IJCTA, 9(20), 2016, pp. 411-419

Text Detection from Live Sports Video

Lalita Kumari*, J. L. Raheja and Vidyut Dey*****

ABSTRACT

Text detection from a video stream is less explored area but very useful for many researches and applications in the field of Artificial intelligence, robotics, and computer vision. In this paper, we proposed a novel approach to detect text, based on Maximally Stable Extremal Regions (MSER) feature and geometrical properties analysis. The method uses MSER feature extraction followed by filtering of text area using basic geometric properties along with stroke width analysis. MSER minimizes regularized variations and is best suited to extract character candidates from a natural scene image by a well-defined closed boundary measure. The proposed methodology is able to detect both overlaid text and scene text from video stream of sports events such as cricket. This proposed approach is capable of handling complex text information efficiently, and showed great success in text detection from live video stream such as cricket video. Experimental results show feasibility of the proposed technique and confirm the practical efficiency of the system.

Keywords: Text detection, Video stream, Sports event video, Video OCR.

1. INTRODUCTION

Text in a video stream is categorized in to two broad types: - scene texts and overlaid texts. The scene texts are the natural text present in video frames. These types of text are part of the scene which appear naturally and captured by camera. The overlaid texts are added purposefully during video production. Text detection in complex nature scenes is one of the hardest problems in computer vision. Robustly locating the text is the first step and very essential task towards many content-based image analysis tasks, but detecting text area in natural scene becomes difficult task because of its complex background, variations in font and text size, orientation of text, text color, and lighting conditions. Text detection from Live sports Event video is a challenging task. Figure 1 shows a sample video frame from a cricket play which gives versatile nature of area of interest.

Text appearances in video frames are very random as well as very much fast. Text position and text size is changing so rapidly that one cannot predict it. Even appeared text moves randomly in every direction. Background of the video becomes too complex due to large variation in illumination, color, object positions, etc. Because of these issues, this problem not only becomes challenging, but also become important for several research and applications.

Appearing text in sports event video are part of the video and are useful for many applications such as video analysis, video annotations, video content searching/retrieval, survey, etc. The first step of any text-based application for image or video sequence is localization. Text location determination and text size calculation is first step for any application of such domain. The generalized video OCR system consists of three major steps: Text detection, recognition, and tracking of text. The first step i.e. text detection is meant for determining the text location and text size in video frame. Second step i.e. text tracking register the detected text object and track movement of the identified text object in video

* Department of Electronics & Communication Engineering, NIT Agartala, Tripura, India, *Email: kumaril2003@yahoo.co.in*

