A Hybrid Approach to Extract Scene Text from Videos

A.Thilagavathy, K.Aarthi, A. Chilambuchelvan
Department of Computer Science and Engineering
R.M.K Engineering College
Chennai, India
atv.cse@rmkec.ac.in
aarthi.kme@gmail.com

***Abstract*— With escalating claim for information indexing and retrieval a lot of attempts have been done on hauling the text from images and videos. Hauling the scene text from image and video is challenging due to complex background, changeable font size, dissimilar style, unknown layout, poor resolution and blurring, position, viewing angle and so on. The primary objective of the proposed system is to detect and haul out the text from digital videos. Through this paper we propose a hybrid approach to haul out the text from videos by integration of the two popular text extraction methods: region and connected component (CC) based method. Primarily fragment the videos into frames and acquire the key frames. Text region indicator (TRI) is being developed to figure out the text prevailing confidence and candidate region by performing binarization. Artificial Neural network (ANN) is used as the classifier to filter out the text and non-text components where Optical Character Recognition (OCR) is used for verification. Text is grouped by constructing the minimum spanning tree using the bounding box (BB) distance.**

***Keywords*— Caption text, Preprocessing, Scene text, Text extraction, Text grouping, and Video frame extraction**

I. Introduction

We are living in the world of brainchild of latest video capturing devices like mobile phones, digital cameras with the high resolutions and updated technology. So the progress of these hand held video/image devices (HID) lead a head start to the capturing of more videos and images with and/ or without text. Text extraction from video has become a recent research area these days as it has many applications like automatic license plate reading, sign detection and translation, mobile text recognition, content based web image search, navigation, text and logo detection in CCTV video feeds. It is also very useful for visually impaired person in daily life to give them access to text and, coupled with the text to speech algorithm, make them read book covers, banknotes, labels on doors, medicine labels and so on.

In content based video and image retrieval we make use of only the low level features from image or video such as color, shape or texture. Even though it is easy to extract these low level features from the images and /or videos, it does not give the full semantic information of the video. So we make use of the high level features like text extracted from the video, which has many benefits:

- Very useful for describing the contents of video sequence.
- Effortlessly extracted and compared to other semantic contents.
- Extracted text is easily synchronized with the image data when the event occurs.
- Labor-intensive logging may not be feasible for large collection of archived video.

There are two types of text that appears in the video: scene text and caption text. *Scene text* is the text that in nature occurs in the area of capturing of video like text on t-shirt, container, CD cover, sign board, text on vehicle. It is also called graphics text. *Caption text or artificial text* on the other hand is the text that is artificially overlaid on the video/image such as the scores in sports videos, subtitles in news video, date and time in the video. It is also called superimposed text.

Extracting scene text from the video is difficult when compared to the caption text due to the following reasons,

- *Viewing angle*: If the angle at which the video was taken is not accurate then it poses difficulties while extracting text from video as it would lead to distortion.
- *Lower resolution*: As there are devices with larger resolution range but OCR can't recognize the video with resolution less than 50 dpi.
- *Unknown layout***:** Caption text is superimposed on the video so it is structured whereas scene text is not as they are of different orientation and alignment.
- *Varying font size, style, and font:* The text may be of varying sizes and of different styles making it difficult to extract.
- *Complex background:* Sometimes the text and background are of same color thus not enabling to detect the text efficiently.
- *Non-uniform lighting***:** Irregular lighting, shadowing, reflection onto object, inter-reflection among object

978-1-4673-0210-4/12/$31.00 ©2012 IEEE

may make the color vary significantly and decrease the analysis performance.

- *Non-planar object*: Text in cans and bottles suffer from distortion.
- *Object distance*: Distance between the text and HID can vary and it may lead to wide variation on size of the same text.

The existing methods to detect and extract the text from video are classified into thresholding based and grouping based methods.

A. Thresholding based method

It would define a global and local threshold for the whole image and for a selected portion of the image respectively. The thresholding based method is further classified into,

*1) Histogram based thresholding :*Mainly used for the monochrome image. It would count for the number of each pixel value in the histogram. Threshold is considering being the value between the two peaks. They are not suitable for the image with complex background.

2)*Adaptive binarization techniques:* The threshold is defined for the parts of the video. Most commonly used is the Niblack's method which considers the mean and standard deviations over the radius of r. It is mainly used for grayscale image.

