Automatic Video Scene Segmentation to Separate Script and Recognition

Bharatratna P. Gaikwad, Ramesh R. Manza, and Ganesh R. Manza

Department of CS and IT, Dr. B.A.M. University, Aurangabad (MS), India
{bharat.gaikwad08,manzaramesh,ganesh.manza}gmail.com

Abstract. Text or character detection in images or videos is a challenging problem to achieve video contents retrieval. In this paper work we propose to improved VTDAR (Video Text Detection and Recognition) Template Matching algorithm that applied for the automatic extraction of text from image and video frames. Video Optical Character Recognition using template matching is a system model that is useful to recognize the character, upper, lower alphabet, digits& special character by comparing two images of the alphabet. The objectives of this system model are to develop a model for the Video Text Detection and Recognition system and to implement the template matching algorithm in developing the system model. The template matching techniques are more sensitive to font and size variations of the characters than the feature classification methods. This system tested the 50 videos with 1250 video key-frames and text line 1530. In this system 92.15% of the Character gets recognized successfully using Texture-based approaches to automatic detection, segmentation and recognition of visual text occurrences in images and video frames.

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

1 Introduction

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]. All are types of video file formats are available on internet, cell phones and easily downloaded. 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. 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 [2]. Optical character recognition (OCR) is one of the most popular areas of research in pattern recognition because of its immense

S.C. Satapathy et al. (eds.), *Proc. of the 3rd Int. Conf. on Front. of Intell. Comput. (FICTA) 2014 – Vol. 2*, Advances in Intelligent Systems and Computing 328, DOI: 10.1007/978-3-319-12012-6_25

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 the elements 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]. Several software is available for editing and shows the videos types as .AVI,.FLV,.DAT,.3GP,.MPEG,.MP4 etc.
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.

1.1 Video Text Detection and Analysis

Video Processing:-
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.
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.

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], [19], [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], [21], [25].

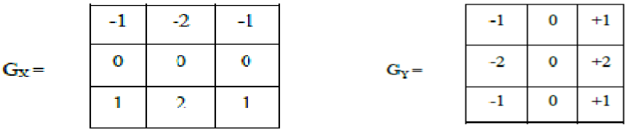
1.2 Survey of Literature

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) Palaiahna kote Shivakumara 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) Rainer Lienhart 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][10].

2 Methodology

2.1 Figures 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].



$$G_X = \begin{array}{|c|c|c|}\hline -1 & -2 & -1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 2 & 1 \\ \hline \end{array} \qquad G_Y = \begin{array}{|c|c|c|}\hline -1 & 0 & +1 \\ \hline -2 & 0 & +2 \\ \hline -1 & 0 & +1 \\ \hline \end{array}$$

Fig. 1. Canny Edge detection operator (a) x direction (b) y direction

1. Compute f_x and f_y

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

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

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

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

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

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

$$G_y(x, y)\frac{-y}{\sigma^2}G(x, y) \tag{4}$$

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

$$= \sqrt{f_x^2 + f_y^2} \tag{5}$$

3. Apply non – maxima suppression.

4. Apply hysteresis thresholding / edge linking .

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],[23].Text in images and video frames can exhibit many variations with respect to the following properties are character font size, width ,hight,alignment,edge,color etc.

2.2 Design a System and Implementing VTDAR Algorithm

The Video Text Detection and Recognition template matching worked on this following Algorithm (VTDAR)

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 by using CC algorithm.

f) Applying edge detection& Binarization algorithm for focuses on text region.
g) Then top-down: extracting texture features of the image and then locating text regions.
h) Bottom-up: separating the image into small regions and then grouping character regions into text regions.
i) Applying simultaneously by space vector for maintain space between two lines as per Image
j) The character image from the detected string is selected.
k) Segmentation: Each character was automatically selected and thresholding using methods.
l) After that, the image to the size of the first template is rescaled.
m) After rescale the image to the size of the first original image then comprising letters with template matching techniques are used and the matching metric is computed.
n) Then the highest match found is stored. If the template image is not match, it might be getting recognized as some other character.
o) The index of the best match is stored as the recognized character.
p) All recognized character showing on Word file.