** Digital System group CSIR/ CEERI, Pilani, Rajasthan, India, *Email: jagdish.raheja.ceeri@gmail.com*

*** Department of Production Engineering, NIT Agartala, Tripura, India, *Email: vidyut.pe@nita.ac.in*

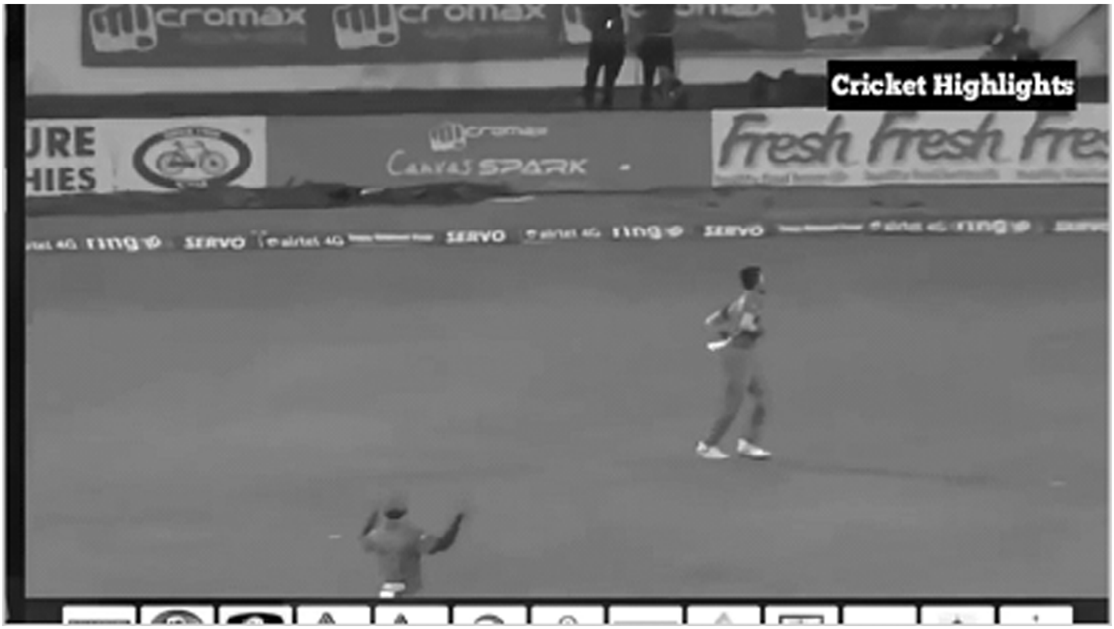


Figure 1: Representation of randomness and challenges of text detection in a sample video frame

sequences. The third step i.e. text recognition perform actual OCR to recognize the text which includes many sub operations such as background subtraction from complex and dynamic scene, contrast adjustment, illumination normalization, etc. Generally embedded text of video frame occurs in very heterogeneous background as well as with large variation of contrast, which makes it very difficult to be recognized by standard OCR software.

Video text detection framework utilizes connection characteristics [15] [22]; texture alike characteristic [12] [16] [20], and edge density information [13] [14]. Connection characteristics based method work on assumption that text regions have uniform colour, shape, size, and spatial layout. The texture alike characteristic of the text region is based on texture in image and work on assumption that text regions have uniform spatial texture. Edge density information based method work on assumption that density of edge density of text regions as completely distinguishable from the background spatially the corner of the text regions.

Discrete Cosine Transform (DCT) coefficient of gray scale image has been applied for text detection in [12]. Lu and Barner proposed in [10], a weighted DCT coefficient to detect text based on the texture information represented by weighted DCT coefficients which further improves text detection performances. Discrete Wavelet Transform (DWT) is also used in text detection to transform images into different sub-bands. In [13] [18] multi resolution based method for text detection has been discussed. It is discussed in [13] [18] that Integration of detected text in different sub-bands, produces performance. Combination of properties such as edge, texture, and shape information are generally used to filter out the false text regions from detected text regions [13] [14] [19].

Although many methods have been proposed for text detection in such problems [13] [17] [18] [19] [21] [22], few of them address structure feature, and few addressed the correlation of video text. But background led to high false detection rate. Based on recent related research, our research includes four steps: (i) Candidate text extraction from video stream using MSER features (ii) refining of candidate text by geometric property analysis of candidate text (iii) Text area detection from recorded sports video (iv)Text detection from locked text area. In this paper we proposed a framework to determine natural scene text along with overlaid text from the sports event video such as cricket video stream data. Our proposed framework is represented by the flowchart shown in figure 2.

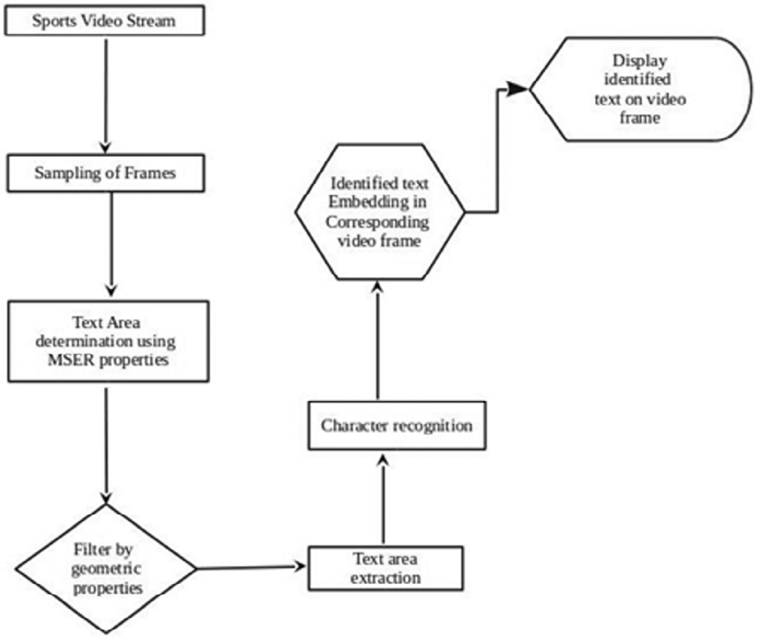

Sports Video Stream
Sampling of Frames
Text Area determination using MSER properties
Filter by geometric properties
Text area extraction
Character recognition
Identified text Embedding in Corresponding video frame
Display identified text on video frame

Figure 2: Generalized flow chart for proposed text detection from sports video

Rest of the paper is organized as follows. Text area localization step is explained in section 2. Geometric properties are analyzed in step 3 to filter candidate text area. Text area detection from sports video frames is explained in section 4. Final text detection step is explained in section 5. Testing and result analysis is discussed in section 6. Quality of the text detected from the live video is measured in positive predictive value (precision) and Negative predictive value (recall).

