An Approach to Detect Text and Caption in Video

Miss Megha Khokhra

[1] *M.E Student Electronics and Communication Department, Kalol Institute of Technology, Gujarat, India*

ABSTRACT

The video image spitted into number of frames, each frame maintains the text. Then the Image is converted into Gray Scale to avoid the text color variation.[5] A single value is corresponding to gray value and detecting the edge. Detecting the edge process is the boundary between two regions with relatively distinct gray-level properties.[2] One is the horizontal direction of the image. Another is the vertical direction of the image. The features to describe text regions are area, saturation, orientation, aspect ratio and position. Then convert into binary image. We propose a simple but efficient methodology for text detection in video frames. The method is based on the gradient information and edge map selection.[1] In the proposed method we first find the gradient of the image and then enhance the gradient information. We later binarized the enhanced gradient image and select the edges by taking the intersection of the edge map with the binary information of the enhanced gradient image. We use the edge detector for generating the edge map. The selected edges are then morphologically dilated and opened using suitable structuring elements and used for text regions. We then perform the projection profile analysis to identify the boundary of the text region. At the end, we implement a false positive elimination methodology to improve the text detection results. Then we make the video from image frames which contain the detection of text. After Detection of text Converted Image into text file and main aim is to convert that text into speech output.

Keyword - *Text detection, Image frames, Caption video, Detect text from video, Edge detection, Gradient Based text detection, Extract Text, Optical Character recognisation, Converted into Text file, Text to Speech etc….*

1. INTRODUCTION

A variety of approaches to text information extraction (TIE) from images and video have been proposed for special applications including page segmentation, address block location, license plate location, and content-based image/video indexing. In spite of such extensive studies, it is still not easy to design a general-purpose TIE system. This is because there are so many possible sources of variation when extracting text from a shaded or textured background, from low-contrast or complex images, or from images having variations in font size, style, color, orientation, and alignment. These variations make the problem of automatic TIE extremely difficult [9].

In text detection and extraction, there are three main categories that are connected component based, texture based, and edge and gradient based methods. It is revealed based on literature review that connected-component based methods are not robust because they assume that text pixels belonging to the same connected region share some common features such as colour or grey intensity. On the other hand, texture based methods tend to be computationally expensive for large databases as they involve expensive operations such as DCT for text detection in images. The edge and gradient based methods have been developed to reduce the number of computations in detecting text in the images. However, the existing methods based on edge and gradient are not robust to complex background as they give more false positives. In addition, selection of threshold values to classify text pixel from non text pixels is another major problem [6-9]. To overcome these problems, the method in [10] introduced the segmentation of text portion with the help of candidate text block identification. A method based on uniform colours in L* a* b* space is also proposed in [5] to locate uniform coloured text in video frames [6].

In this Paper we propose a simple methodology to extract text from unconstrained complex background and low intensity video frames. The proposed method is based on the gradient information and edge map selection. In this methodology we first take the gradient of the whole image and then enhance the gradient information using a suitable mask. We later binarized the enhanced gradient image and select the edges by taking the intersection of the edge map with the binary information of the enhanced gradient image. We use the canny edge detector for generating the edge map. The selected edges are then passed through morphological operator using suitable structuring elements. Based on this morphologically operated information we select the text region. We perform the projection profile analysis to identify the boundary of the text region. Then we implement a false positive elimination methodology to reduce the number of false detection in a frame [1]. At the end we save detected text into text and converted it in to text to speech [5].

2. BACKGROUND

2.1 Text Detection

Text detection and recognition in images and video frames, which aims at integrating advanced optical character recognition (OCR) and text-based searching technologies, is now recognized as a key component in the development of advanced image and video annotation and retrieval systems. Unfortunately, text characters contained in images and videos can be any gray-scale value (not always white), low-resolution, variable size and embedded in complex backgrounds. Experiments show that applying conventional OCR technology directly leads to poor recognition rates. Therefore, eAcient detection and segmentation of text characters from the background is necessary to 3ll the gap between image and video documents and the input of a standard OCR system.