Architecture of Video Scene Segmentation and Recognition system

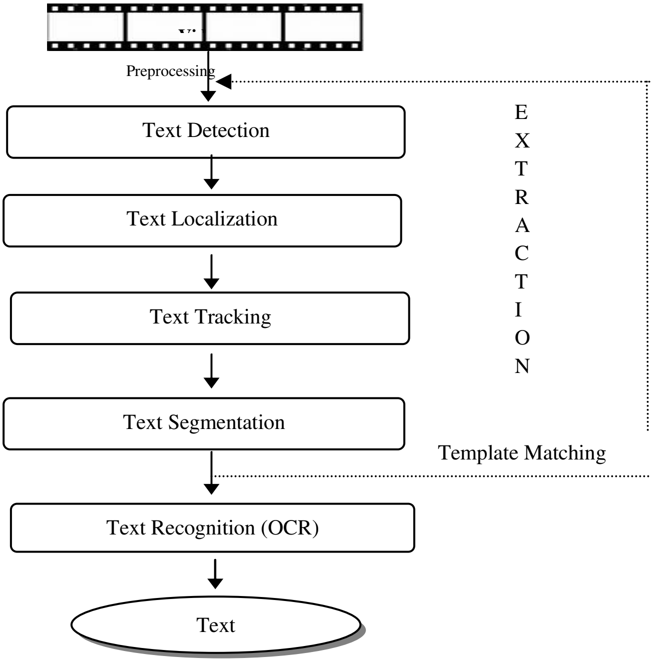


Fig. 2. System for character detection and recognition from video/image

Character Recognition:- 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 1. These include 26 capital letters, 26 small letters, 10 numeric digits and 18 special characters table 1. These include 26 capital letters, 26 small letters, 10 numeric digits and 18 special characters [26].

Table 1. Videos frame template classes

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	“	;	,	.	#	&	@	(
)	-	%	!	:	‘	$	?	+	/

$$\text{Recall}=\frac{\text{Correct Detected}}{(\text{Correct Detected}+\text{Missed Text Lines})} \tag{6}$$

Whereas precision is defined as:

$$\text{False alarm rate}=\frac{\text{Number of falsely detected text}}{\text{Number of detected text}} \tag{7}$$

$$\text{Precision}=\frac{\text{Correct Detected}}{(\text{Correct Detected}+\text{False Positives})} \tag{8}$$

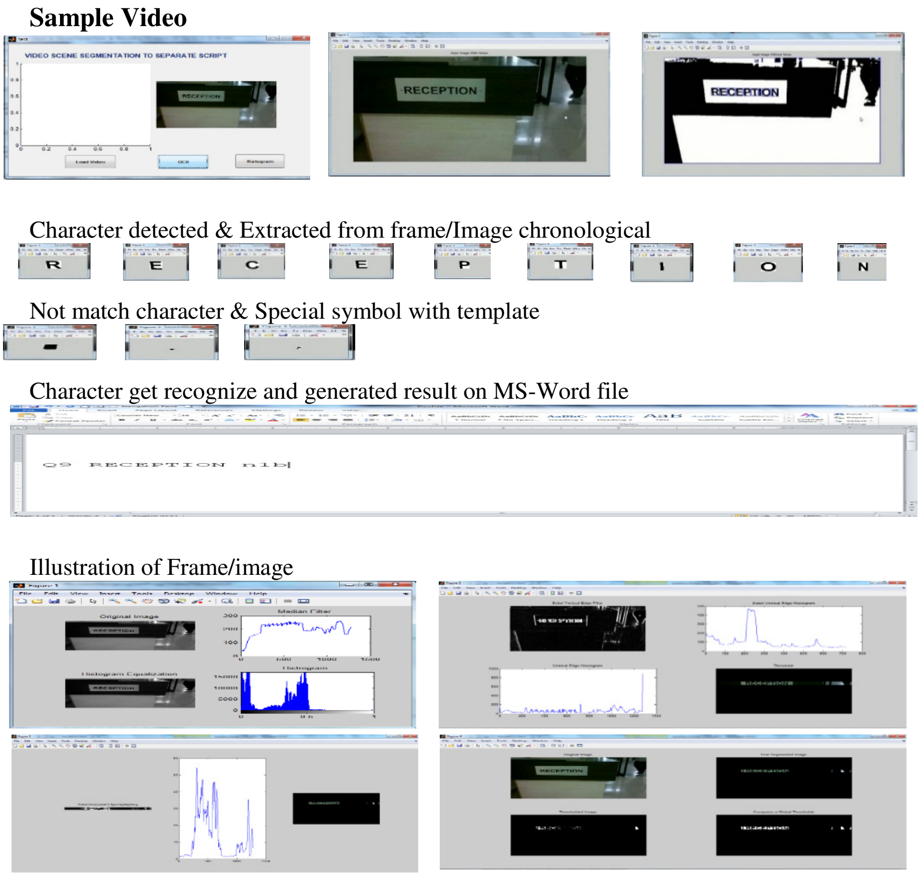
3 Experimental Implementation and Result Analysis

There are several performance evaluations to estimate the VTDAR 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 [27]. The performance metrics are as follows:

False Positives: False Positives (FP) / False alarms are those regions in the image which are actually not characters of a text, but have been detected by the algorithm as text.

