Automatic Video Scene Segmentation to Separate Script for OCR

Bharatratna P. Gaikwad
Department of CS and IT
Dr. B. A.M. University,
Aurangabad (MS),India

Ramesh R. Manza
Department of CS and IT
Dr. B. A.M. University,
Aurangabad (MS), India

Ganesh Manza
Department of CS and IT
Dr. B. A.M. University,
Aurangabad(MS),India

ABSTRACT
In Text or character recognition in images or video frames is a difficult problem to achieve video data. This paper proposes improved template matching algorithm that applied for the automatic extraction of text from image and video frames. Optical character recognition using template matching is a system model that is useful to recognize the character, digits& special character by comparing two images of the alphabet. The objectives of this system model are to develop a model for the Optical Character Recognition (OCR) system and to implement the template matching algorithm in developing the system model .The template matching techniques are more profound to font and size variations of the characters than the feature classification methods. This system tested the 35 videos with 700 video frames for each video. Empirical result of this system precision rate is 91.52% for automatic character gets recognized images and video frames. Experimental results show the relatively high accuracy of the new developed robust algorithm when it is tested on several size characters and text.

General Terms
Video Processing with using template matching algorithm

Keywords
Video Processing, text detection, localization, tracking, segmentation, Template Matching, OCR

1. INTRODUCTION
The Popularity of digital video is emerging at an explosive rate. The rapid growth of video data leads to an urgent demand for efficient and true content-based browsing and retrieving systems. In response to such needs, various video content analysis schemes are using with one or a combination of image, audio, and textual information in the video [1]. A variety of approaches to text information extraction from images and video have been proposed for specific applications including page segmentation , address block location , license plate location , and content-based image/video indexing [2].In the extraction of this information involves the detection, localization, tracking, extraction, enhancement and recognition of text from the images and video frames are provided. Text in images and video frames carries important information for visual content understanding and retrieval. Optical character recognition (OCR) is one of the most popular areas of research in pattern recognition because of its immense application potential. The two fundamental approaches to OCR are template matching and feature classification. In the template matching approach, recognition is based on the correlation of a test character with a set of stored templates. In the feature classification method, features are extracted from a standard character image to generate a feature vector. A decision tree is formed based on the presence or absence of some of theelements in the feature vector. When an unknown character pattern is encountered, this tree is traversed from node to node till a unique decision is reached. The template matching techniques are more sensitive to font and size variations of the characters than the feature classification methods. However, selection and extraction of useful features is not always straight forward [5].

2. VIDEO PROCESSING
A. VideoProcessing Shot: Frames recorded in one camera operation form a shot.
Scene: One or several related shots are combined in a scene.
Sequence: A series of related scenes forms a sequence.

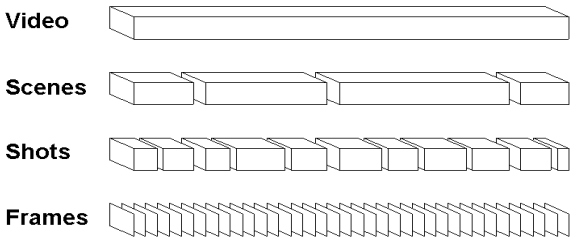


Fig 1: Videos Scenes sequences

Video: A video is composed of different story units such as shots, scenes, and sequences arranged according to some logical structure (defined by the screen play). These concepts can be used to organize video data. The video consists of sequence of images (video frames). In the first step, we convert video into all frames and saved as JPEG images [7].Several software is available for editing and shows the videos as following are extension of video types its extension and acronym.

Table 1.List of different videos formats

Sr.No.	Extension	Specifics
1	.3G2	3G Mobile Phone Video
2	.3GP	3G Mobile Phone Video
3	.ASF	Advanced Systems Format Video
4	.AVI	Audio Video Interleave
5	.BIK	Bink Video
6	.BIN	Binary DVD Video
7	.DAT	VCD Video
8	.DIVX	DivX Movie

9	.DV	Digital Video
10	.FLV	Flash Video
11	.GXF	General eXchange Format
12	.M1V	MPEG-1 Video
13	.M2TS	MPEG-2 Transport Stream Videos
14	.MPEG2	MPEG-2 Video
15	.MPEG4	MPEG-4 Video

Extracting text information from videos generally involves three major steps:

• Text detection: Find the regions that contain text.

