International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Multi-Oriented Text Localization and Classification from Natural and Video Scenes

Namitha Nandakumar
M.Tech Computer Science and Engineering,
Sree Narayana Gurukulam College of Engineering, Kadayirippu
Ernakulam, Kerala

***Abstract*— Localization of text from video as well as natural scenes is not a straightforward chore. If it comes to the case of text in different orientation, it certainly becomes a tedious task. Multi-oriented text can be referred to as the scene texts that make text localization from the scene more exigent because of the non-uniformity of the text features. Being much challenging task, conventional text localization methodologies will not perform well for multi-oriented text detection. The scheme projected in this work can be made use for localizing the characters aligned in any direction and overlaid on any complex background. The work also recommends a technique to catalog text extracted from the natural scene images. The very common Canny edge map, that sketch even minuscule details in a scene is used for edge detection in both natural and video scenes. To traverse the text aligned in multiple orientation, I make use of boundary growing technique. This technique works on the basis of nearest neighbor theory. Text classification framework make use of boundary clustering, stroke segmentation and string fragment classification. Multi oriented text localization is experimented in video and natural scenes whereas text classification framework is evaluated on natural scene images.**

Keywords*— Canny edge detection, boundary growing method, boundary clustering, stroke segmentation, string fragment classification (key words)

I. INTRODUCTION

The field of information retrieval has emerged to be the key focus of most of the researchers who focus on image processing and content retrieval today. Localization and detection of text from video frames as well as natural images is an emerging area in this significant field. Text information from images serve as vital information in different applications. Text localization is not an easy job and it is regarded as a major problem in Content Based Image Retrieval. Optical Character Recognition technique is mainly being deployed for text detection and extraction. Design of OCR techniques allows us to transform text images into readable text codes. In the case of image frame with complex background, performance of OCR techniques seems to be very poor. As CBIR requires human intervention, performance of CBIR also mortifies with large dataset.

A video can be considered as a sequence of image frames running continuously. Hence, the methods proposed to retrieve text from video scenes can be used for natural scene images also. There are several methods in the literature for text detection and extraction in video and image frames. These processes can be generally classified into three types, to be exact, texture-based, connected component-based and edge and gradient based methods. Nearly all of the methods give attention to detection of text in only one orientation, mostly horizontal direction.

Scene images containing text information are separated into two categories based on the intricacy of the background. Images with characters and strings set in high resolution on a simple background falls into first category. It generally includes the close-up shots of objects, like book covers and wrappers. Image scene that attaches text into more compound backgrounds comes under second category. Text characters seem to be in regular structure in the print pattern, in these two classifications.

Text localization is rather a common topic in analysis of document on camera based images with a explosion. Pan et al. [1, 2] put forwards a hybrid approach to localize text in scene images based upon histogram of leaning gradients (HOG) and conditional random field (CRF). To group text components, geometrical property of connected component analysis is used. As the approach focus on scene images, performance of the same on video text degrades because of low resolution and complexity of background of video. Epshtein et al. [11] proposed a method that makes use of stroke width of the text component for detection of text. Video scene text cannot be extracted with much accuracy by using stroke width of character. Current works cited in [3-11] on text extraction from scene images point out that these methods take for granted big fonts like caption text in videos and characters with high contrast for detection of text. content retrieval from video based images is not supposed to be carried out with these methods, because of the features of video like dissimilarity in contrast, colour bleeding, distinct fonts and sizes, multiple orientations, complexity of background, outlook deformation, movement of text and background movements. [3-6].

Roy et al. [12] mentions a scheme for multi-oriented text localization from camera images. This scheme performs well if scene contains text with clear character shape. Thus methods for extraction of text from camera images and natural scene photographs will not be accurate for text detection from video images. Jain and Yu [8 9] suggest a conventional text detection algorithm derived from connected component analysis. Connected components are selected based on color, and if the components satisfy certain geometrical features it is considered as text candidate. This scheme fails if text with multiple colour characters is included in a text line. Li et al. [13] make use of

concept of texture features like mean, second and third order central moments in the wavelet domain.