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.



Sample Video

Character detected & Extracted from frame/Image chronological

Not match character & Special symbol with template

Character get recognize and generated result on MS-Word file

Illustration of Frame/image

Fig. 3. a) Median filter & Histogram b) Sobel vertical Edge filter & Histogram, Threshold c) Sobel Horizontal Edge d) Computes global threshold

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

Test Result and Analysis: The following table compares recognition result of improved template matching method and traditional template matching method, the test result is shown in below table 2&3.

4 Performance Evaluation VTDAR Algorithms

Every character set the text box as per the character ,digit, 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 90.96% and a precision of 92.15%.As seen from the table 2 & 3 using the improved template matching method, the average recognition rate and Recognition speeds of upper, lower letters, numeric and special characters have been enhanced.

Table 2. Experimental result for the proposed Algorithm

Images/Frames	1250
Text Lines	1530
Correct Detected	1410
False Positives	120
Recall (%)	90.96%
Precision (%)	92.15%

Table 3. Characters recognition test table

Test group	Recognition Result
Uppercase	92.05%
Lowercase	92.22%
Digits	93.05%
Special Character	91.30%

Video Databases (Video Text Detection and Recognition):-

We testing database for VTDAR, database sources that can be down loaded from the location of web sites. There are also several research institutes that are currently working on this problem. Testing for VTDAR we having creating our own database and also downloaded videos sample approximately 50 videos are used for testing sample .The below in the link. Data set, location, language [18]. Our experiments are conducted on 50 video sequences having a 320x240 frame size with a frame rate of 25fps. Some of these video clips are captured with E7 Nokia Mobile; others are downloaded from above listed standard database. Test data for VTDAR Own database by using E7 Nokia mobile https://sites.google.com/site/ bharatgaikawad2012/videos, http://documents.cfar.umd.edu/LAMP/, http://www.cfar.umd.edu/~doermann/UMDTextDetectionData.tar.gz

Comparison Study of Video text detection and recognition (VTDAR) with Tesseract and Transym OCR .On the basis of below table 4, we have compared the recognition rate of VTDAR with Tesseract and Transym OCR .Tesseract and Transym both tools are proprietary Optical Character Recognition, we tried to tested all set 1250 of videos frames or images with all 3 different techniques or tools. Empirical result of VTDAR precision rate is 92.15 %. Experiments show that the recognition rates of VTDAR are compared to branded tools as Tesseract 3.02, Transym OCR 3.3 which is approximately similarly.

Table 4. Percentage of recognition with different Techniques/Tools

Sr.No.	Techniques/Tools	No. of Test Frames/images	Text Lines	No. of Correct Detected	False Positives	% of Recognition
1	VTDAR	1250	1530	1410	120	92.15%
2	Transym OCR 3.3	1250	1530	1425	105	93.13%
3	Tesseract 3.02	1250	1530	1450	80	94%

Graphically illustration of Character recognition result as following in figure 4.

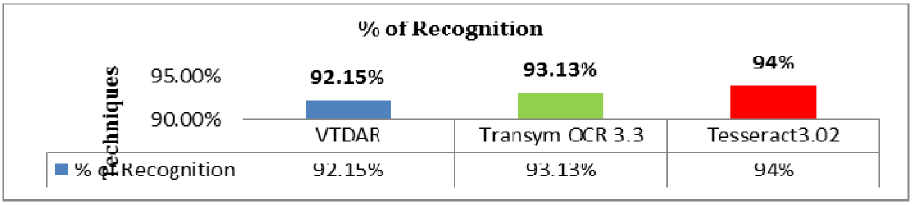


Fig. 4. Rate of percentage for character recognition from video frames or image

5 Conclusions

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 50 videos with 1250 video frames and 1530 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 7.85% characters could not get detected but some other character get recognized . The overall empirical performance of this system recognizing rate is 92.15%successfully. Empirically show that the recognition rates of VTDAR are compared to branded tools as Tesseract3.02, Transym OCR 3.3 which is approximately similarly.

