# Bandlet Based Video Completion Scheme After Selective Text Removal

**Gayathri R, Smitha P. S**

***Abstract:*** ***This paper presents a semi-automatic video text detection and removal along with a video completion scheme. In the video text detection stage, accurate edge locations are detected using a new type of image representation called as bandlets. Text locations are found by taking Stroke Width Transform (SWT) of the edge map and are grouped using Connected Components (CCs). Motion analyses of the video frames are done in order to preserve the spatial and temporal consistency of the video. After removing the unwanted text regions, an automatic inpainting scheme is employed to fill in the regions with appropriate data. The proposed inpainting scheme takes advantage of both structural and hybrid inpainting techniques. Evaluation of the approach is done using the user prepared video dataset along with ICDAR competition results. The experimental results demonstrate the effectiveness of both video text detection approach and completion technique, thereby the entire video.***

***Index Terms****: Bandlets, Connected Components, Spatial and temporal consistency, Stroke Width Transform*

## I. INTRODUCTION

Digital videos are common nowadays. Embedded texts in such videos provide useful information such as those appearing in news and other television broadcastings. But all of these texts may not be necessary as they hide some important portions of the video. Consider a case in which the user wants a video as such without the embedded texts in it. In such cases there should be an easy way to remove them and use the video accordingly. This highlights the need for an automatic approach to remove those texts and complete the video. Automatic text detection and video completion consists of mainly two steps:-1] Text detection and extraction 2] an automatic video completion scheme after the text removal. Although many methods have been proposed over the past few years, text detection is still a challenging problem because of the unconstrained colors, sizes and alignments of the characters. Moreover, scene text is affected by lighting conditions. The detection of texts with many sizes still poses a problem. So here, we propose a method where bandlets are used for text detection and the regions are inpainted using structural and hybrid features.

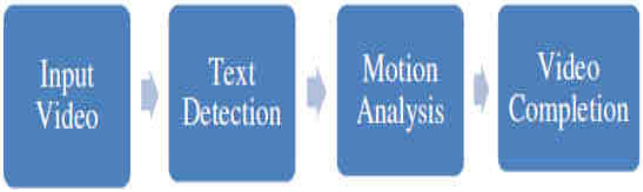


**Fig. 1: System Block Diagram**

**Revised Version Manuscript Received on June 29, 2015.**

**Gayathri R**, M.Tech Student, SCT College of Engineering, Pappanamcode, Trivandrum, Kerala, India.

**Smitha P. S**, Asst. Prof., SCT College of Engineering, Pappanamcode, Trivandrum, Kerala, India.

Fig 1 shows the overall proposed system. Input video is divided into frames and text detection is performed in each frame using bandlets. Motion Analysis of the video frames is done inorder to maintain the temporal and spatial consistency. Video Completion is performed in successive frames using patch based inpainting techniques.

## II. PREVIOUS WORKS

Existing text detectors are broadly classified into two main groups: texture (also called region) based and connected component (CC) based methods. Texture based approaches view text as a particular texture that is distinguishable from the background. Basically, features of various regions are retained .Finally, the text candidates are classified using an unsupervised or a supervised classifier and text blocks are generated according to some geometric features.CC based approaches extracts regions from image and uses geometric constraints to rule out candidate non-text candidates. The laplacian approach [3] consists of four steps: text detection, connected component classification, connected component segmentation, and false positive elimination. In the first step, we identify candidate text regions by using Fourier-Laplacian filtering. The second step uses skeletonization to analyze each CC in the text regions. Simple CCs are retained while complex CCs are segmented in the third step. False positives are removed in the last step.Edge-based methods are proposed. Liu et al.[4] extract statistical features from the Sobel edge maps of four directions and use K-means to classify pixels into the text and nontext clusters. Although this method is robust against complex background, it fails to detect low contrast text and text of small font sizes. It is also computationally expensive due to the large feature set. Another method is to design two filters to enhance the edges in text areas. This method uses various threshold values to decide whether to enhance the edges in a certain region, and thus may not generalize well for different data sets. For video completion, the challenge is making the completed regions consistent in the spatio-temporal domain. Several approaches have attempted to resolve this. An approach proposed in [4] uses structure repair and texture propagation. To repair the structure regions, the structure interpolation uses the new model's rotated block matching to estimate the initial location of completed regions and later refine the coordinates of completed regions. The information in the neighboring frames then fills the structure regions. To complete the structure regions without tedious manual interaction, the structure extension utilizes the spline curve estimation. Afterwards, derivative propagation realizes the texture region completion. Another approach in [7] uses bandlet based edges for video completion. It uses a 3D video volume completion technique to complete the missing regions of video. Apart from all these approaches, our technique uses bandlet bases to detect the text locations