Text information appears like text string in natural scene images and the characters of the string are almost of same size, aligned arrangement and constant color. In order to extract the text components, layout analysis can be done based on the mentioned features. [14], [15–16]. Other features like boundary-based structural analysis, stroke width consistency, and character alignment also plays a key role in text localization [17-18]. Edge distribution, closed component boundary, gradient variation, and edge based filter response acquired from boundary maps are some text features to detect and validate text regions.

Yi and Tian[19] put forward a new framework to localize text segment from scene images with intricate backgrounds and compound text appearances. The scheme proposes three main concepts: boundary clustering, stroke segmentation, and string fragment classification. In BC, a new bigram-color-uniformity-based method s proposed in order to model text and the attachment surface. Edge pixels are clustered based on color pairs and their spatial positions in the boundary layers. Stroke segmentation is carried out at every boundary layer by assigning color in order to extract character candidates. Algorithms are proposed to merge the structural analysis of text stroke along with the color assignment and to filter out false positives. Gabor-based text features are the basis of string fragment classification.

Here, a novel method is implemented that can be used for extracting multi-oriented text from both video frames as well as natural scene images[30]. Canny edge mapping algorithm is implemented to cut out even the minute detailing of the video as well as the natural scene image. Then, boundary growing method is applied to traverse the text candidates in different orientation. This will be perhaps the easiest method to traverse multi-oriented text from video scenes. Classification of the detected text is also done on the basis of neural network concept, mainly by deriving concepts of Yi and Tian[19]. Text region is calculated by making use of features at pixel, character and string level. Boundary clustering is used at the pixel level, stroke segmentation at character level and string fragment classification at string level. The images are trained using ANN to envisage whether the scene contains any text region or not.

Rest of the paper is arranged as follows. Section II illustrates the proposed method in detail. Section III portrays the technique of multi-oriented text extraction from video and natural scenes. Section IV describes classification of text and non text image. Conclusion and future work is mentioned in section V.

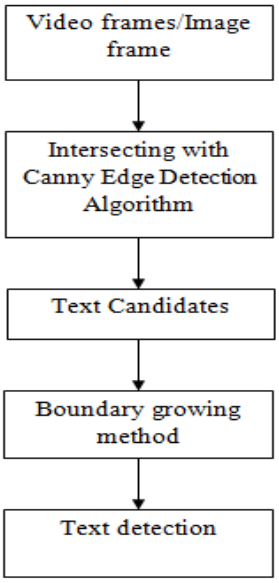
II. PROPOSED METHOD

A. *Detection of Text from Video And Scene Images*

Here by this work, a compact method for text detection is detected that work well for video scenes and natural scene images. The following block diagram gives an idea of the first phase of the proposed method.

B. *Classification Of Text And Non Text Images*

The block diagram depicted below shows the major steps in the classification of text and non text image patches. Major steps included in the process of classifying the text and non text images are boundary clustering, then stroke width analysis and move on to classification of the string fragment detected from the above mentioned processes. These three processes collectively form the second phase of the work.



a. Fig 1: Block diagram for detection of text from video and natural scenes

b.

