A Comparative Approach for Overlay Text Detection and Extraction from Complex Video Scene

Anupama S. Budhewar, Ravindra C. Thool

Department of Computer Science and Engineering,
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded,
budhewar.anupama@gmail.com, rcthool@yahoo.com

Abstract—Overlay text brings important semantic clues in video content analysis such as video information retrieval and summarization, since the content of the scene or the editors intention can be well represented by using inserted text. Most of the previous approaches to extracting overlay text from videos are based on low-level features, such as edge, color, and texture information. However, existing methods experience difficulties in handling texts with various contrasts or inserted in a complex background. In this paper, we propose a novel framework to detect and extract the overlay text from the video scene. Based on our observation that there exist transient colors between inserted text and its adjacent background, a transition map is first generated. Then candidate regions are extracted by a reshaping method and the overlay text regions are determined based on the occurrence of overlap text in each candidate.

In this paper represents comparative approach between edge and change in intensity models that are computationally fast and invariant to basic transformations like horizontal, vertical and scaling. We demonstrate that change in intensity context can be used to detect overlay text regions. Intensity context describes all boundary points of the shape with respect to any single boundary point.

Index Terms—Optical Character Recognition (OCR).

I. INTRODUCTION

With the development of video editing technology, there are growing uses of overlay text inserted into video contents to provide viewers with better visual understanding. Most broadcasting videos tend to increase the use of overlay text to convey more direct summary of semantics and deliver better viewing experience. For example, headlines summarize the reports in news videos and subtitles in the documentary drama help viewers understand the content. Sports videos also contain text describing the scores and team or player names [1]. In general, text displayed in the videos can be classified into scene text and overlay text [2]. Scene text occurs naturally in the background as a part of the scene, such as the advertising boards, banners, and so on. In contrast to that, overlay text is superimposed on the video scene and used to help viewers understanding. Since the overlay text is highly compact and structured, it can be used for video indexing and retrieval [6]. Overlay text brings important semantic clues in video content analysis such as video information retrieval and summarization, since the content of the scene or the editors intention can be well represented by using inserted text [8]. Most broadcasting news videos tend to increase the use of overlay text that usually represent names of anchors persons, places, persons, or description of news in crisp. Moreover in sports news it may be name of player, type of sport, location, score and many more.

However, overlay text extraction for video optical character recognition (OCR) becomes more challenging, compared to the text extraction for OCR tasks of document images, due to the numerous difficulties resulting from complex background, unknown text color, size and so on. There are two steps involved before the overlay text recognition carried out, i.e., detection and extraction of overlay text. First, overlay text regions are roughly distinguished from background. The detected overlay text regions are refined to determine the accurate boundaries of overlay text strings. To generate a binary text image for video OCR, background pixels
are removed from the overlay text strings in the extraction step. Although many methods have been proposed to detect and extract the video text, a few methods effectively deal effectively with different color, shape, and multilingual text. Digital video libraries and archives of immense size are becoming accessible over data networks. Efficient video retrieval and browsing has become crucially important. Understanding the semantic contents of the video and using them for indexing is inevitable. Automatic indexing and retrieval of video information based on content is very challenging research area [1]. A vast variety of techniques are proposed in literature for video analysis ranging from extraction of low-level features to high-level semantic features and all these techniques are based on color, texture, shape, sound, text, and objects. Of all the available techniques of the video annotation, only text analysis is useful for the high-level semantic directly, while other techniques require an extra effort to produce high-level semantics [6].