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

and uses a patch based algorithm to inpaint the missing regions produced by selective text removal.

## III. TEXT DETECTION

The input video is divided into frames. Each frame is analyzed. The image intensities are linearly adjusted to enhance the contrast. Video text detection consists of mainly 3 steps.1] Edge detection using bandlets,2]Stroke Width Transform and CC generation

### 1] Edge detection using bandlets

Bandlet transform [2], [6] is performed on the original frame, and for each segmentation square $S$ the bandlet coefficients are generated. For each $S$, the resulting coefficients are grouped in low-pass (approximation) and high-pass filtering results similar to the 1D wavelet transform. We discard the approximation part and only process the high-pass coefficients. The first-order derivatives of the fine-detail bandlet coefficients are computed. By applying a contextual filer, we find local maximum of the resulting gradient signal since many meaningful edges can be found in the local maxima of the gradient not only in the global maxima. Then, in order to improve the quality of the edge image a two level thresholding is employed. For each point $x_i$ in the gradient signal, we check if $x_i$ is a local maximum and its value is greater than a threshold $T$. If so, $x_i$ is kept as an edge point coefficient otherwise it will be discarded. Hence, a window with size $2L+1$ centered at $x_i$ is set. Then, the binary indicator of edge points in the gradient signal is generated as follows

$$M_i = \begin{cases} 1 & \text{if } g_i > T_G \wedge g_i > g_j,\ \forall j \in [i-L, i-1] \wedge \\ & g_i > g_j,\ \forall j \in [i+1, i+L] \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $g_i$ represents the gradient value for $x_i$ and $g_j$ indicates gradient value of neighboring pixels of $x_i$ that exist is the window. $M$ is a map of local maxima of the gradient signal. The corresponding locations of 0's of $M$ in the bandlet fine (high-pass) coefficients are set to 0, for all the bandlet squares $S$. Then, the inverse bandlet transform is performed in order to have the final edge locations of the original image. Inorder to determine the threshold T, a two level thresholding is employed. First, the edge detection is performed using a low value T for which an edge map $E_l$ is performed. Next, the edge detection is performed using a high value T for which an edge map $E_h$ is performed. The edge locations in $E_l$ and $E_h$can be combined to get good results. For each edge component $C_h$ in $E_h$, we inspect $E_l$ and find if any edge component $C_l$ in $E_l$ coincide with $C_h$. If so, $C_l$ is taken and saved in the final image.


**Fig. 2 (a): Input video frame**

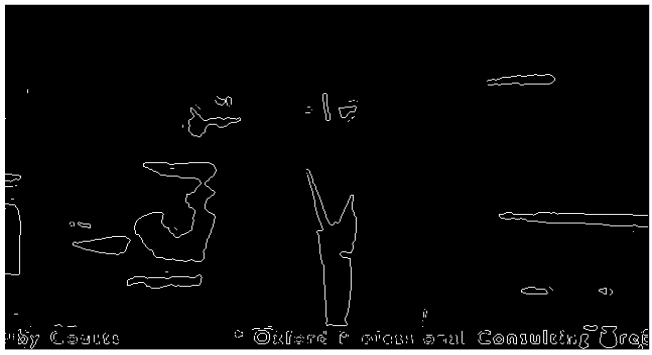

**Fig. 2 (b): Bandlet based edge detection**

### 2] Stroke Width Transform and CC generation

The Stroke Width Transform (SWT) value of each pixel is roughly the width of the stroke that contains the pixel. A stroke is defined as a part of the image that forms a band of constant width. We determine the stroke width using a novel approach based on distance transform, which differs drastically from the SWT proposed in [1].The proposed method guarantees that the SW information is provided at every pixel of the original CC with any stroke shape. The main steps are the following:-