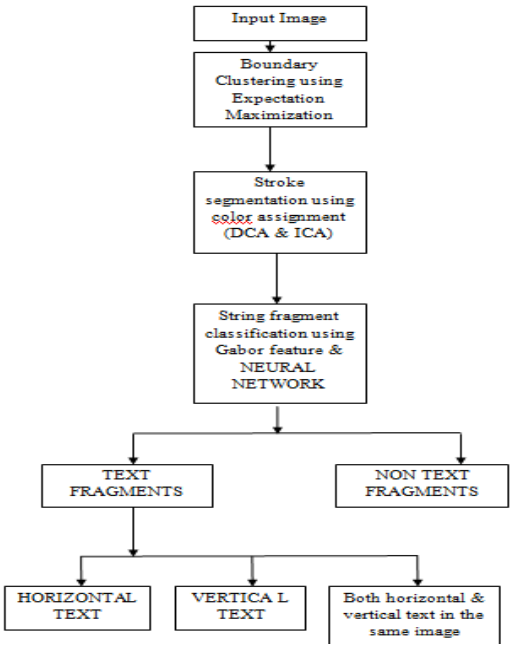


Fig 2: Block diagram for classification of text images and non text images

III. MULTI ORIENTED TEXT LOCALIZATION IN VIDEOS AND NATURAL SCENES

Localization and extraction of text in videos and images is carried out by applying the renowned edge detection algorithm, Canny Edge Detection Algorithm and the boundary of the text is traced by Boundary Growing Method. Both the methods work well for images and videos.

A. Canny Edge Detection Algorithm

Boundaries of an object, whether it's an image or any physical object, are characterized by edges. In the case of image processing, edge detection of an image is a crucial step. Edges in images will be characterized as areas that show high intensity contrasts. There will be a clear-cut difference in intensity per pixel as we move on to the edges of an image, when we analyze an image pixel by pixel. Edge detection algorithm that we use should preserve the structural features of an image while filtering out the useless data. That is the reason why, Canny edge detection algorithm is used in this work to trace the edges of an image.

Canny edge mapping is an optimal algorithm. It is crucial that edges in images should never be missed out and that non-edges should not be responded. The edge points should be well contained. The distance between the actual edge and the edge pixel detected by the algorithm should be always minimum.

Based on these conditions, the canny edge detection method first analyzes the image to remove noise. Next step is to find image gradient in order to show up regions with high spatial derivatives. It then trail down these regions and curbs all pixels that does not reach the maximum. Then the gradient array is condensed by hysteresis, which is used to track the remaining unsuppressed pixels. Two thresholds are used in hysteresis. If the magnitude undergoes the first threshold, set it to zero. That is, it is made as a non-edge. If the magnitude falls above the higher threshold, it is considered to be an edge. Value is set to zero, if magnitude lies between two thresholds.

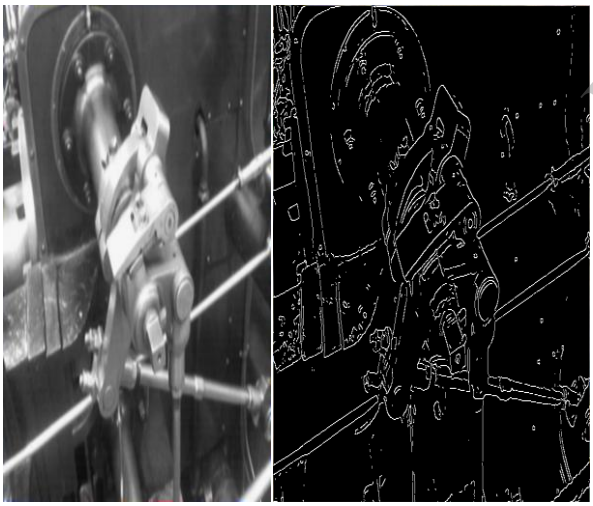


Fig. 3: Example of Canny edge mapping

B. Boundary Growing Method for Traversing Multi-Oriented Text

The key crisis of multi-oriented text detection is to set a bounding box along the text line by traversing the detected text pixels by means of canny edge detector. Due to the complex background, text traversing becomes exigent in case of videos. Despite of usual projection methods, an innovative idea is introduced by[] known to be Boundary Growing Method (BGM), which is based on nearest neighbor theory. The assumption that text lines enclosed in images appear with regular spacing between characters and words along a single direction all the time serve as the basis of the mentioned methodology. A text candidate image is scanned from top left pixel to bottom right pixel. If a component strikes while scanning, a bounding box is fixed for the component. The boundary will grow till it reach an adjacent bounding box is reached. The bounding box that seems to be adjacent to the earlier one will be the boundary of the next text pixel. The process is continued throughout the entire text line. The spacing between individual characters, words and text lines marks the end of text line. Coverage of the background information results in generation of false positives. Geometrical properties like aspect ratio, edge ratio etc are usually used for elimination of false positives. Its not possible to ensure that false positives can be eliminated in all means. The figure shown below depicts the process of growing bounding boxes along the text direction.