2. DETERMINATION OF CANDIDATE TEXT

Text area localization from complex background is a challenging task as it has multiple colors and object of different shapes. To determine text in natural scene image, first step is to partition the scene image in such a way that each portioned block contains character candidate. Character candidate is determined by various features such as uniform local gradient, uniform color of character components, uniform stroke width of character components, etc. This step is done by structural analysis in search of independent text character or group of adjacent character in horizontal or vertical direction in complex natural scene image. Maximally Stable Extremal Region (MSER) feature can be efficiently used for object recognition in natural scene images. These different identified objects are verified latter by geometrical properties of text.

MSER regions are connected areas which are characterized and identified by almost uniform intensity, and is surrounded by contrasting background. These areas are constructed by using a specific process of trying multiple thresholds. MSER regions are finally identified as those regions those that maintain unchanged shapes over a large set of thresholds. MSERs are defined by intensity in the region along with outer border, and hence it is very useful to determine object regions. For an object in scene image, local binarization should be stable over a large range of threshold, and is selected as an object. With help of MSER features, both large and small object is identified as it uses multi-scale detection. Above all, MSER feature can identify object in very effectively (in worst case O(n)) [8]. MSER is like Watershed that is focused on finding stable connected-components over the largest possible size. This achieved by applying a threshold value from 0 to maximum, one step at a time. It is similar to flooding the basins in Watershed-speak. Smaller regions are merged to form larger regions if area remains ‘stable’. The region becomes a candidate when the area growth rate reaches a local stationary point.

3. GEOMETRIC PROPERTIES ANALYSIS OF CANDIDATE TEXT AREA

A fast and efficient text recognizer needs to narrow down the obtained candidate text locations by filtering out the localized text area in the image. Much analysis on dataset has been performed in past research to obtain distinctive features which are capable of distinguishing the text area and non-text area from available object area. Many text features are computed, analyzed, and recommended to consider for distinguishing between character object and non character objects. Some of those features are as follow:

Solidity: Solidity specify the proportion of the pixels that are present in the convex hull and are also in the region. i.e. solidity is what fraction of the actual area the region is. Convex hull, represented by equation-(I) in figure 3, is determined from the shape of the candidate object for text area filtering. Area of this transformed convex shape is used to determine ratio of candidate text area (represented by equation-(II) in figure 3. Solidity ranges between 0 and 1. Analysis of large dataset point to higher solidity for expected text object.

Eccentricity: It specifies the eccentricity of the ellipse that has the same second-moments as the region. Eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length, represented by equation-(III) in figure 3. Its value can be any fractional number between 0 and 1. (0 and 1 are special cases; if an ellipse has eccentricity 0 then it becomes circle, and if ellipse has eccentricity 1 then it becomes a line segment). Figure 4 shows the eccentricity of text object. Analysis of large dataset point to abandon out the higher value of eccentricity for expected text object.

Extent: Extent describes the ratio of pixels in the region to pixels in the total bounding box. It is computed by dividing area of object by the area of the bounding box. Extent is represented by equation-