• Text segmentation: Segment text in the detected text regions. The result is usually a binary image for text recognition.

• Text recognition: Convert the text in the video frames into ASCII characters.

3. VIDEO TEXT DETECTION AND ANALYSIS

3.1 Video Text Recognition Scheme

A. Pre Processing

A scaled image was the input which was then converted into a gray scaled image. This image formed the first stage of the pre-processing part. This was carried out by considering the RGB color contents of each pixel of the image and converting them to grayscale. The conversion of a colored image to a gray scaled image was done for easier recognition of the text appearing in the images as after grayscale conversion, the image was converted to a black and white image containing black text with a higher contrast on white background [12].

B. Detection and Localization

In the text detection stage, since there was no prior information on whether or not the input image contains any text, the existence or nonexistence of text in the image must be determined. However, in the case of video, the number of frames containing text is much smaller than the number of frames without text. The text detection stage seeks to detect the presence of text in a given image. Text localization methods can be categorized into two types: region-based and texture-based. Select a frame containing text from shots elected by video framing, this stage used region Based Methods for text tracking. Region based methods use the properties of the color or gray scale in a text region [1] [24].

C. Tracking and Segmentation

When text was tack, the text segmentation step deals with the separation of the text pixels from the background pixels indirectly separate single character from whole text. The output of this step is a binary image where black text characters appear on a white background. This stage included extraction of actual text regions by dividing pixels with similar properties into contours or segments [2][9][22].

D. Recognition

This stage included actual recognition of extracted characters , The result of recognition was a ratio between the number of correctly extracted characters and that of total characters and evaluates what percentage of a character were extracted correctly from its background. For each extraction result of correct character [4][25].

3.2 Survey of literature

Please use a 9-point Times Roman font, or other Roman font

1) Jie Xi and et.al. has work on Text detection, tracking and recognition to extract the text information in news and commercial videos. He has used Techniques morphological opening procedure on the smoothed edge map. They got the text detection rate is 94.7% and the recognition rate is 67.5% [7].

2)PalaiahnakoteShivakumara and et.al. has work on elimination of non-significant edges from the segmented text portion of a video frame to detect accurate boundary of the text lines in video images. They got percentage 93% [8].

3)RainerLienhart and et.al. has worked on the text localizing and segmenting text in complex images and videos, It is able to track each text line with sub-pixel accuracy over the entire occurrence in a video. They got percentage text recognition 69.9% [9].

4)Qixiang Ye and et.al.has worked on the detection and verification of English Text and Chinese text from images and video frames, He has used Techniques for detection is based on Sobel edges feature and the verification uses the wavelet-based features and svm classifier. They got percentage detection rate English 93.9% [10].

Y. Zhong and et.al. has worked on to automatically localize captions in JPEG compressed images and the I-frames (intra-frames) of MPEG compressed videos. Caption text regions are segmented from background images. He has used Techniques discrete cosine transform (DCT) to use the coefficients directly from compressed images and video as texture features to localize text regions. The texture features are extracted directly from domain using the quantized DCT coefficients and morphological operation are used to remove isolated noisy block and merge disconnected text blocks. The Limitation of their work is the font size of characters or the gap between the characters in the text is too big such that there is no strong texture present in a MPEG block and the contrast between the background and the text is too weak so that the text energy is not sufficiently high.According to them this work may extends in future to the compressed domains of the two color frames (Cr and Cb) to extract text with high color contrast. If segment text with a larger font size to compute texture features at a larger scale [13].

A chronological listing of some of the published work on different approaches used for text extraction is presented in Table 2[17].In this table 2 the edge base and texture base recognition rate get escalation as comparatively to region based.