i] The Euclidean distance transform is applied to label each foreground pixel with the distance to its nearest background pixel. The ridge values of the distance map correspond to half the width of the stroke.

ii] Then, we propagate the stroke width information from the ridge to the boundary of the object. The method bypasses the need to locate ridge pixels by iteratively propagating the stroke width information, starting from the maximum value to the minimum value of the distance map.

iii] We exclude CCs with a large standard deviation. The rejection criterion is std/mean > 0.5, which is invariant to scale changes. This threshold was obtained from the training set of the ICDAR competition database. When CCs are generated, filtering of the same must be done inorder to filter out non-text candidates. Those with nearby strokewidths are combined together inorder to obtain a text block. Again these text blocks are grouped together to form letter candidates.

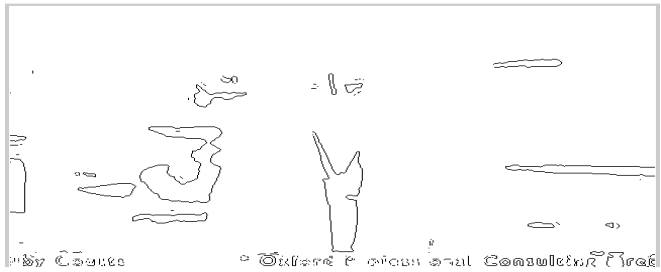

**Fig. 3 (a): Stroke width image**

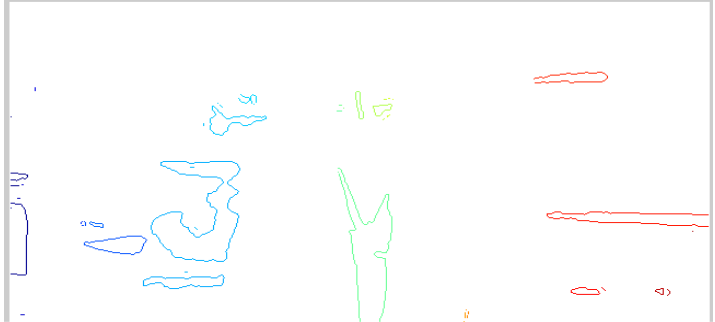

**Fig. 3 (b): Connected Component Generation**



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

## IV. MOTION ANALYSIS

Once the text locations are detected within each frame of the video sequence, a mechanism is needed to distinguish the video texts from the natural texts that may exist in a frame. Considering that an embedded video text appears in a consequence of frames with specific motion properties compared to the rest of the video, we employ a tracking and motion analysis scheme in order to specify the video text regions. The detected text locations in each image are considered as different objects and CAMSHIFT algorithm is performed on each of them. It is worth noting that CAMSHIFT starts from a text object of the current frame if the text object has not been already tracked in the sequence. Therefore, the large set of text locations of all the frames is reduced to a set of tracked text objects in the video. For each text object we have the spatial and temporal locations. In the next step, the local motion field of each text object and the global motion field of the video are estimated using Lucas-Kanade optical flow computation algorithm.