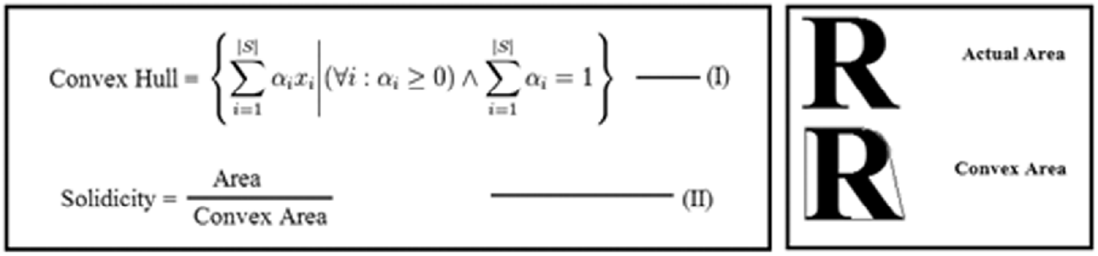

$\text{Convex Hull} = \left\{ \sum_{i=1}^{|S|} \alpha_i x_i \middle| (\forall i : \alpha_i \geq 0) \wedge \sum_{i=1}^{|S|} \alpha_i = 1 \right\}$ —— (I)
$\text{Solidicity} = \frac{\text{Area}}{\text{Convex Area}}$ —— (II)
R
Actual Area
R
Convex Area

Figure 3: convex hall and solidicity representation

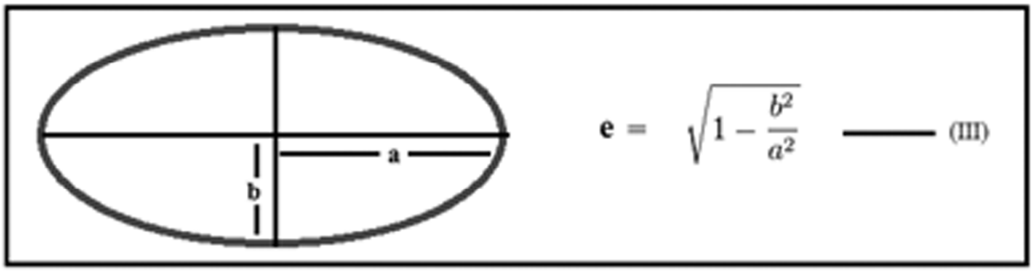


Figure 4: representation for eccentricity

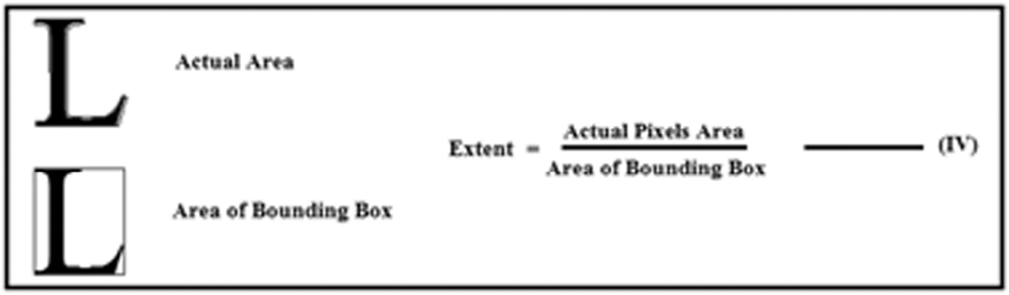

L
Actual Area
L
Area of Bounding Box
$\text{Extent} = \frac{\text{Actual Pixels Area}}{\text{Area of Bounding Box}}$ —— (IV)

Figure 5: representation of extent

(IV) in figure 5. Statistical analysis of large dataset indicates that extent of a candidate text object play important role to filter out text object from non-text object.

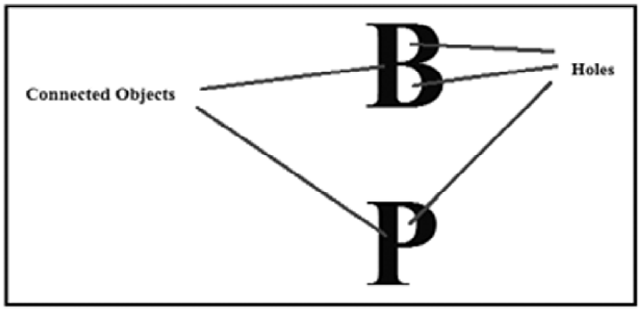
Euler Number: Euler number is computed in a binary image which represents difference between total number of objects present in the area, and total number of holes (represented in figure 6). Euler number is important feature of a binary image, which describe the topological structure of image. Euler number of binary image is represented by equation (vi) in figure 7. Here N is number of connected objects and H is number of holes or disconnected background.