B. Grouping based method

It will group the text pixel according to some criteria in order to extract the extract the text. It is further classified into following types,

1)*Region based method*: It would detect and extract the text using the texture analysis. It is further classified into two groups: top-down and bottom-up. Top-down will consider the entire image and then will move towards the smaller parts by considering the grayscale value. Bottom-up approach on the other hand will start with the smaller parts and then will merge into a single image. The connected component (CC) based method and edge based method are of bottom approach which makes use of the edges. It is expensive and slow too.

2) *Learning based approach:* It mainly makes use of the neural networks. Some of classifiers it makes use of are Multilayer perceptron (MLP), single layer perceptrons (SLP) and so on.

3) *Clustering based approach:* It would group the text into clusters based on the color similarity. Some of the commonly used methods are K-means clustering, Gaussian mixture model (GMM), density based and so on.

To overcome the difficulties to detect and extract the text from video we proposed a system that makes use of the hybrid approach. First we need to identify the text region indicator which provides the information whether the text is present or not i.e. text prevailing confidence and the scale of the candidate region in the selected key frames. Then we need to construct the conditional random field (CRF) to realize the text and non-text components by making use of some properties. At last we need to group the letters into words and words into lines by using the learning based minimum spanning tree.

In the following section we discuss about the related works in section II, system overview in section III and about the index terms in the section IV, V, VI, VII respectively, conclusion in section VIII and future contribution in section IX.

II. RELATED WORK

The overview of the existing text localization methods was provided in the Section I. In [6] Jung *et al developed the text information extraction system (TIE)* with four stages: text detection (finds the text region in the frame), text localization (groups the text region and generate bounding boxes), text extraction and enhancement (extract the text using some classifier and enhance it) and recognition (verify the extracted text with OCR).An approach proposed by Liu *et al* [1] makes use of the two text extracting methods: region based and connected component (CC) based method. It uses region based method for segmentation and CC for filtering the text and non-text components. It used feature descriptor: histogram of oriented gradients (HOG) and a boosted cascade classifier: WaldBoost for constructing the text region indicator.

In [2] Hu *et al* proposed a corner based approach to detect and extract the caption text from videos. It assumed that the text has a dense corner points which will help us to figure out the presence of text in the videos. In [3] Weinman *et al* developed the unified processing which consists of information namely: appearance, language, similarity to other characters, and a lexicon. By integrating the character identity and similarity information in a unified model the recognition is improved and at last the lexicon is used through the application of the sparse method. In [4] Dumont *et al* projected an approach where the video is segmented into a one second video segments and they are clustered using the agglomerative hierarchical clustering approach. The significant frames are selected for further processing.

In [5] Tsai *et al*, they proposed the method to detect the sign from videos. It uses connected component analysis to detect the candidate region. It developed a single detector for all the color rather than a separate detector for each color. The radial basis function network is used as the classifier. A rectification method was proposed to rectify a skewed road sign to its correct shape so the deformation is surmounted. Then, all its embedded texts can be well recognized. In [7] Nicolas *et al* implemented the Conditional Random Field (CRF) in the document analysis. It takes into account both the local and contextual feature. These features are extracted and feed as input to the Multilayer Perceptron (MLP).In [8] *et al* Epshtein suggest that the one popular feature that distinguishes the text from the other component is stroke. It uses a Stroke Width Transform (SWT) to calculate per pixel the width of the stroke. It will combine the adjacent pixel only if they have the same stroke and finally the text lines are grouped based upon some criteria like stroke width, height ratio

In [9] Tan *et al* used the sharp edges, straight appearances of edges and proximity of the edge distribution in the text block. He proposed Fourier-statistical features (FSF) in RGB space to detect the text from the video frames. It would divide the frame into block and so it is easy to detect the text in block level rather in the frame level. It makes use of the arithmetic filter (AF) and mean filter (MF) to get rid of the noise from the image. Canny edge and Sobel operator are used for edge detection. In [10] Chasanis *et al* proposed a novel approach to obtain the key frames from the video by using the clustering techniques especially K-means algorithm. It used Needleman–Wunsch global sequence alignment algorithm for checking the frame similarity by building the substitution matrix using color similarity. In [11] Lyu *et al* developed an approach to extract text from videos using edges and thresholding. Sobel operator is used to detect the candidate regions.