References

1. Hua, X.-S., Wenyin, L., Zhang, H.-J.: Automatic Performance Evaluation for Video Text Detection. In: Sixth International Conference on Document Analysis and Recognition (ICDAR 2001), Seattle, Washington, U.S.A, September 10-13, pp. 545–550 (2001)
2. Junga, K., Kimb, K., Jain, A.K.: Text information extraction in images and video: a survey. Published by Elsevier Ltd. (2003)
3. Canny, J.: A Computational Approach to Edge Detection. IEEE Trans. Pattern Analysis and Machine Intelligence 8, 679–714 (1986)
4. Kim, H.K.: Efficcient automatic text location methodand content-based indexing and structuring of video database. J. Visual Commun. Image Representation 7(4), 336–344 (1996)
5. Zhong, Y., Jain, A.K.: Object localization using color, texture, and shape. Pattern Recognition 33, 671–684 (2000)
6. Antani, S., Kasturi, R., Jain, R.: A survey on the use of pattern recognition methods for abstraction, indexing, and retrieval of Images and video. Pattern Recognition, 945–965 (2002)

7. Jie, X., Hua, X.-S., Chen, X.-R., Wenyin, L., Zhang, H.: A Video Text Detection and Recognition System. In: IEEE International (2009)
8. Shivakumara, P., Huang, W., Tan, C.L.: Efficient Video Text Detection Using Edge Features. In: The Eighth IAPR Workshop on Document Analysis Systems (DAS 2008), Nara, Japan, pp. 307–314 (2008)
9. Lienhart, R., Stuber, F.: Automatic text recognition in digital videos. In: Praktische Informatik IV, University of Mannheim, 68131 Mannheim, Germany
10. Ye, Q., Gao, W., Wang, W., Zeng, W.: A Robust Text Detection Algorithm in Images and Video Frames. In: IEEE ICICS-PCM, pp. 802–806 (2003)
11. Aghajari, G., Shanbehzadeh, J., Sarrafzadeh, A.: A Text Localization Algorithm in Color Image via New Projection Profile. In: IMECS, Hong Kong (2010)
12. Ghorpade, J., Palvankar, R.: Extracting Text from Video. Signal & Image Processing, An International Journal (SIPIJ) 2(2) (2011)
13. Gaikwad, B., Manza, R.R.: Critical review on video scene segmentation and Recognition. International Journal of Computer Information Systems (IJCIS) 3(3) (2011)
14. Manza, R.R., Gaikwad, B.P.: A Video Edge Detection Using Adaptive Edge Detection Operator. CiiT International Journal of Digital Image Processing (2012), doi: DIP012012006, ISSN: 0974–9691 & Online: ISSN: 0974-9586
15. Manza, R.R., Gaikwad, B.P., Manza, G.R.: Use of Edge Detection Operators for Agriculture Video Scene Feature Extraction from Mango Fruits. Advances in Computational Research 4(1), 50–53 (2012)
16. Manza, R.R., Gaikwad, B.P., Manza, G.R.: Used of Various Edge Detection Operators for Feature Extraction in Video Scene. In: Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering, ICACEEE 2012 (2012) ISBN: 978-981-07-1847-3
17. Sumathi, C.P., Santhanam, T., Priya, N.: Techniques and challenges of automatic text extraction in complex images: a survey. Journal of Theoretical and Applied Information Technology 35(2) (2012)
18. Spitz, A.L.: Determination of the Script and Language content of Document Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(3) (1997)
19. Sharma, S.: Extraction of Text Regions in Natural Images. Masters Project Report (Spring 2007)
20. Mollah, A.F., Majumder, N.: Design of an Optical Character Recognition System for Camera based Handheld Devices. IJCSI 8(4(1)) (2011)
21. Su, Y.-M., Hsieh, C.-H.: A Novel Model-Based Segmentation Approach To Extract Caption Contents On Sports Videos. In: IEEE International Conference on Multimedia & Expo, pp. 1829–1832 (2006)
22. Leon, M., Vilaplana, V., Gasull, A., Marques, F.: Caption Text Extraction for Indexing Purposes Using a Hierarchical Region-Based Image Model. In: Proceedings of the 16th IEEE International Conference on Image Processing, pp. 1869–1872 (2009)
23. Zhong, Y., Zhang, H., Jain, A.K.: Automatic Caption Localization in Compressed Video. In: International Conference on Image Processing, vol. 2, pp. 96–100 (1999)
24. Liu, X., Wang, W.: Extracting Captions From Videos Using Temporal Feature. In: Proceedings of the International Conference on ACM Multimedia, pp. 843–846 (2010)

25. Lilo, B., Tang, X., Liu, J., Zhang, H.: Video Caption Detection and Extraction Using Temporal Information. In: International Conference on Image Processing, vol. 1, pp. I297–I300 (2003)
26. Gaikwad, B.P., Manza, R.R., Manza, G.R.: Video scene segmentation to separate script. In: Advance Computing Conference (IACC). IEEE xplore IEEE (2013) 978-1-4673-4527-9
27. Gaikwad, B.P., Manza, R.R., Manza, G.R.: Automatic Video Scene Segmentation to Separate Script for OCR. International Journal in Computer Application (IJCA) (2014) ISBN: 973-93-80880-06-7