2.2 Video Text detection

Video text containing important and reliable semantic information can contribute significantly to video retrieval and understanding. Video text detection is the first and important step for retrieving video text. [6]

In today's world, video plays an important role as a media delivered over TV broadcasting, internet and wireless network. It is often required to automatically detect and extract the text information from these video frames. [8]

The video image spitted into number of frames, each frame maintains the text. Then the Image is converted into Gray Scale to avoid the text color variation. A single value is corresponding to gray value and detecting the edge. Detecting the edge process is the boundary between two regions with relatively distinct gray-level properties. One is the horizontal direction of the image. Text another is the vertical direction of the image. [5] Video Images contains two types of texts: (1) caption text which is manually edited, and (2) scene The quality of video degrades when it is transmitted through a medium and this makes the video frames of low intensity and low contrast. These are the main problems of working with video images. [3]

Generally, video which exists in video naturally. Since caption text is edited, we can expect such texts to be of a good clarity or contrast, which are usually aligned in either horizontal or vertical direction. On the other hand, scene text is a part of a video frame/image and its nature is unpredictable. Thus it suffers from non-uniform illumination, perspective distortion, low resolution, low contrast, varying font types and font sizes, multiple colors and arbitrary orientations, etc. The presence of such types of texts in video adds more complexity to the text detection and tracking problem. Therefore, accurate text detection and tracking in unconstrained environments is still an elusive goal for researchers.

Though video images often suffer in degradation during transmission through various media, text portions can always be distinguished due to its discriminative pixel values with respect to the background. Text portions in an image always have distinct intensity values with respect to its background. The differences in the pixel values of an image are noted in the gradient of that image. Based on this observation our proposed method performs using gradient information and text edge map selection. Generally the approach to detect text and caption in videos consists of the following steps:

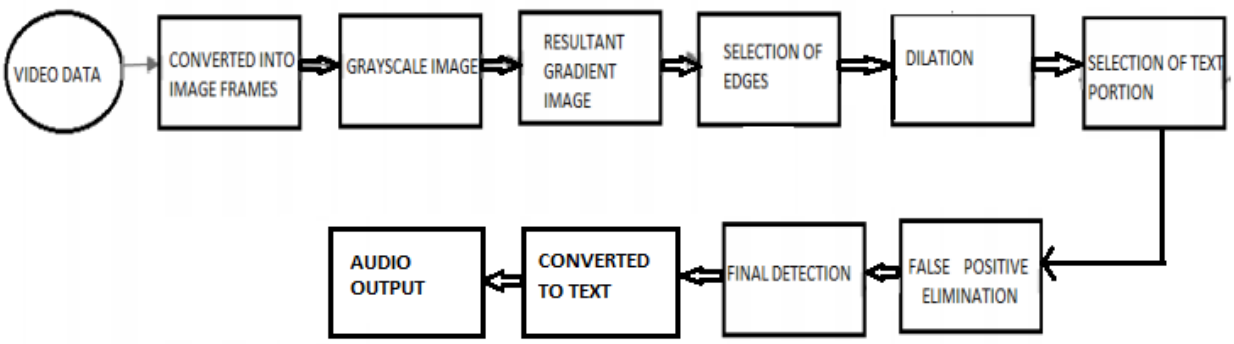

Fig -1: Block Diagram for Text Detection in Video

3. THE PROPOSED METHOD

We describe the entire proposed methodology subsequently in the order where sub-section A describes gradient based procedure; edge based procedure is described in sub-section B; uniform color based procedure are discussed in C [1].

3.1 Gradient Based Procedure

We here propose an improved method for creating gradient image because our intention was to create more and more edge pixels in the text portion of the image. In this method we first create horizontal and vertical gradient image of the original image and then merge them to get the resultant gradient image. To get the horizontal gradient image (HGI) of the original image (I) we consider the corresponding gray image (GI) of that image. The gradient value of a particular pixel (i, j) of HGI is calculated by taking the difference of the immediate lower pixel (i+1, j) with the current pixel (i, j) of the gray image (GI). HGI (i , j) = GI (i , j) – GI (i+1 , j) ∀(i, j) ∈ GI and (i +1, j) ∈ GI Similarly the gradient value of a particular pixel (i, j) of the vertical gradient image (VGI) of the image (I) is calculated by taking the difference of the immediate right pixel (i, j+1) with the current pixel (i, j) of the gray image in the following way. VGI (i,j) = GI (i, j) − GI (i, j +1)| ∀(i, j) ∈ GI and (i +1, j) ∈ GI The horizontal and vertical gradient images are then merged to find the Resultant gradient image (RGI). RGI (i, j) = HGI (i, j) + VGI (i, j) ∀(i, j) ∈ HGI IVGI The resultant gradient images are shown in Figure 2.