In [12] Chen *et al* uses Laplacian of Gaussian (LOG) edge detector for predicting the edge sets and apply the affine rectification to overcome the deformation due to the improper camera angle. Gabor transform is used to obtain the local features. In [13] Chen and Yuille used Adaboost classifier where the weak classifiers are applied to train for a strong classifier in order to construct the fast text detector. This region identified by the classifier is given as input to the binarization and followed by CC analysis. In [14] Kim *et al* presented an approach where they used SVM as a classifier and then perform the continuously adaptive mean-shift (CAMSHIFT) to identify the text regions.

In [15] Lienhart *et al* used Complex-valued multilayer feedforward network trained to detect text at an unchanging scale and position. In [16] Li et al make use of the neural network as classifier and the extracted text is compared with the successive frames. It will identify the presence of the text in 16 X 16 windows only and SSD (Sum of Squared Difference) for frame similarity. In [17] Zhong *et al* extract text from compressed video using the discrete cosine transform (DCT). It applies horizontal thresholding to obtain the noise. In [18] Liu *et al* used edge detection algorithm to obtain the text color pixels. Connected component analysis is done to obtain the text confidence. In [19] Zhu *et al* employs Non-linear Niblacks method to perform the gray scale conversion and then fed into the classifier which is trained by Adaboost algorithm for filtering the text and non-text regions.

III. SYSTEM OVERVIEW

The main objective of the proposed system is to haul out the text from the videos. The proposed system consists of four stages namely: video frame extraction, preprocessing, text extraction and text grouping. The figure 1 shows the architecture diagram for the proposed system. Firstly we will divide the uploaded videos into frames based on the shots. By performing the frame similarity we will get rid of the redundant frames and this set in motion the selection of the key frames which contains the scene text.

Next we need to perform the pre-processing stage in order to identify the text prevailing confidence and its scale in the key frames. This would lead to the region where the text is present i.e. candidate region. In order to identify the presence of the text in the key frame we need to apply the adaptive thresholding (binarization) and perform the morphological operations to remove the noise. After the text region is detected we perform the connected component analysis where it performs both horizontal and vertical projection in the key frame to detect the text.

In Connected Component Analysis (CCA) the CRF model is used to classify the candidate region into two classes: {text, non-text}. The Artificial neural network is trained to be a classifier to filter out the text and non-text components. The extracted text is passed to the OCR (Optical character recognition) for character confirmation. At last the texts are grouped into words and in turn into lines by using the horizontal and vertical bounding box distances by building minimum spanning tree.

Algorithm for hybrid approach

```
Input: Video which contains scene text
Output: Text extracted from video
Begin
1. Split video into frames, get rid of temporal redundancy
   and get hold of key frames.
2. Carry out image segmentation by means of binarization
   to acquire binarized image,  $Z= \{z_{1...} z_n\}$.
3. Obtain the text prevailing confidence and the candidate
   region by projecting the binarized image in two
   dimensions.
     For (x=0; x<=n; x++) {
         For (y=0; y<=n; y++) {
             Generate the bounding box (BB). }}
4. Train ANN to haul out the text from the candidate
   region.
5. Construct MST for text grouping.
   For (z=1; z<=n; z++) {
      Match up to the BB distance of the neighboring ones
      and assemblage the text.  }
End.
```

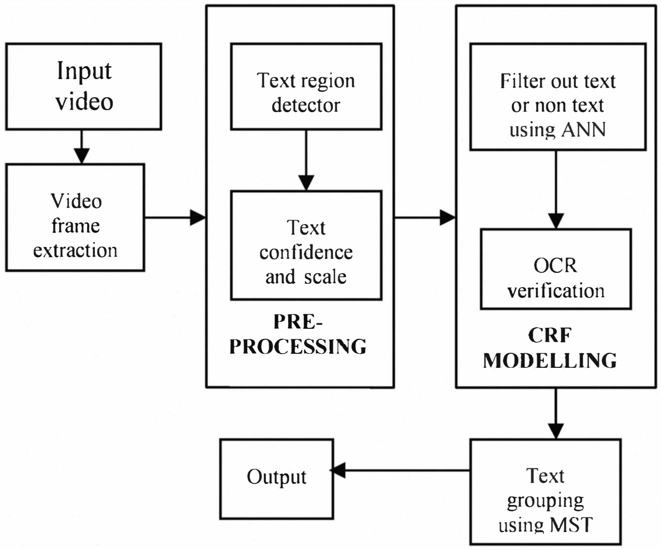

Figure 1 Block diagram for text extraction from videos

IV. VIDEO FRAME EXTRACTION

After uploading the videos containing the text firstly we need to split it into frames by reducing the rate of the video to 1 or 0.1 second. Experimental results showed that when we consider the frame rate to be 0.1 second it will lead to the redundant frames. In order to overcome this temporal redundancy we make use of the pixel by pixel comparison where every pixel of a frame is compared with respect to the other frames. When we find the inter frame space difference between frames to be high it indicate that those frames are similar, so we store only one frame and discard all the remaining frames. Thus by using the pixel by pixel comparison we would be able to eliminate the redundant frames and will result in the distinct and unique frames (non-redundant video frames set). Secondly, there is the need to choose the key frame from these non-redundant frames set. The key frames are those frames which contain the text.