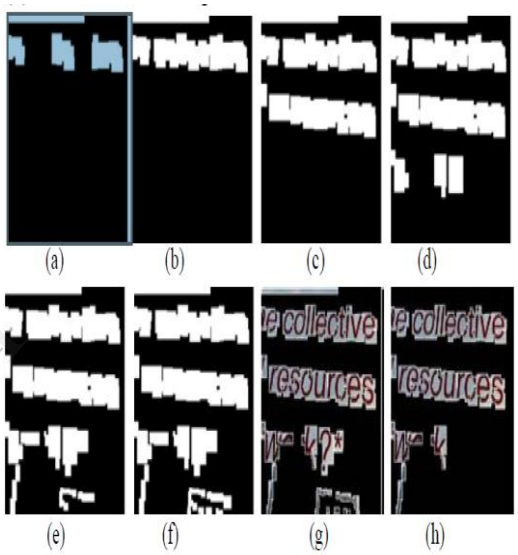


Fig. 4: Boundary Growing Method

The same methodology is being implemented for video as well as natural scene images for detection of text from the scenes. The only difference is that in the case of video, we have to repeat the whole method for each of the frame being computed. While when an image is loaded, there is only one frame and the result is calculated from that single frame.

IV. CLASSIFICATION OF TEXT IMAGE AND NON TEXT IMAGE

In order to classify the images, we first have to identify and extract the text lines from the input image. For the same, here we use Canny edge detection and Sobel product. Canny detection will provide an initial map of the boundary of the object. Sobel operations on the text line will extract high resolution pixels. Canny edge detection along with sobel operation yields a correct mapping of the text candidates in the loaded image. As soon as the text boundaries are marked the image will then undergoes several processes as follows, in an intention to make the classification of text and non text images.

The core of the section includes:
1) A color pair clustering algorithm which is based on GMM and EM algorithm
2) Structural analysis of stroke boundary.
3) String fragment classification.

A. Boundary Clustering

Text structure analysis and geometrical modeling largely depend on the boundary of the text line. We can derive the object boundary in a natural image by making use of the color difference of two identical regions, which are the object and its adjacent backgrounds. To analyze the text character boundary, employ the concept of color uniformity and spatial positions. A clustering algorithm is implemented to separate text characters from the edge line of background.

a. Color Uniformity and Spatial Position

There are several clustering algorithms based on color uniformity are in use now for text segmentation. Most of these algorithms does not take care of the difference in color of the neighboring pixel that lies around the object. Te color difference can be taken as an accurate measure for analyzing the texture of the object as it is robust to changes in lighting.

Text information seems generally to be attached to a plane carrier. This surface exhibits uniform color for pixels close to the character boundary. Bigram color uniformity can be used to define the uniformity in the color of pixel that stands for the text character and attachment surface by means of a color pair.

b. EM-Based BC

Edge pixels are characterized by color pairs. Two pixels which have maximum color variation is identified among all color pairs. Make use of these color values for observing the color pair across the sides where the edge pixel is positioned. ***clL*** stand for the color component with lower intensity and ***clH*** for higher intensity component. ***clL*** and ***clH*** have three dimensional value in RGB space. These values symbolize the colors of text and attached plane. Coordinates of edge pixel *Pe is* used for observing spatial positions. By cascading the color values, an observation vector is defined. To normalize the dimensions of color pair and spatial position observation, the coordinates *spx* and *spy* are made three dimensions, to normalize dimensions of the color pair and the spatial observation. *Pe* can be described by observation vector x = [***clL***, ***clH***, ***spx***, ***spy***].