(a) (b) (c)

Fig -2 (a) Horizontal Gradient Image (b) Vertical Gradient Image

(c)Resultant Gradient Image

3.2 Edge Based Procedure

In this stage we first apply canny edge detector to identify the edges in the image. We then select edges of the image by taking the intersection of the binarized information of the enhanced gradient image with the edge map. The original canny edge image and the selected edge image are shown in Figure 3.

Fig -3 (a) Original Canny Image **Fig.3: (b) Selected Canny Image**

3.3 Uniform Color Based Procedure

In this stage we take the intersection of the images we got after opening morphological operation and the binarized gradient image. We next perform projection profile analysis (horizontal projection followed by vertical projection) to determine the boundary of the text region. We also applied the projection profile analysis for the second time within the rectangle (text boxes determined after the first projection profile) to make the boundary more accurate which further helps us to reduce number of inaccurate text Boundary. We next applied the false positive removal methodology. In that method they used a scaling factor α for scaling both the internal and external boxes of the detected text block. In that algorithm set the scaling factor α to 0.2. We did a simple modification in this method. In our case we set two different scaling factors a and b for scaling internal and external boxes of the original box respectively. We set a=0.4 and b=0.2 experimentally on the dataset of images we considered.

3.4 Text to Speech

There are about 45 million blind people and 135 million visually impaired people worldwide. Disability of visual text reading has a huge impact on the quality of life for visually disabled people. Although there have been several devices designed for helping visually disabled to see objects using an alternating sense such as sound and touch, the development of text reading device is still at an early stage. Existing systems for text recognition are typically limited either by explicitly relying on specific shapes or colour masks or by requiring user assistance or may be of high cost. Therefore we need a low cost system that will be able to automatically locate and read the text aloud to visually impaired persons. The main idea of this project is to recognize the text character and convert it into speech signal. The text contained in the page is first pre-processed. The pre-processing module prepares the text for recognition. Then the text is segmented to separate the character from each other. Segmentation is followed by extraction of letters and resizing them and stores them in the text file. These processes are done with the help of MATLAB.

4. EXPERIMENTS AND RESULTS

4.1 Comparison

To give an objective comparison of all the above methods, we define the following performance metrics, i.e. detection rate, false positive rate and misdetection rate. The metrics are defined as follows. The detected text blocks are represented by their bounding boxes. A text block is considered as truly detected if the bounding box covers all or some characters in the same text block. If a truly text block has some characters not covered by the bounding box,

it is considered as a misdetection. To judge the correctness of the text blocks detected, we manually count Actual Text Blocks (ATB) in the images in the dataset.

Also we manually label each of the detected blocks as one of the following categories:

Truly detected text blocks (TDB): a detected block that contains text fully or partially.

Falsely detected text blocks (FDB): a detected block that does not contain text.

Text block with missing data (MDB): a truly detected text block that misses some characters.

Based on the number of blocks in each of the categories mentioned above, the following metrics are calculated to evaluate the performance of the methods:

Detection rate (DR) = Number of TDB /Number of ATB.

False positive rate (FPR) = Number of FDB / Number of detected text blocks (Truly + falsely).

Misdetection rate (MDR) = Number of MDB / Number of TDB.