Euler Number is computed locally in two ways, 4-connected, and 8-connected. 4-connected Euler no and 6-connected is represented by equation (vii) and equation (viii) respectively for a binary image. Here V is total number of foreground pixels (1 in binary image) and E is total number of two consecutive ones scanned either horizontally or vertically. D is total number of two consecutive ones at diagonal position, represented by equation (viii) of figure 6. F is total number of 2X2 pixels size box containing 1 only, represented by equation (ix) of figure 7. T is total number of 2X2 pixel box containing 3 ones and 1 zero, represented by equation (x) of figure 7. Euler Number is proved as important feature in image analysis. It is used to suspect an object for being a text object.

Aspect ratio: A candidate region is also filtered out by use of aspect ratio of the object. Aspect ration is used to identify object of similar shape. Aspect ration of candidate text vary slightly on one script to another. To filter out the wrong candidate object of very different height width ratio, such as tree, pillar, etc. in natural scene, aspect ratio play vital role. It is calculated from ration of height and width of bounding box, containing candidate object.

4. TEXT AREA DETECTION IN SPORTS EVENT VIDEO

Text area localization in each frame of sports event video such as live cricket, football event is very challenging task. It requires determining text area in fast moving video and that also to be finished in real



Euler Number = N - H ———— (VI)

4-connected Eular no = V - E + F ———— (VII)

8-connected Eular no = V - E - D + T - F ———— (VIII)

Figure 6: Euler number representation

$$D = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1}\left(a_{i,j} \times a_{i+1,j+1}\right) + \sum_{i=1}^{r-1}\sum_{j=1}^{c-1}\left(a_{i+1,j} \times a_{i,j+1}\right)$$

$$F = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1}\left(a_{i,j} \times a_{i,j+1} \times a_{i+1,j} \times a_{i+1,j+1}\right)$$

$$T = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} 1 \left| \left(a_{i,j} \times a_{i,j+1} + a_{i,j} \times a_{i+1,j} + a_{i+1,j+1} \times a_{i+1,j} + a_{i+1,j+1} \times a_{i,j+1}\right) = 2\right.$$

Figure 7: equation for euler number

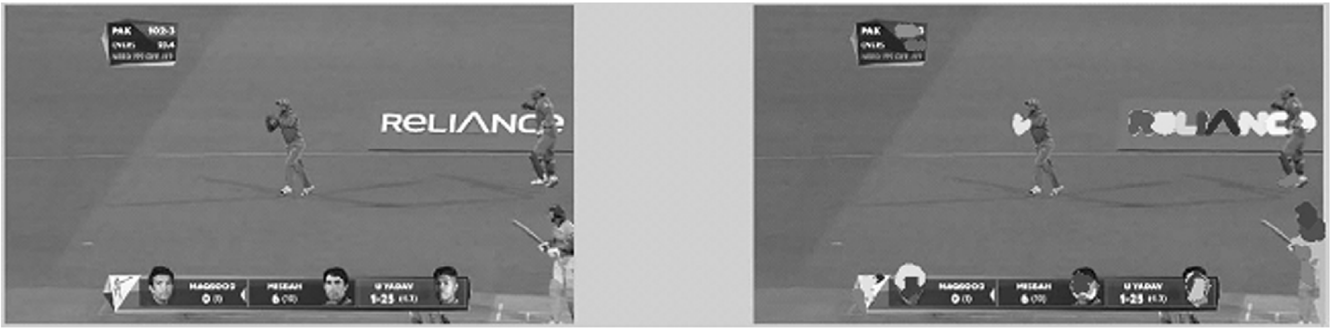


Figure 8: Video frame 1 showing actual video frame with detected text area in that frame

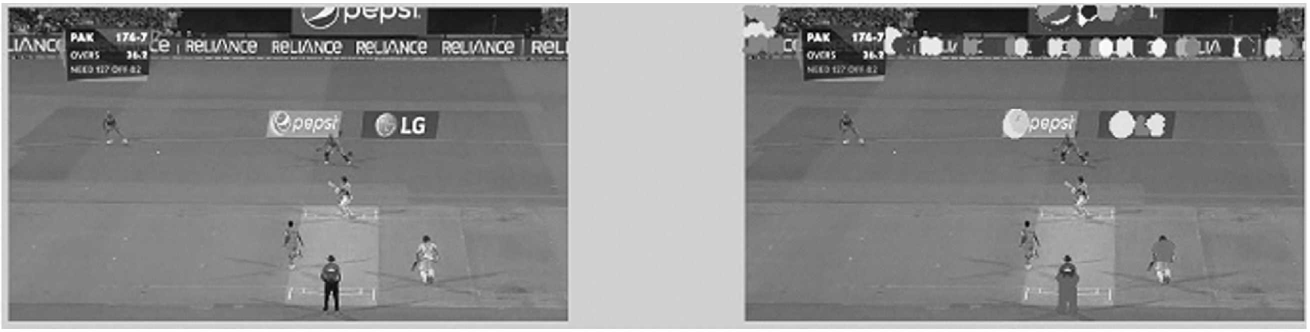


Figure 9: Video frame 2 showing actual video frame with detected text area in that frame