Table 2.listing the survey of text extraction using different approaches

Technique	Author	Year	Title	Method	Accuracy (%)
Region-Based	Leon	2009	Caption text extraction for indexing purposes using a hierarchical region-based image model	Hierarchical region-based	86.3
	Debapratim	2009	A Bottom -Up Approach of Line Segmentation from Handwritten Text.	Bottom-Up Approach of Line Segmentation.	92
Edge Based	XinZhang	2010	A Combined Algorithm for Video Text Extraction	Transition Map, Canny Operator	90.74
	Xiaoqing Liu	2006	Multiscale Edge-Based Text Extraction From Complex Images	Multiscale Strategy, Clustering	96.6
Texture Based	Chu Duc	2009	Robust Car License Plate Localization using a Novel Texture Descriptor	Hough transform	96.7
	Bassem	2006	A New Approach For Texture Features Extraction: Application For Text Localization In Video Images.	Hough Transform technique combined with a next remitysegment's neighborhood analysis.	96

4. METHODOLOGY

4.1 Canny Edge Detector

Among the several textual properties in an image, edge-based methods focus on the 'high contrast between the text and the background'. The edges of the text boundary are identified and merged, and then several heuristics are used to filter out the non-text regions. Usually, an edge filters (e.g. canny operator) is used for the edge detection, and a smoothing operation. The Canny method finds edges by looking for local maxima of the gradient of I. The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds, to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges [13][15]. This method is therefore less likely than the others to be fooled by noise, and more likely to detect true weak edges [3][16].

$G_X =$

-1	-2	-1
0	0	0
1	2	1

$G_Y =$

-1	0	+1
-2	0	+2
-1	0	+1

Fig 2: Canny Edge detection operator (a) x direction (b) y direction

The canny edge detection algorithm is easy to implement, and more efficient than other algorithms. From this edge detected images, text region is identified [3].

Table 3. Different properties of text in images and Video frames

Category	Detail Properties	Sub-classes
Geometry	Size Regularity in size of text	Font Size, width, height, Bold, Italic.
	Alignment	Horizontal/vertical/center
		Straight line with skew (implies vertical direction)
		Curves
		3D perspective distortion
	Inter-Character distance	Aggregation of characters with uniform distance
	Strokes	Different stroke density and statistics
Color		Gray
		Color (monochrome, polychrome)
		Text color is dark or light
Motion		Static
		Linear movement
		Static
		Free movement
Edge		Strong contrast (edges) at text boundaries
Compression		Un-compressed image
		JPEG, MPEG-compressed image.

1. Compute f_x and f_y

$$f_x = \frac{\partial}{\partial x}(f * G) = f * \frac{\partial}{\partial x} G = f * G_x (1)$$

$$f_y = \frac{\partial}{\partial y}(f * G) = f * \frac{\partial}{\partial y} G = f * G_y \quad (2)$$

$G(x, y)$ is the Gaussian function $G_x(x, y)$ is the derivate of

$G(x, y)$ with respect to x:

$$G_x(x, y) = \frac{-x}{\sigma^2} G(x, y) (3)$$

$G_y(x, y)$ is the derivate of $G(x, y)$ with respect to y:

$$G_y(x, y) \frac{-y}{\sigma^2} G(x, y)$$

(4)2. Compute the gradient magnitude magn(i, j)

$$= \sqrt{f_x^2 + f_y^2} \quad (5)$$

3. Apply non − maxima suppression.
4. Apply hysteresis thresholding / edge linking .

4.2 The basic steps of the connected-component text extraction algorithm

1. Convert the input frame or image into a gray image.
2. Convert the gray image into an edge base image by using edge base operator as Canny and sobel. With using this operator so we get accurate detecting shape of the character and entire text.
3. Compute the horizontal and vertical projection profiles of candidate text regions using a histogram with an appropriate threshold value.
4. Use geometric properties of text such as width to height ratio of characters to eliminate possible non-text regions.
5. Binarize the edge image enhancing only the text regions against a plain white background with object black colors.
6. When the sentences finish in image then given space at requisite locations by applying space vector to differentiate the line for proper result as similar to the input image space in two lines [19].

Text in images and video frames can exhibit many several properties in the table 3.

Table 3 shows a list of some properties that have been we used in our research work as Edge- Maximum timestext and scene text are appear in video frames are unclear edges so we can apply edge detector operator for easily detect edge and increase text borderline, thereby resulting in strong edges at the boundaries of text and background.