Observation points are clustered in order to extract the text boundaries from images. To scrutinize the distributions of observation points of edge pixels, GMM is used. In order to calculate K observation points, K means clustering is used. Means of K observation points is taken as *K* variances σi . A group of Gaussian distributions is initialized. Expectation of GMM is characterized by

$$P(x|\mu,\sigma) = \sum_{i=1}^{K} wiPi(x|\mu i,\sigma i)$$

where **x** stands for observation points, *wi* the weights of *i* th Gaussian distribution in the set, and μi and σi signifies the mean and variance of the *i* th Gaussian distribution.

EM algorithm is then applied to acquire the likelihood approximation of GMM parameters. In this process, the GMM parameters are iteratively rationalized from their original values derived by *K*-means clustering.

Boundary layer is assembled from the *K* Gaussian distributions. If the edge spawns the greatest likelihood in *i* th Gaussian distribution, it will be allocated into the *i* th boundary layer *Bi.*

$$\mathbf{Xi} = \{x\,j \in \mathbf{x} | \forall\, k \in [1, k],\ pi\,(x\,j\,|\mu i,\ \sigma i\ ) \geq pk(x\,j\,|\mu k,\ \sigma k\ )\}$$

$$Bi\ (p) = \begin{cases} 1, & \text{if } x\,p \in \mathbf{Xi} \\ 0, & \text{otherwise} \end{cases}$$

The probable value μi of *i* th Gaussian distribution presents a mean color pair {***cliL***, ***cliH***} to name all edge pixels at layer *Bi*.

B Stroke Segmentation

Initial evidence location of string and structure of character is given by the text boundaries. Sometimes the background interferences in images make the text boundary break into small segments. In order to localize the text with much accuracy, the mean color pair in each of the boundary layers is used to label the set of connected components as candidate text characters. Stroke is used as basic units for character labeling. A character composes of strokes with related width and multiple orientation.

Stroke can be defined as a associated image region with similar color and half-closed edge that keeps constant distance in a direction whereas remains extensible in perpendicular path. This consistent distance is termed as stroke width. The extensible direction is termed as stroke orientation. In order to retain the stroke of text candidates color assignment and structural analysis are carried out.

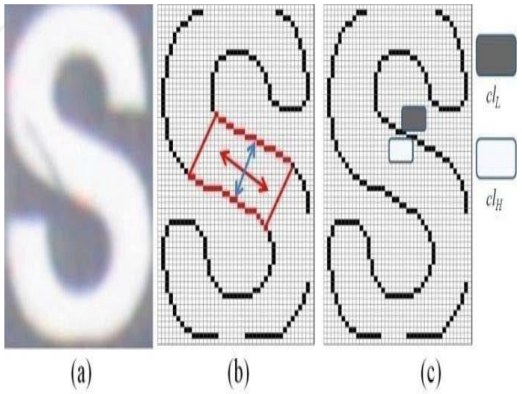


Fig 5: (a) Image patch of a text character. (b) Stroke is marked by red boundary, and red arrow indicates stroke orientations and blue arrow shows the stroke width. (c) Color assignment based on the mean color pair {***clL***, ***clH***} in the current boundary layer.

Color assignment allocates each pixel to the closer value of the mean color pair {***clL***, ***clH***} on aboundary layer by,

$$\mathbf{c}^* = \begin{cases} \mathbf{clL} & \|\mathbf{clL} - \mathbf{c}\| \leq \|\mathbf{clH} - \mathbf{c}\| \\ \mathbf{clH} & \|\mathbf{clL} - \mathbf{c}\| > \|\mathbf{clH} - \mathbf{c}\|. \end{cases}$$

where **c** denotes the original color of one pixel in RGB space.