time. Completing the text area determination in relatively less time requires optimal algorithm development for specific domain of video. In order to detect text area each frame of video is analyzed fast with optimized algorithm using MSER properties followed by geometric properties. Figure 8, and figure 9 shows text area detected from a frame of video using our proposed method. This picture clearly detects text area from the scene text and as well as overlaid text.

In the both image part (a) that is 8.a and 9.a shows actual video frame and part (b) that 8.b and 9.b shows the gray-scale image with detected text area using MSER method followed by filtering the area using geometric properties analysis.

5. TEXT DETECTION IN LIVE SPORTS EVENT VIDEO

After detecting text area from the video frame it is further processed to detect text from these detected text areas. For this further stroke width is determined and compared with other text area to determine refine the

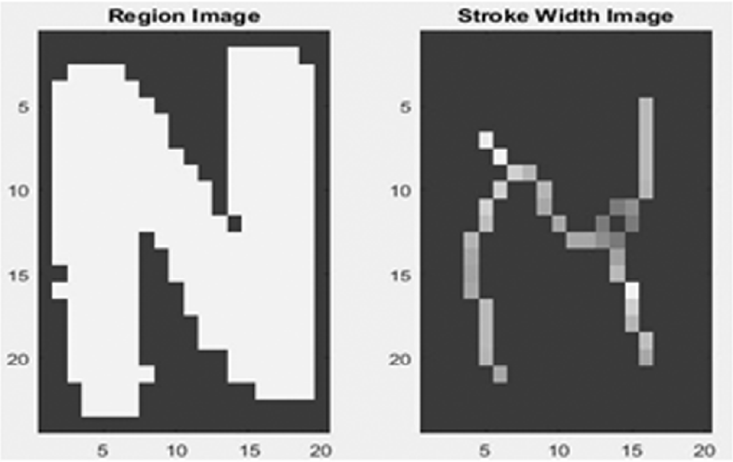


Figure 10: stroke width calculation for text detection

text area. Latter using property of stroke width and it variation, actual text couture is defined, shown in figure 10, to further determine actual text. As SWT method is not alone sufficient to detect text area in dynamic situation as describe here, machine learning approach is further used to recognize text area and non text area. In our proposed framework, we have used SVM classifiers for final classification of text area and non text area.

6. RESULT ANALYSIS

After detecting the text from video stream it is analyzed and compared with recently published research method and its result found assertive. A set of sports video mainly cricket video is used for testing purpose which has been taken from youtube. For training and testing purpose we have used ICDAR 2015 Dataset available for text detection from video stream of different qualities along with self created dataset of cricket only. Details of the testing video file are shown in table1.

Sports video used for this result analysis is ICC World Cup cricket match highlight video having frame rate of 25fps. Each frame had resolution of 1280 × 714 pixels with data rate of 1580 kbps. As most of the

Description	*Frame Rate*	*Resolution*	*Data rate*
ICC World Cup cricket match highlight video	25fps	1280 × 714 pixels	1580 kbps
Asia Cup Match Highlights- Bangladesh vs. India	25fps	1280 × 720 pixels	1260 kbps
Champions trophy - India vs. South Africa	25 fps	1280 × 720 pixels	1697 kbps
ICDAR 2015 Dataset -	30 fps	720 × 480 pixels	1474 kbps
ICDAR 2015 Dataset -	24 fps	1280 × 960 pixels	7567 kbps
ICDAR 2015 Dataset -	30 fps	1280 × 720 pixels	2867 kbps
ICDAR 2015 Dataset -	24 fps	1280 × 720 pixels	1807 kbps