II. LITERATURE REVIEW

Now see methods to detect and extract overlay text using low-level features such as edge, color and texture information and experience difficulties in handling text with various contrasts. The commonly adopted method is to apply an edge detector to the video frame and then identify regions with high edge density and strength. Some employ the high-frequency wavelet coefficients and connected components to detect the text regions. Most of the existing methods are based on approaches such as edge, color, and stroke. Few of them are as follows: M.R. Lyu, J. Song, and M. Cai implemented a method, there are two main approaches the first is language dependent and other language independent[2]. In that character recognition is based on language or character dependent English language is giving more accurate results. It is elaborated more as follows According to the linguistic classification, English, French, and Spanish belong to alphabetic literal, whereas Chinese, Japanese, and Korean belong to ideograph. Their differences in the following four aspects affect the video text processing. Zhong Y. et al .[4] extracted text as those connected components that follow certain size constraints. In [7] stroke model based character extraction from grey level document images are proposed. The method extracts the text with known stroke width but fails to distinguish noise in linear form from true strokes. Liu Y. et.al.[1] proposed a hybrid approach to detect the text regions and expectation maximization algorithm to binarize the text regions. Kasar et.al. [3] Proposed an edge based connected component approach to detect the characters in camera based document images and used a specialized binarization to separate the characters from the background. The survey indicates that all the above methods fail to isolate the foreground text in document images having complex background.

N. Otsu implemented another approach [11]in 1979. It is color based text extraction method. It is simple and efficient method. It is not robust for text extraction with similar color background. Otsu method [14] is a widely used color-based text extraction method due to its simplicity and efficiency of the algorithm. However, Otsu method is not robust to text extraction with similar color of background due to the use of global thresholding. To solve this problem, the detected text regions are divided into several blocks and then Otsu method is applied locally to each block, such as the adaptive thresholding introduced in [7], where a dam point is defined to extract text strings from background. T. Sato implemented another method in Jan.1998 [4]. It considers stroke of text in horizontal, vertical, up-right, and up-left directions and generate the edge map along each direction. Apply an interpolation filter based on vertical, horizontal, left diagonal, right diagonal directions to enhance the performance of text extraction.

The concept of text composed of uniform color was introduced by L. Agnihotri and N. Dimitrova [5] in Jun.1999. Color-based approaches assume that the video text is composed of a uniform color. In the approach by Agnihotri et al. [4], the red color component is used to obtain high contrast edges between text and background. In [5], the uniform color blocks within the high contrast video frames are selected to correctly extract text regions. K. C. K. Kim et al introduced the concept of cluster color based on Euclidean distance in the RGB space and used 64 cluster color channels for text detection [9]. Color-based approaches assume that the video text is composed of a uniform color. In the approach by Agnihotri [5], the red color component is used to obtain high contrast edges between text and background. In his approach, the uniform color blocks within the high contrast video frames are selected to correctly extract text regions. C. Liu, C. Wang, and R. Dai presented an approach based on edge [13]. Edge based approach is very useful for overlay text detection since text region contains rich edge information. This method performs well if there is no complex background. Use a modified edge map with strength for text region detection and localize the detected text regions using coarse-to-fine projection. They also extract text strings based on local thresholding and inward filling X. Liu and J. Samara bandu presented[14] based on the silent point detection and the wavelet transformation has also been used to detect the text regions. Use multiscale edge detector to detect the text regions. They compute the edge strength, density, and orientation variance to form the multiscale edge detector. Texture-based approaches, such as the salient point detection and the wavelet transform, have also been used to detect the text regions. Bertini et al. [10] detect corner points from the video scene and then detect the text region using similarity of corner points between frames. Sato et al. [11] apply an interpolation filter based on vertical, horizontal, left diagonal, right diagonal directions to enhance the performance of text extraction. Gllavata et al. [2] employ the high-frequency wavelet coefficients and connected components to detect the text regions. However, since it is almost impossible to detect text in a real video by using only one characteristic of text, some methods take advantage of combined features to detect video text [12], [13]. In this paper we are using the color models for finding out the intensity of each pixel. With the help this we can remove the minimum intensity pixels and
highlight only high intensity pixels. To find out the intensity following color models are used.

III. Overlay Text Detection using Change in Intensity

In this section it represents the different methods for overlay text detection and extraction based on low level features. Overlay text brings important semantic clues in video content analysis such as video information retrieval and summarization. The implemented method is robust to different character size, position, contrast and color. Text in images and video frame carries important information for video content understanding and video retrieval.