V. PRE-PROCESSING

In this stage we will design the TRI to identify the candidate region using the hybrid approach where the transition map generation which will distinguishes the text from background. It will provide the text prevailing confidence and length of the candidate region which will lead to image binarization. It is assumed that the text will have a certain set of properties that distinguishes itself from the background. Some of the properties are alignment (vertical / horizontal), inter-character space; motion (static, 2D or 3D).

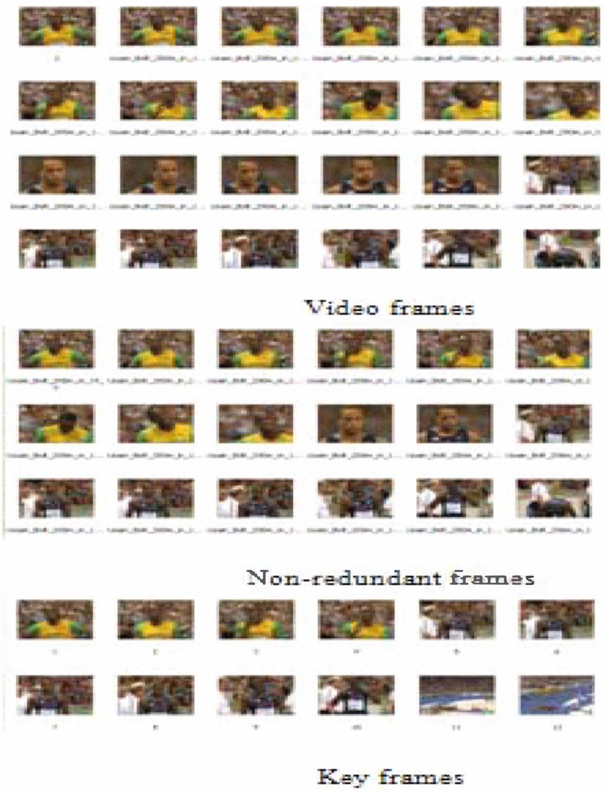

Figure 2 A video splitting and key frame selection

To obtain binarized image we will convert the colored frame into gray scale and apply the adaptive thresholding. The threshold value is selected conditional upon the minimum size of text region. The Niblack binarization technique will convert the gray scale image into binarized image. In Niblack's binarization algorithm the threshold value is selected according to the mean and standard deviation by sliding the small window over the key frame. We project the binarized image in two dimensions by examining the successive pixels to perform CC. Let X= $\{X_{ij}\}$ and T be the gray scale image and threshold value respectively, b_i and b_j be the binary value corresponding to the black and white of the binarized image,

$$X_{ij=}\begin{cases} b_{i,} \text{ if } X_{ij} < T \\ b_j, \text{ if } X_{ij} > T \end{cases} \qquad (1)$$

The hybrid approach is applied on the gray scaled image to obtain the edges using the canny edge detector. Later the noises are removed by using the morphological operation and the text region is detected by performing the horizontal and vertical projection which leads to the text prevailing confidence and the scale of the candidate region. The candidate regions are selected based on the aspect ratio (height and width ratio) so it will help us to discard the non-text regions and further will reduce the false alarm rate also.