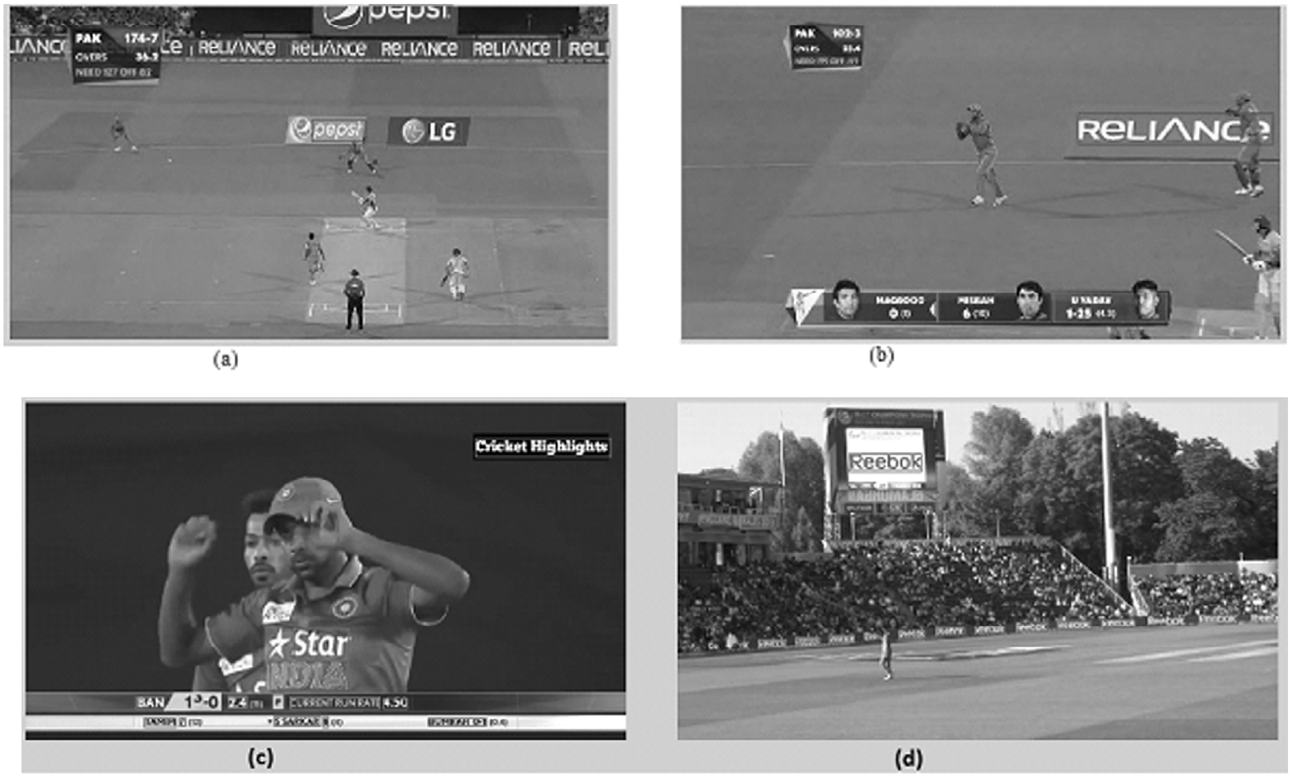


Figure 11. Detected scene text in cricket video frames.
(a), (b) is taken from India vs Pakistan ICC World Cup cricket match highlight video frame,
(c) shows a test frame from Bangladesh vs. India Asia Cup Match Highlights,
(d) shows a test frame from India vs. South Africa ICC Champions trophy.

existing method for detection from a video is based on overlaid text only and not for naturally occurring text (scene text), our proposed framework result sits with distinction (due to scene text detection from sports video). Figure 11 shows the detected scene text from cricket video event.

Qualities of text detected from the live sports video is measured in the term of standard matrices precision and recall. Result comparison table represents precision and recall for the proposed method in table 2. It describes the precision and recall for overlaid text and scene text. Further its efficiency is also compared with other existing methods. Although it takes little more time than the time required to process the video frames in real time, experimental result showed great robustness and its efficiency is appreciable. It is able to detect the text in cricket sports video. Although presented framework is tested on a limited dataset, further improvement in its algorithm is required to detect text from other sports video in real time. Test dataset for this framework required to be of high quality (minimum720p). Video frame size, sport domain, and text language are main parameters which affect the quality of detected text i.e. precision and recall of the detected text. The text detection framework discussed in this paper, has been tested/validated on cricket domain video only and therefore further improvement in terms of video quality and video type is left for future work.