The results in terms of the above metrics are reported in Table 1. The table shows the performance of the proposed method in comparison with the existing methods where we can see that the detection rate, false positive and misdetection rates of the proposed method are higher than the three existing methods.[1]

Table -1: Performance of proposed the existing methods

Method	DR	FPR	MDR
Proposed Method	High	Low	Low
Edge based	80.0	18.3	20.1
Gradient based	71.0	12.0	10.0
Uniform color based	51.3	27.3	37.3

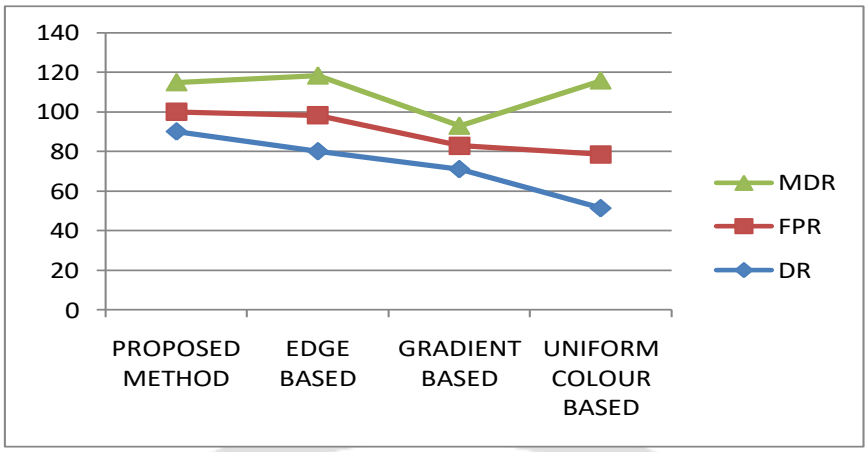

Chart -1 Comparison

4.2 Result of this Paper

Fig 4 RESULT 1

Fig 4 RESULT 2

5. CONCLUSION

We have developed a video text and caption detection system. Viewing the corner points as the fundamental feature of character and text in visual media, the system detects video text with high precision and efficiency. We built up several discriminative features for text detection on the base of the corner points. These features can be used flexibly to adapt different applications. We also presented a novel approach to detect moving captions from video shots. Optical flow based motion feature is combined with the text features to detect the moving caption. Over 90% detection ratio is attained. The results are very encouraging. Most of the algorithms presented in this paper are easy to implement and can be straightforwardly applied to caption extraction in video programs with different languages. Our next focus will be on the word segmentation and text recognition based on the results of text detection.

6. REFERENCES

[1]. Chong Yu, Yonghong Song , Quan Meng, Yuanlin Zhang, Yang Liu "Text Detection And Recognition In Natural Scene With Edge Analysis"[2015]

[2]. Liang Wu, Palaiahnakote Shivakumara, Tong Lu, Member, IEEE, and Chew Lim Tan "A New Technique For Multi-Oriented Scene Text Line Detection And Tracking In Video"[2015]

[3]. A. Dutta, U. Pal, Shivakumara, Ganguli, Bandyopadhya and L. Tan, "Gradient based Approach for Text Detection in Video Frames", CVPR Unit, Indian Statistical Institute, Kolkata, India,[2009]

[4]. B.H.Shekar and Smitha M.L. "Morphological Gradient based Approach for text localization in video/scene images"[2014]

[5]. Vladimir Y. Mariano Rangachar Kasturi "Locating Uniform- Colored Text In Video Frames"[2000]

[6]. Xiaodong huang,"A novel of approach to detecting scene text in video"[2011]

[7]. Ping Hu, Weiquiang Wang, Ke Lu,"Video text detection with text edges and convolutional neural network".[2013]

[8]. Xiangrong Chen and Hongjiang Zhang "Text Area Detection from Video Frames", Microsoft Research China, [2000]

[9]. Ze-Yu Zuo, Shu tian, Wei- Yi Pei, Xu-Cheng Yin , " Multi –Strategy Tracking Based Text Detection in Scene Videos."[2015]

[10]. Christian Wolf AND Jean-Michel Jolion, "Model based text detection in images and videos," in Proc. IEEE Conf. Comput. Vis. Pattern Rec 2004.

[11]. Xu-Cheng Yin, Senior Member,IEEE,Ze-Yu Zuo,Shu Tian, and Cheng-Lin Liu,Fellow,IEEE." Text Detection,Tracking and Recognition in Video : A comprehensive Survey."

[12]. Guozhu Liang, Palainhnaktoke Shivakumara, Tong lu, Member IEEE and Chew lim tan, Senior Member,IEEE."Multi-spectral Fusion based Approach for Arbitrarily-Oriented Scene Text Detection in Video Images".