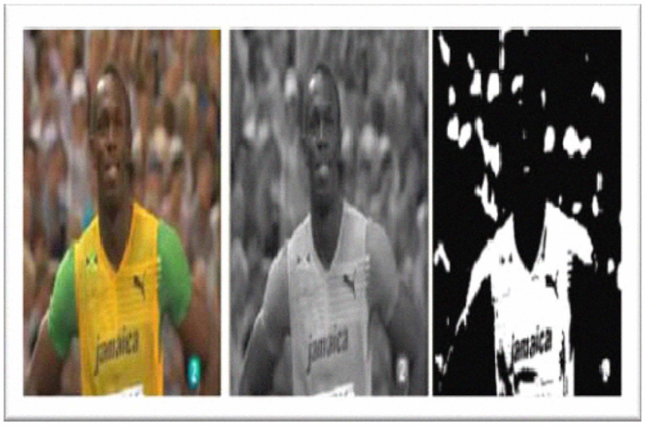

Figure 3 A sample color video frame converted into binarized frame

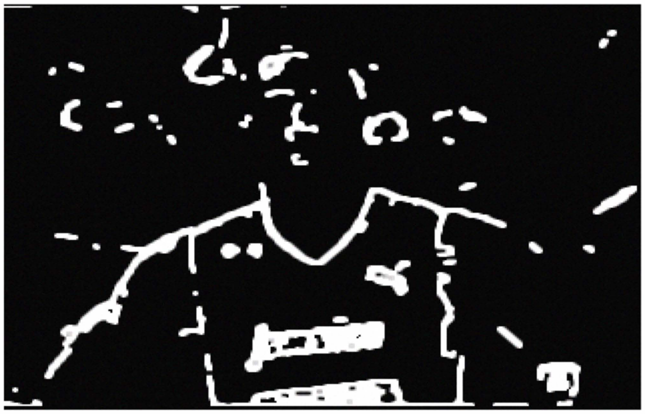

Figure 4 A key frames after pre-processing

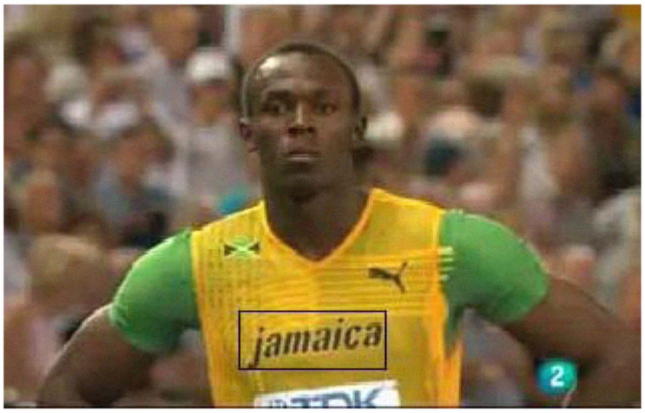

Figure 5 A text localization examples

VI. TEXT EXTRACTION

In CCA, we use CRF model to classify the candidate region into text and non-text. The CRF is the graphical model with depends upon the Markovian property. The binarized image Z= $\{z_1 \ldots z_n\}$ is projected in two dimensional to construct the component locality graph by assuming that the neighboring text will have the same width and height. It calculates the Euclidean distance between the centroid of the two components along with the height and width of the bounding box. If it satisfies the following rule then it is added to the graph,

$$\text{Dist}\ (Z_i, Z_j) =< 2 * \min\ [\max\ (W_{i,} H_i), \max\ (W_j, H_j)] \qquad (2)$$

The CRF takes into account both unary (height and width, aspect ratio, confidence) and binary (shape difference, spatial distance) components features. The Artificial Neural Network (ANN): Multi-layer Perceptron (MLP) is being engaged as a classifier to categorize the text and non-text components. The MLP is trained using back propagation by the means of gradient descent technique. It is trained by supervised learning where the hidden layer of the network contains the matrix value of the alphabets (both upper and lowercase) in the data array which is used for comparing the text from video frame. Thus it would classify the input and obtains the text. The extracted text is then passed on to the OCR (Optical character recognition) for the character confirmation.

VII. TEXT GROUPING

Once the texts are extracted we need to group it into words and in turn to lines by construction the Minimum Spanning Tree (MST) using Kruskal algorithm. The tree is built on the basis of the bounding box distance between the texts both horizontally and vertically for the word and line partition respectively. It uses the spatial distance which is the distance between the BB. The edge cuts are used as the partition in the tree construction which will lead us to the text localization.

VIII. CONCULSION

In this paper we proposed a hybrid approach to extract the extract the text from the videos. We split the videos into frames and identify the text containing video frames. Then project the binarized image to obtain the text prevailing confidence and the length of the candidate region. ANN is trained to classify the text and non-text component in the candidate region which is followed by the OCR confirmation. Text is grouped by constructing MST on the basis of the bounding box distance. Thus our proposed system will extract the text from videos effectively.

