXISTING METHOD

Image type	**Recall Rate %**			
	Proposed Method	**Liu method [6]**	**Kumar et el.[4]**	**Gllavata *et al.*[3]**
News Images	**93.95**	**74.03**	**89.30**	**87.18**
TV Images	**99.02**	**71.56**	**90.37**	**88.34**
Video images	**98.72**	**60.44**	**94.13**	**75.17**

TABLE IV. PERFORMANCE OF THE PRECISION RATE IN PROPOSED AND OTHER EXISTING METHOD

Image type	**Precision Rate %**			
	Proposed Method	**Liu method [6]**	**Kumar et el.[4]**	**Gllavata *et al.*[3]**
News Images	**96.13**	**80.85**	**96.69**	**94.80**
TV Images	**93.11**	**64.78**	**96.06**	**96.62**
Video images	**97.59**	**52.57**	**91.76**	**95.15**

From Fig. 8 ~ 10, it can see that the performance of our proposed method and other existing method on a wide variety of video frames. The performance of every method has been evaluated based on its precision and recall rates obtained overall.

$$\text{Recall Rate} = \frac{\text{Correctly Detected Words}}{\text{Correctly detected Words+False Negative}} * 100 \quad (9)$$

$$\text{Precision Rate} = \frac{\text{Correctly Detected Words}}{\text{Correctly detected Words+False Positive}} * 100$$

(10)

According to table I~VI results obtained show that the proposed experiment result obtained a Racall rate of 93.95%, 99.02% and 98.72% and precision rate of 96.13%, 93.11% and 97.59% for News, commercial and video text images respectively. It can be noted from table II that Liu et al. [5] obtained an Racall rate of 74.03%, 71.56% and 60.44% and precision rate of 80.85%, 64.78% and 52.57%, Kumar et el. Obtained an Racall rate of 89.30%, 90.37% and 94.13% and precision rate of 96.69%, 96.06% and 91.76% and Gllavata *et al*. Obtained an Racall rate of 87.18%, 88.34% and 75.17% and precision rate of 94.80%, 96.62% and 95.15% for News, commercial and video text images respectively. The proposed algorithm achieved a precision rate 95.49% and a recall 96.77% for the 72 video frames test images with average computed time 0.360 second /frames higher than other existing method.

TABLE V. AVERAGE PERFORMANCE OF THE PROPOSED AND EXISTING METHOD

Methods	Total Images	Recall Rate %	Precision rate %	Average time (s)
Proposed Method	**72**	**96.77**	**95.49**	**0.360**
Liu Method		**70.00**	**67.77**	**7.58**
Kumar et al.		**90.88**	**95.18**	**2.3**
Gllavata *et al.*		**84.60**	**95.48**	**19**

An implementation of a typical Xiaoping Liu et al. [6] and Gllavata *et al*. [3] with my previous algorithm Kumar et al. [4] for Extraction the text in video images is done to verify the effectiveness of the proposed method. The compared experimental and performance results are shown in Table VI and V. Table I, II, III, IV and can conclude that the performance and processing time of the proposed method is far better than the implemented method in recall rate, precision rate and average time parameters.

TABLE VI. PERFORMANCE OF THE PROPOSED AND OTHER EXISTING METHOD

Methods	Recall rate %	Precision Rate %	Average Time (Image/s)
Proposed	**96.77**	**95.49**	**0.360**
Gllavata et al. [3]	**84.60**	**95.48**	**19**
Kumar et al.[4]	**90.88**	**95.18**	**2.3**
Liu et al. [6]	**70.27**	**65.85**	**7.58**
Li *et al.* [8]	**91.1%**	**-**	**12.9**
Ye *et al.* [13]	**90.8%**	**-**	**10.1**
Liu *et al.* [15]	**91.3%**	**-**	**11.7**
Lyu et al. [16]	**82.13**	**90.22**	**-**
Ye *et al.* [17]	**94.2**	**-**	**8.3**
Lienhart et al. [18]	**94.3**	**-**	**2.2**

IV. CONCLUSION

This paper gives an algorithm which is efficient for text extraction is complex video text images. The experimental results show that the chosen features are efficient with respect to the comparisons between text regions and non-text regions. Experimental results obtained a 96.77 % recall rate and precision rate 95.49 by the average text extraction time is 0.360 second, which is far better than other previous and also presents robustness in text localization from low quality images. Still some text in images remains being extracted by proposing methods, which are more wide spaced text specially written in large font size and large gap between texts .In future, the algorithm will be solved and satisfied spaced font and color effects in text detection.