4.3 Design a System and Implementing Algorithm

The template matching worked on this following Algorithm
a) Load the video (E.g. Avi, Mpeg etc.).

b) Then video is converted into frames with frames name

from"img-1 to Img-N "till the video will be come to an end.

c) Template is made of Upper case, Lower case, Special character & digit with size 24x42 size.

d) Applying OCR techniques, select the frame among one of them (E.g.img-50).

e) Image is Converted to gray scale and then converted to binary.

f) Then top-down: extracting texture features of the image and then locating text regions only from complete frames and non-text regions is skip on the basis on texture properties.

g) Bottom-up: separating the image into small regions and then grouping character regions into text regions.

h) The character image from the detected string is selected.

i) Segmentation: Segment those regions in the image which are actually text characters are present.Each character was automatically selected as per size of template design.

j) After that, the image to the size of the first template is rescaled.

k) After rescale the image to the size of the first original image then comprising letters with template and the matching metric is computed.

l) Then the highest match found is stored. If the template image is not match, it might be getting recognized as some other character.

m) The index of the best match is stored as the recognized character.

n) All recognized character showing on Word file.

4.4 Template Matching Method

The template matching technique is a process in video processing for outcome small parts of an image which match a template image. The corrected text region (TR) is binarized using a simple yet efficient binarization technique developed by us before segmenting it. The algorithm has been given below. In his method, the arithmetic mean of the maximum (G_{max}) and the minimum (G_{min}) gray levels around a pixel is taken as the threshold for binarizing the pixel. In the present algorithm, the eight immediate neighbors around the pixel subject to binarization are also taken as deciding factors for binarization. This type of approach is especially useful to connect the disconnected foreground pixels of a character [19].

```
Begin
for all pixels (x, y) in a TR

    if intensity(x, y) < (G min + G max )/2, then

       mark (x, y) as foreground
     else
          if no. of foreground neighbors > 4, then
mark (x, y) as foreground
          else
              mark (x, y) as background
          end if
          end if
end for
```

Architecture of Video Scene Segmentation to Separate Script for OCR

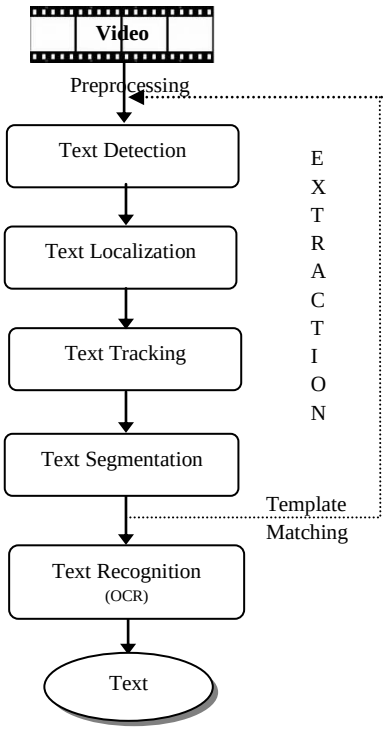
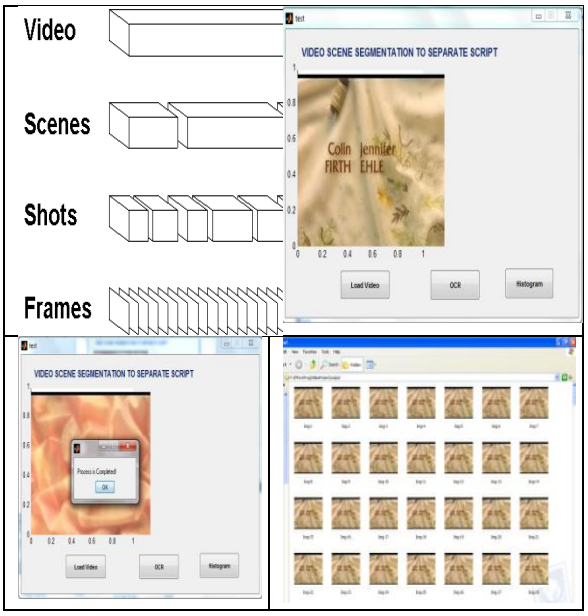


Fig 3: System for text detection and recognition from video/image

Above figure 3 shows the outline of workflow of our system for automatic Video Scene Segmentation to Separate Script for OCR.

Followingin figure 4 are the screenshot of our system for **Load Video**:-This Button is used for Loading the Video of (*.MPEG,*.AVI) and Converted into frames at destination folder, every frame save with extension JPG. **OCR**: This Button is used for optical character recognition; open Frame files then detect the character, Tracking, segmentation and lastly recognition i.e. OCR.