C. String Fragment Classification (SFC)

The result of stroke segmentation is text and attached surface extracted as connected component. Layout analysis is performed to validate text among these connected components. The attachment plane appears as background board whereas text appears as text string. Groups of components are identified that have the probability to form text strings. Gabor-based text features are used to train SVM-based classifier of string to settle on whether a candidate string fragment is text piece or not.

a. Training Set

Perform the closest character grouping on scene images with text information to acquire a set of image pieces, in which string fragments are considered as positive samples, while non text outliers are considered as negative samples. Synthetic text characters and fragments are generated to create extra positive samples and negative samples.

These image patches are attuned to the standard size in order to train the string fragment classifier. If the width-to-height ratio seems to be greater than 1:1 ratio and less than 3:1, then the patch is normalized into width of 86 pixels and height of 48 pixels. If the ratio is greater than 3:1, then it is cut down vertically into overlapped patches ratio 2:1, and then it is normalized. Gabor-based text features are extracted based on this training set to train a classifier of string fragment in SVM model.

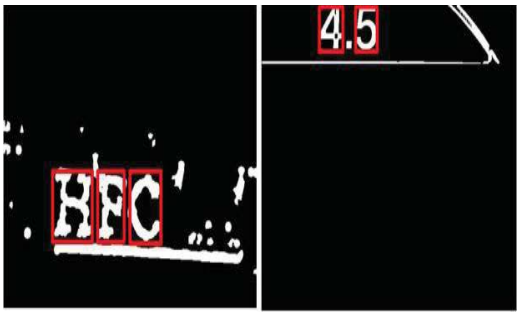


Fig. 6: Example of adjacent character grouping.

c. Gabor-Based Features

Gabor-based features are employed to eliminate the false-positive string fragments. Gabor filter can be used to examine the grouping and distributions of stroke components that are associated to the textures of text. Gabor filter responses are employed in this work in order to spot out pixels of interest (POI) for text feature extraction.

To characterize the stroke width and the stroke orientation from a pixel-based view, Gabor filter is adaptively created at each pixel of the string fragment for utmost Gabor responses. First, calculate edge map and the distance transform (DT) map of string fragments. DT map of a pixel point to its adjacent edge pixel *Pe* and the distance between them, *dp*. If *pixel* is positioned within a stroke, the line *PPe* should be vertical to the stroke direction. The pixel *P* is termed as the source pixel of Gabor filter. The compatible filter is probable to turn out the maximum Gabor reaction on the stroke of its source pixel. The compatible Gabor filter is then rotated by $\pi/2$ at each pixel to get an anticompatible Gabor filter and resultant Gabor response map. Compute the variation between two Gabor response maps, where the local maximum pixels at the map of absolute difference are extracted as POI. Feature extraction focuses on the POIs in the feature maps of stroke width, gradient and stroke distribution. Then the extracted feature is named Gabor-based text feature. The POIs model only the inner arrangement of text by the positions and orientations of the strokes. Statistics of stroke orientations is made based on the POIs in the training set collected from the string fragments. To divide the image pieces into three horizontal partition regions, a block pattern is used.

V. CONCLUSION AND FUTURE WORK

In this work, a new video and natural scene multi-oriented text detection method is proposed that makes use of the most convenient Canny edge detection algorithm and the boundary growing method. Classification of the text image and non text image is also done based on neural network concepts like EM based clustering, Stroke segmentation and String Fragment Classification. This work focused mainly on multi oriented text. For the detection of multi oriented text, only few methods are in use till now. This work is considered to be the most simplest method that make use of the popular methodology, canny edge map and boundary growing method to texture the boundary of the text in a scene and localize the same in a short span. A training set is developed to evaluate the performance of the proposed method. Experiments show that the proposed method is the most convenient method and outperforms the existing methods for localizing the text from video as well as natural scene images and classification of text and nontext images. Plans to extend this method to detection of curve shaped text lines with good recall, precision, F-measure and low computational time is being developed behind the curtain. Design of a more sophisticated algorithm to model the structure of text characters and strings and extension of the framework to localize nonhorizontal text strings in deformed surfaces, and design word recognition algorithm to read text information from text regions is also under consideration.