## V. VIDEO COMPLETION

Spatial patch blending is used here. In classic patch-based inpainting methods,the reconstruction of an image is a kind of patchwork. Patches are iteratively extracted from the image, cut up, and the remaining pieces are pasted inside the region to complete the given image. The main idea of the spatial patch blending is to point out the fact that parts of the individual patches are discarded during sequential compositing,but these parts contain valuable information that could have been used if a different insertion order had been used. In this method, the scrapped offcuts are kept and spatially blended in order to reduce seams between the pieces of patches pasted side by side. This method is defined as a pixelwise process. After patch based blending, frame adjustment process is performed.It uses a panoramic mosaic of the video to regulate the luminance of each frame and reduce the intensity of flicker in the video.

## VI. EXPERIMENTAL RESULTS

As there is no standard data sets available, video clippings have been collected from news clippings, sports videos, environment videos etc .For comparison purposes, ICDAR competition results as well as other techniques using sobel and canny edge detectors are used.

### VI.1 Performance Measures

We consider the performance of the text detection algorithm at a block level. Three types of blocks is considered.

i) Truly Detected Block (TDB): A block that actually contains text candidates. It can be one candidate or higher.

ii)Falsely Detected Block (FDB):A block that is detected as a text block but does not contain any text candidates.

iii) Text block with missing data (MDB): A block with missing text candidates.

The performance measures can be defined as follows:

A) Recall (R) = TDB/ATB,

B) Precision (P) = TDB/ (TDB+ FDB)

C) F-measure (F) = 2 * P * R / (P + R)

D) Misdetection Rate (MDR) = MDB/TDB.

Video completion scheme is tested with respect to the mean error rate between the actual frame of the input video and its inpainted frame. Mean Square Error between the frames can be plotted and different techniques are analyzed. The videos contain several news clippings, sports videos etc.

**Fig. 4(a): Hardcoded Video frame**

**Fig. 4(b): Inpainted Video frame**

## VII. CONCLUSION

In this paper, video texts are detected and removed. The missing regions due to text removal are then inpainted thereby completing the video. The technique helps in the removal of unwated texts obstructing different parts of the video. The inpainting algorithm maintains the video in its original way and it can be used according to each user's priority.

## REFERENCES

[1] B Epshtein, E Ofek, Y Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform", Micsosoft Corporation.

[2] S Mallat,G Peyre, "Orthogonal Bandlet Bases for Geometric Images Approximation" in *Communications on Pure and Applied Mathematics*, Vol.000, (2000).

[3] P Shivakumara, T Q Phan, C L Tan, "A Laplacian Approach to Multi-Oriented Text Detection in Video" in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol .33, No.2, Feb 2011.

[4] T H Tsai, C L Fang, "Text-Video Completion Using Structure Repair and Texture Propagation", in *IEEE Transactions on Multimedia*, Vol .13, No.1, Feb 2011.

[5] Y F Pen, X Hou, C L Liu, "A Hybrid Approah to Detect and Localize Texts in Natural Scene Images", in *IEEE Transactions on Image Processing*, Vol.20, No.3, March 2011.

[6] A Mosleh, N Bouguila, A B Hamza, "Automatic Inpainting Scheme for Video Text Detection and Removal", in *IEEE Transactions on Image Processing*, Vol. 22, No.11, Nov 2013.

[7] A Mosleh, N Bouguila, A Ben Hamza, "Bandlet-based sparsity regularization in video inpainting" in *J. Vis. Commun. Image R.* 25 (2014) 855–863.



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

**Gayathri R** is currently doing M.Tech. Degree in Signal Processing with the Department of Electronics and Communication Engineering, SCT College of Engineering, Pappanamcode, Trivandrum, Kerala. She received the B.Tech degree from the University of Kerala, Thiruvananthapuram, in 2013 in Electronics and Communication Engineering. Her research interests include areas in signal and image processing.

**Smitha P. S** is currently working as an Assistant Professor under Electronics and Communication Department in SCT College of Engineering, Pappanamcode.. She did her M.Tech (2011) in Communication Engineering from National Institute of Technology, Karnataka, India. She is also the second rank holder of the M.Tech degree. She did her B.Tech (2002) in Electronics and Communication Engineering from University of Kerala, India.

.



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*