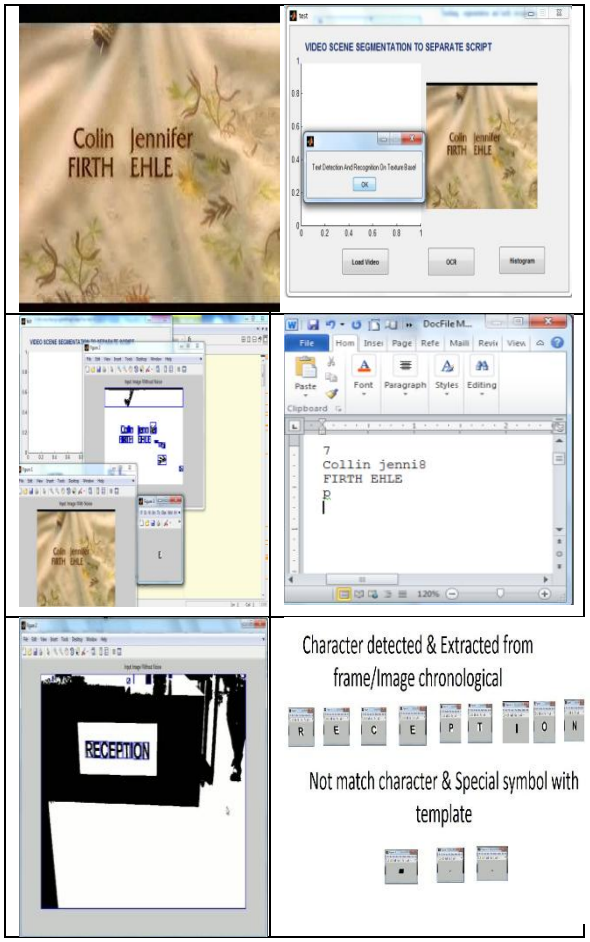


Fig 4: (a) Units for Video data (b) Load video system (c) Video process completed (d) Video into frames (eg.Img-1 to N) (e)Select Video frame (eg.img-50) (f) Detect text (g) Text tracking & Segment texte(h)Character recognizes in word file (i) Select other video frame (eg.Img-150) (j)Text or Character get recognizes.

4.5 Character Recognition System

Among the 256 ASCII characters, only 94 are used in document images or frame and among these 94 characters, only 80 are frequently used. In the present scope of experiment, we have considered 80 classes recognition problem. These 80 characters are listed in Table 4. These include 26 capital letters, 26 small letters, 10 numeric digits and 18 special characters table 4. These include 26 capital letters, 26 small letters, 10 numeric digits and 18 special characters [26] [21].

Table 4. Videos Frame Template Classes

Images/Frames	700
Text Lines	850
Correct Detected	778
FalsePositives	72
Recall (%)	90.92%
Precision (%)	91.52%

5. PERFORMANCE EVALUATION

There are several performance evaluations to estimate the algorithm for text extraction. Most of the approaches quoted here used Precision, Recall to evaluate the performance of the algorithm. Precision, Recall rates are computed based on the number of correctly detected characters in an image, in order to evaluate the efficiency and robustness of the algorithm. The performance metrics are as follows:

$$\text{Recall} = \frac{\text{Correct Detected}}{(\text{Correct Detected} + \text{Missed Text Lines})} \quad (7)$$

Whereas precision is defined as:

$$\text{False alarm rate} = \frac{\text{Number of falsely detected text}}{\text{Number of detected text}} \quad (8)$$

$$\text{Precision} = \frac{\text{Correct Detected}}{(\text{Correct Detected} + \text{False Positives})} \quad (9)$$

False Negatives

False Negatives (FN)/ Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

Precision rate

Precision rate (P) is defined as the ratio of correctly detected characters to the sum of correctly detected characters plus false positives.

Recall rate

Recall rate (R) is defined as the ratio of the correctly detected characters to sum of correctly detected characters plus false negatives.

The following table compares recognition result of improved template matching method and traditional template matching method, the test result is shown in below table 6.