ACKNOWLEDGMENT

I thank everyone who supported me to do a research in this field and to publish a paper on this base.

REFERENCES

[1] X. Chen, J. Yang, J. Zhang and A. Waibel, "Automatic Detection and Recognition of Signs from Natural Scenes", *IEEE Transactions on Image Processing*, 2004, pp 87-99.

[2] Y. F. Pan, X. Hou and C.L. Liu, "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images", *IEEE Transactions on Image Processing*, 2011, pp 800-813. .

[3] K. Jung, K.I. Kim and A.K. Jain, "Text information extraction in images and video: a survey", *Pattern Recognition*, 2004, pp. 977-997.

[4] D. Crandall and R. Kasturi, "Robust Detection of Stylized Text Events in Digital Video", In Proc. of *ICDAR* 2001, pp 865-869.

[5] K. L Kim, K. Jung and J. H. Kim. "Texture-Based Approach for Text Detection in Images using Support Vector Machines and Continuous Adaptive Mean Shift Algorithm". *IEEE Transaction on PAMI*, 2003, pp 1631-1639.

[6] F. Wang, C. W. Ngo and T. C. Pong, "Structuring low quality videotaped lectures for cross-reference browsing by video text analysis", *Pattern Recognition*, 2008, pp 3257-3269.

[7] U. Bhattacharya, S. K Parui and S. Mondal, "Devanagari and Bangla Text Extraction from Natural Scene Images", In Proc. of *ICDAR* 2009, pp 171-175.

[8] Y. F. Pan, X. Hou and C. L. Liu, "Text Localization in Natural Scene Images on Conditional Random Field", In Proc. of *ICDAR 2009*, pp 6-10.

[9] A. K Jain and B. Yu, "Automatic text location in images and video frames", *Pattern Recognition*, 1998, pp 2055-2076.

[10] K. Jung and J.H. Han, "Hybrid Approach to Efficient Text Extraction in Complex Colour Images", *Pattern Recognition Letters*, 2004 pp 679–699.

[11] B. Epshtein, E. Ofek and Y. Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform", In Proc. of *CVPR*, 2010, pp 2963-2970

[12] P. P .Roy, U. Pal, J. Liados and F. Kimura, "Multi-Oriented English Text Line Extraction using Background and Foreground Information", In Proc. of *DAS*, 2008, pp 315-322.

[13] H. Li, D. Doremann and O. Kia, "Automatic text detection and tracking in digital video", *IEEE Transactions on Image Processing*, 2000, pp 147-156.

[14] Y. F. Pan, X. Hou and C. L. Liu, "Text Localization in Natural Scene Images on Conditional Random Field", In Proc. of *ICDAR 2009*, pp 6-10.

[15] H. Anoual, S. Elfkihi, and A. Jilbab, "Features extraction for text detection and localization," in *Proc. 5th Int. Symp. I/V Commun. Mobile Netw.*, 2010, pp. 1–4.

[16]] Y. F. Pan, X. Hou and C.L. Liu, "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images", *IEEE Transactions on Image Processing*, 2011, pp 800-813.

[17] U. Bhattacharya, S. K Parui and S. Mondal, "Devanagari and Bangla Text Extraction from Natural Scene Images", In Proc. of *ICDAR* 2009, pp 171-175.

[18] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in nature sceneswith Stroke width transform," in *Proc. IEEE Conf. Comput. VisionPattern Recogn.*, Jun. 2010, pp. 2963–2970.

[19] C. Yi and Y.Tian." Localizating text in scene images by boundary clustering, stroke segmentation ans string fragment classification",
IEEE Transactions on Image Process. Vol. 21.no. 9.Sep.2012
.