IX. FUTURE CONTRIBUTION

In this paper we focused only on the English character extraction from the videos. In future we will make an effort to widen it to as many languages as possible. Thus multilingual text extraction method will be useful to us in various applications like the road sign detection and translation and so on. In future we will try to extract the vertical text from videos. We will enhance our proposed system to extract the text from the deformed image. We will also try to improvise the proposed system in video retrieving using the text present in the videos.

REFERENCES

[1] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, Senior Member, IEEE, “A Hybrid Approach to Detect and Localize Texts in Natural Scene Images”, IEEE transactions on image processing, vol. 20, no. 3, March 2011.

[2] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Member, IEEE, Yuxiao Hu, Member, IEEE, Yuncai Liu, Member, IEEE, and Thomas S. Huang, Life Fellow, IEEE “Text from Corners: A Novel Approach to Detect Text and Caption in Videos”, IEEE transactions on image processing, vol. 20, no. 3, march 2011.

[3] J. Weinman, E. Learned-Miller, and A. Hanson, “Scene text recognition using similarity and a lexicon with sparse belief propagation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, 2009.

[4] Emilie Dumont and Bernard Mérialdo “Split-Screen Dynamically Accelerated Video Summaries”, ACM *TVS’07,* September 29, 2007, Augsburg, Bavaria, Germany.

[5] L. W. Tsai J. W. Hsieh C. H. Chuang Y. J. Tseng K.-C. Fan, C. C.Lee1,”Road Sign Detection Using Eigen Colour”.

[6] K. Jung, K. I. Kim, and A. K. Jain, “Text information extraction in images and video: A survey,” *Pattern Recogn.*, vol. 37, no. 5, pp. 977–997, 2004.

[7] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, “Document image segmentation using a 2-D conditional random field model,” in *Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR’07)*, Curitiba, Brazil, 2007, pp. 407–411.

[8] Boris Epshtein Eyal Ofek Yonatan Wexler, Microsoft Corporation, “Detecting Text in Natural Scenes with Stroke Width Transform”.

[9] Palaiahnakote Shivakumara, Trung Quy Phan, and Chew Lim Tan, Senior Member, IEEE,“New Fourier-Statistical Features in RGB Space for Video Text Detection”, IEEE transactions on circuits and systems for video technology, vol. 20, no. 11, November 2010

[10]] Vasileios T. Chasanis, Aristidis C. Likas, and Nikolaos P. Galatsanos, “Scene Detection in Videos Using Shot Clustering and Sequence Alignment”, IEEE transactions on multimedia, vol. 11, no. 1, January 2009,89.

[11] Michael R. Lyu, Fellow, IEEE, Jiqiang Song, Member, IEEE, and Min Cai, “A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction”, IEEE transactions on circuits and systems for video technology, vol. 15, no. 2, February 2005, 243

[12] X. L .Chen, J. Zhang, and A. Waibel, “Automatic Detection and Recognition of Signs from Natural Scenes”, IEEE transactions on image processing vol13, no.1, pp.87-99, Jan 2004

[13] X.R.Chen and A.L.Yuille, “Detecting and Reading Text in Natural Scenes”, in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR’04), Washington, DC, pp. 366–373, 2004.

[14] K. I. Kim, K. Jung, and J. H. Kim, “Texture-based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm,” IEEE transaction on. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1631–1639, 2003.

[15] Rainer Lienhart, Member, IEEE, and Axel Wernicke, “Localizing and Segmenting Text in Images and Videos”, IEEE transactions on circuits and systems for video technology, vol. 12, no. 4, April 2002.

[16] H. P. Li, D. Doermann, and O.Kia, “Automatic Text Detection and Tracking in Digital Video,” IEEE transaction on Image Processing, vol. 9, pp. 147–156, Jan. 2000.

[17] Yu Zhong, Hongjiang Zhang, and Anil K .Jain, Fellow,“Automatic Caption Localization in Compressed Video”, IEEE transactions on pattern analysis and machine intelligence, vol. 22, no. 4, April 2000.

[18] Y. X. Liu, S. Goto, and T. Ikenaga, “A Contour-based Robust Algorithm for Text Detection in Color Images,” IEICE transaction. Inf. Syst., vol. E89-D, no. 3, pp. 1221–1230, 2006.

[19] K. H. Zhu, F. H. Qi, R. J. Jiang, L. Xu, M. Kimachi, Y. Wu, and T. Aizawa, “Using Adaboost to Detect and Segment Characters From Natural Scenes,” in Proc. 1st Conf. Caramera Based Document Analysis and Recognition (CBDAR’05), Seoul, South Korea, pp. 52–59., 2005.