REFERENCES

[1] Kumar, Anubhav. "An efficient text extraction algorithm in complex images." *Contemporary Computing (IC3), 2013 Sixth International Conference on*. IEEE, 2013.

[2] Shivakumara, P. ; Weihua Huang ; Chew Lim Tan ;"An Efficient Edge based Technique for Text Detection in Video Frames"., The Eighth IAPR International Workshop on Document Analysis Systems, DAS '08,pp. 307 – 314,2008,IEEE.

[3] J. 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,IEEE.

[4] Kumar, Anubhav, Awanish Kr Kaushik, and R. L. Yadav. "A robust and fast text extraction in images and video frames." *Advances in Computing, Communication and Control*. Springer Berlin Heidelberg, 2011. 342-348.

[5] Kumar, Anubhav, Awanish Kr Kaushik, R. L. Yadava, and Divya Saxena. "An Edge-Based Algorithm for Text Extraction in Images and Video Frame." *Advanced Materials Research* 403 (2012): 900-907.

[6] Xiaoqing Liu, Jagath Samarabandu: Multiscale Edge-Based Text Extraction from Complex Images. International Conference on Multimedia and Expo,icme, pp. 1721-1724,IEEE, (2006).

[7] Zhong Ji ; Jian Wang ; Yu-Ting Su ; "Text detection in video frames using hybrid features" ., International Conference on Machine Learning and Cybernetics,pp. 318 – 322, 2009,IEEE.

[8] Xiaojun Li; Weiqiang Wang; Shuqiang Jiang; Qingming Huang; Wen Gao; "Fast and effective text detection"., 15th IEEE International Conference on Image Processing,pp. 969 – 972, ICIP 2008,IEEE..

[9] 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,IEEE.

[10] Shi Jianyong ; Luo Xiling ; Zhang Jun ; "An Edge-based Approach for Video Text Extraction"., International Conference on Computer Technology and Development, ICCTD '09,pp. 331 – 335,2009,IEEE.

[11] V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: An automatic system to detect and recognize text in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1224–1229, Nov. 1999

[12] X. Gao and X. Tang *et al.*, "Automatic news video caption extraction and recognition," in *Proc. LNCS 1983: 2nd Int. Conf. Intell. Data Eng. Automated Learning Data Mining, Financial Eng., Intell. Agents*, K. S. Leung *et al.*, Eds., Hong Kong, 2000, pp. 425–430.

[13] Q. Ye, Q. Huang, W. Gao and D. Zhan, Fast and robust text detection in images and video frames, Image and Vision Computing, 2005,23,pp:565-576.

[14] K. Kim, K. Jung, and J. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1631–1639, Dec. 2003.

[15] Q. Liu, C. Jung, S. Kim, Y. Moon and J.Kim," Stroke filter for text localization in video images," in *Proc. Int. Conf. Image Process.,* Atalanta, GA, USA, Oct. 2006, pp. 1473-1476.

[16] Michael R. Lyu, Jiqiang Song, Min Cai. "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction" *IEEE Transaction on circuits and systems for video technology*, IEEE,vol. 15, No. 2, Feburary 2005, pp 243-255.

[17] Qixiang Ye, Qingming Huang, Wen Gao and Debin Zhao,Fast and Robust text detection in images and video frames,Image and Vision Computing,23,2005.

[18] R. Lienhart, A. Wernicke, Localizing and segmenting text in images and videos, IEEE Transactions on Circuits and Systems for Video Technology 12 (2002) 256–268.

[19] M. Zhao, S. Li, J. Kwok, Text detection in images using sparse representation with discriminative dictionaries, Image and Vision Computing, Elsevier 28 (12) (2010) 1590–1599.

[20] Anubhav Kumar, "An Robust and Fast Algorithm for Text Extraction in Video Text Images", Paper published in IEEE 2nd International Conference on Current Trends in Technology, NUiCONE 2011, PP-1-5.