Table 6.Characters Recognition Test Table

Test group	Recognition Result
Uppercase	92.00%
Lowercase	91.08%
Digits	92.00%
Special Character	91.00%

Every character set the textbox as per the character, digit, and special character size is detected correctly, all character is completely surrounded by a box, some character is not match with template data set then showing other character, so a detected text box is considered as a false alarm, if no text appears in that box.

The text localization algorithm achieved a recall of 90.92% and a precision of 91.50%.As seen from the table 5 & 6, using the improved template matching method, the average recognition rate and Recognition speeds of upper, lower letters, numeric and special characters have been enhanced.

1	2	3	4	5	6	7	8	9	0
A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z	a	b	c	d
e	f	g	h	I	j	k	l	m	n
o	p	q	r	S	t	u	v	w	x
y	z	“	;	,	.	#	&	@	(
)	-	%	!	:	‘	$	?	+	/

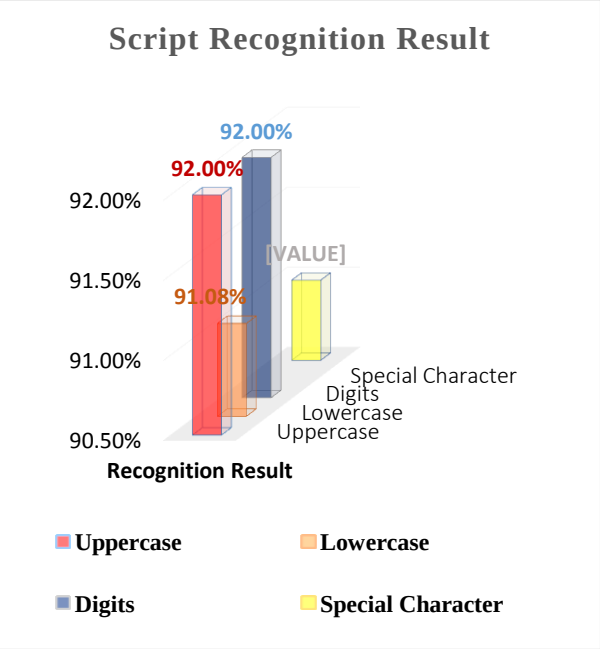


Fig 5: Script Recognition Rate from video frames

Graphically representation of script recognition rate shown in figure 4 with upper, lower case .digit and special character.

6. CONCLUSION

There are many cases this system are useful for video text information extraction system, vehicle license plate extraction, text based video indexing, video content analysis and video event identification .In this work, we have new approach for character recognition system based on template matching. This system tested the 35 videos with 700 video frames and 850 text lines .The system is texture-based approaches to automatic detection, segmentation and recognition of visual text occurrences in images and video frames. The characters are recognized automatically on run-time basis, In a few cases in which 8.50% characters could not get detected but some other character get recognized. The overall empirical performance of this system recognizing rate is 91.52% successfully.

7. REFERENCES

[1] Xian-Sheng Hua, Liu Wenyin, Hong-Jiang Zhang.2001,” Automatic Performance Evaluation for Video Text Detection,” Sixth InternationalConference on Document

Analysis and Recognition (ICDAR2001), pp 545-550, Seattle, Washington, U.S.A., September 10-13, (2001).

[2] KeechulJunga, Kwang In Kimb, Anil K. Jain.2003"Text information extraction in images and video: a survey,"Published by Elsevier Ltd.(2003).

[3] Canny, J.1986, "A Computational Approach to Edge Detection,"IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, November (1986).

[4] H.K. Kim, ECcien.1996," automatic text location methodand content-based indexing and structuring of video database,"J. Visual Commun. Image Representation 7 (4) 336–344(1996).

[5] Y. Zhong, A.K. Jain.2000," Object localization using color, texture, andshape," Pattern Recognition 33 671–684(2000).

[6] S. Antani, R. Kasturi, R. Jain.2002," A survey on the use of pattern recognition methods for abstraction, indexing, and retrieval ofImages andvideo, "Pattern Recognition 35 945–965(2002).

[7] Xi Jie, Xian-Sheng Hua, Xiang-Rong Chen, Liu Wenyin, HongJiang Zhang.2009" A Video Text Detection and Recognition System, "IEEE International(2009).