Text Type	*Avg. Precision*	*Avg. Recall*
Overlaid Text	0.9	0.8
Scene Text	0.85	0.7
Text on moving Player	0.7	0.65

REFRENCES

[1] Haojin Yang, Bernhard Quehl, Harald Sack, A framework for improved video text detection and recognition. Multimedia Tools Appl , Volume:69, pp217–245, (2014)

[2] Weinman, J.J.; Butler, Z.; Knoll, D.; Feild, J. “Toward Integrated Scene Text Reading”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Volume: 36, Issue: 2, pp. 375 – 387, (Feb. 2014)

[3] Neumann, L.; Matas, J. “Scene Text Localization and Recognition with Oriented Stroke Detection”, Computer Vision (ICCV), IEEE International Conference on, pp. 97 – 104, (2013)

[4] Chucai Yi; YingLi Tian “Localizing Text in Scene Images by Boundary Clustering, Stroke Segmentation, and String Fragment Classification”, Image Processing, IEEE Transactions on, Volume: 21, Issue: 9, pp. 4256 – 4268, (Sept. 2012)

[5] Chucai Yi; YingLi Tian, “Text String Detection From Natural Scenes by Structure-Based Partition and Grouping”, Image Processing, IEEE Transactions on, Volume: 20, Issue: 9, pp. 2594 – 2605, (Sept. 2011)

[6] Anthimopoulos M, Gatos B, Pratikakis I, “A two-stage scheme for text detection in video images”. J Image Vison Computing, Volume:28, pp. 1413–1426, (2010)

[7] Sun L, Liu G, Qian X, Guo D, “A novel text detection and localization method based on corner response”. in Proc ICME, (2009)

[8] Nister, D. and Stewenius, H., “Linear Time Maximally Stable Extremal Regions”, ECCV, (2008).

[9] Zhang J, Goldgof D, Kasturi R, “A new edge-based text verification approach for video”. in Proc. ICPR, (2008)

[10] Lu S, Barner K, “Weighted DCT coefficients based text detection”. in Proc. ICASSP, pp. 1341-1344, (2008)

[11] Halin AA, Rajeswari M, Ramachandram D, Automatic overlaid text detection, extraction and recognition for high level event/concept identification in soccer videos.”, International conference on computer and electrical engineering, pp 587-592, (2008)

[12] Qian X, Liu G, Wang H, Su R, “Text detection, localization and tracking in compressed videos”. Signal Processing: image Communication Volume:22, No: 9, pp. 752–768, (2007)

[13] Lyu M, Song J, Cai M, “comprehensive method for multilingual video text detection, localization, and extraction” IEEE Trans Circuits and Systems for Video Technology Volume:15, No:2, pp. 243–255, (2005)

[14] Ngo C, Chan C, “Video text detection and segmentation for optical character recognition”. Multimedia Systems Volume:10, No:3, pp. 261–272, (2005)

[15] Jung K, Kim K, Jain A, “Text information extraction in images and video: a survey. Pattern Recognition” Volume:37, pp. 977–997, (2004)

[16] Lee C, Jung K, Kim H, “Automatic text detection and removal in video sequences”. Pattern Recogn Lett Volume:24, pp. 2607–2623, (2003)

[17] Tang, X., Gao, X., Liu, J., “A Spatial-Temporal Approach for Video Caption Detection and Recognition”. IEEE Trans On Neural Networks, special issue on Intelligent Multimedia Processing, pp. 961–971 (2002);

[18] Zhang, H.J.: “Content-based video analysis, retrieval and browsing”. Microsoft Research Asia, Beijing (2001)

[19] Chen, D., Bourlard, H., Thiran, J.-P.: “Text Identification in Complex Back-ground Using SVM”. In: CVPR, vol. II, pp. 621–626 (2001)

[20] Zhong Y, Zhang H, Jain A, “Automatic caption localization in compressed video”. IEEE Trans Pattern Analysis and Machine Intelligence Volume:22, No:4, pp. 385–392, (2000)

[21] Sato, T., Kanade, T., Kughes, E.K., Smith, M.A., Satoh, S.: “Video OCR: Indexing digital news libraries by recognition of superimposed captions”. ACM Multimedia Syst (Special Is-sue on Video Libraries) Volume:7, No:5, pp. 385–395, (1999)

[22] Li, H.P., Doemann, D., Kia, O.: “Text extraction, enhancement and OCR in digital video”. In: Proc. 3rd IAPR Workshop, Nagoya, Japan, pp. 363–377, (1998)