[8] P. Shivakumara, W. Huang and C. L. Tan.2008: Efficient Video Text Detection Using Edge Features, The Eighth IAPR Workshop on Document Analysis Systems (DAS2008), Nara, Japan, pp 307-314(2008).

[9] Rainer Lienhart and Frank Stuber,"Automatic text recognition in digital videos," University of Mannheim, PraktischeInformatik IV, 68131 Mannheim, Germany.

[10] Qixiang Ye, W. Gao, W. Wang and W. Zeng.2003,"A Robust Text Detection Algorithm in Images and Video Frames," IEEE ICICS-PCM, pp. 802-806, (2003).

[11] G. Aghajari, J. Shanbehzadeh, and A. Sarrafzadeh.2010," A Text Localization Algorithm in Color Image via New Projection Profile,"IMECS Hong Kong (2010).

[12] JayshreeGhorpade, Raviraj Palvankar.2011,"Extracting Text From Video," Signal & Image Processing , An International Journal (SIPIJ) Vol.2, No.2, (2011).

[13] Bharatratna GaikwadRamesh R. Manza.2011," Critical review on video scene segmentation and Recognition ," International Journal of Computer Information Systems (IJCIS), Vol 3, and Number 3, (2011).

[14] Ramesh R. Manza and Bharatratna P. Gaikwad.2012,"A Video Edge Detection Using Adaptive Edge Detection Operator," Issue: January 2012, DOI: DIP012012006, CiiT International Journal of Digital Image Processing: ISSN: 0974–9691 & Online: ISSN: 0974-9586.

[15] Manza R.R., GaikwadB.P., Manza G.R. 2012,"Use Of Edge Detection Operators For Agriculture Video Scene Feature Ex-Traction From Mango Fruits," Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Vol 4, Issue 1, 2012, pp.-50-53.

[16] Manza Ramesh R., Bharatratna P. Gaikwad, Manza Ganesh R.2012,"Used of Various Edge Detection Operators for Feature Extraction in Video Scene," ICACEEE-Jan-2012Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering ,ISBN: 978-981-07-1847-3(2012).

[17] C.P.Sumathi,T. Santhanam, N. Priya.2012 ,"Techniques and challenges of automatic text extraction in complex images: a survey", Journal of Theoretical and Applied Information Technology, 31st January 2012. Vol. 35 No.2

[18] Keechul Jung, Kwang In Kim, Anil K. Jain.2004,"Text Information Extraction in Images and Video: A Survey ",the journal of the Pattern Recognition society, 2004.

[19] SnehaSharma.2006,"Extraction of Text Regions in Natural Images",Masters Project Report, Spring 2006/07.

[20] AyatullahFarukMollah,,NabamitaMajumder.2011, "Design of an Optical Character Recognition System for Camerabased Handheld Devices ",IJCSI ,Issues, Vol. 8, Issue 4, No 1, July 2011.

[21] Yih-Ming Su, Chaur-Heh Hsieh.2006, "A Novel Model-Based Segmentation Approach To Extract Caption Contents On Sports Videos", IEEE International Conference On Multimedia And Expo,pp:1829 - 1832 .

[22] Miriam Leon, Veronica Vilaplana, AntoniGasull, Ferran Marques.2009 , "Caption Text Extraction For Indexing Purposes Using A Hierarchical Region-Based Image Model",Proceedings Of The 16th IEEE International Conference On Image Processing, pp:1869-1872.

[23] u Zhong, Hongjiang Zhang, And Anil K. Jain.1999,"Automatic Caption Localization InCompressed Video", International Conference On Image Processing, pp: 96 - 100 Vol.2.

[24] XiaoqianLiu,Weiqiang Wang.2010 ,"Extracting Captions From Videos Using TemporalFeature",Proceedings Of The International Conference On Acm Multimedia ,pp:843-846.

[25] Bo Lilo, Xaoou Tang, Jianzhuang Liu, And Hongiiang Zhan.2003 ,"Video Caption Detection And Extraction Using Temporal Information", International Conference On Image Processing, Vol.1 , pp:I 297-300

[26] Bharatratna P. Gaikwad , Ramesh R.Manza,Manza R. Ganesh.2013 , "Video scene segmentation to separate script", Advance Computing Conference (IACC), 2013 IEEE xploreieee , 978-1-4673-4527-9.