In general, text displayed in the videos can be classified into scene text and overlay text. Scene text occurs naturally in the background as a part of the scene, such as the advertising, boards, banners, and so on. In contrast to that, overlay text is superimposed on the video scene and used to help viewers understanding. Since the overlay text is highly compact and structured, it can be used for video indexing and retrieval. However, overlay text extraction for video optical character recognition (OCR) becomes more challenging, compared to the text extraction for OCR tasks of document images, due to the numerous difficulties resulting from complex background, unknown text color, size and so on. There are two steps involved before the overlay text recognition is carried out, i.e., detection and extraction of overlay text. First, overlay text regions are roughly distinguished from background. The detected overlay text regions are refined to determine the accurate boundaries of overlay text strings. To generate a binary text image for video OCR, background pixels are removed from the overlay text strings in the extraction step. Although many methods have been implemented to detect and extract the video text, few methods can effectively deal with different color, shape, and multilingual text.

To solve this problem, the detected text regions are divided into several blocks and then Otsu [14] method is applied locally to each block, such as the adaptive thresholding introduced in, where a dam point is defined to extract text strings from background. On the other hand, some filters based on the direction of strokes have also been used to extract text in the stroke-based methods.

In this approach a method for new overlay text detection and extraction method using the transition region between the overlay text and background is represented. First, it generates the transition map based on our observation that there exist transient colors between overlay text and its adjacent background. Then the overlay text regions are roughly detected by computing the density of transition pixels and the consistency of texture around the transition pixels. The detected overlay text regions are localized accurately using the projection of transition map with an improved color-based thresholding method to extract text strings correctly. It generates the transition map and refines the detected text regions.

The overall procedure for overlay text detection & extraction is based on overlay text detection model as shown in Figure 1 [1], where each module is applied on n, where n is number of frames.

Fig. 1. Block Diagram of Overall procedure of the Text Detection Method

The different steps are as follows,

A. Transition Map Generation

To effectively determine whether a pixel is within a transition region, the modified saturation is first introduced as a weight value based on the fact that overlay text is in the form of overlay graphics.

B. Candidate Region Extraction

If a gap of consecutive pixels between two nonzero points in the same row is shorter than 5% of the image width, they are filled with 1s. If the connected components are smaller than the threshold value, they are removed. The threshold value is empirically selected by observing the minimum size of overlay text region. Then each connected component is reshaped to have smooth boundaries.

C. Overlay Text Region Determination

The next step is to determine the real overlay text region among the boundary smoothed candidate regions by some useful clues, such as the aspect ratio of overlay text region. Since most of overlay texts are placed horizontally in the video, the vertically longer candidates can be easily eliminated. The density of transition pixels is a good criterion as well. LBP is a very efficient and simple tool to represent the consistency of texture using only the intensity pattern [16].

D. Overlay Text Region Update

The overlay text region update method can reduce the processing time efficiently [19]. Overall procedure of extraction method discussed. And two opposite scenarios, in which either the overlay text is darker than the surrounding background[20].
Since it is confirmed that the overlay text is always bright in each text region, it is safe to employ Lyus method to extract characters from each overlay text region. First, each overlay text region is expanded wider by two pixels to utilize the continuity of background. This expanded outer region is denoted as ER. Then, the pixels inside the text region are compared to the pixels in ER so that pixels connected to the expanded region can be excluded. We denote the text region as TR and the expanded text region as ETR, i.e. Next, sliding-window based adaptive thresholding is performed in the horizontal and the vertical directions with different window sizes, respectively[21].

E. Modified Dam Point Labeling

Compared to the Lyus method, the height of expanded text region is not normalized in our method. Let and denote gray scale pixels on ETR and the resulting binary image, respectively. All are initialized as White. The window with the size of is moving horizontally with the stepping size 8 and then the window with the size of is moving vertically with the stepping size. If the intensity of is smaller than the local thresholding value computed by Otsu method in each window, the corresponding is set to be Black. The process of applying the sliding windows for adaptive thresholding[22].

F. Inward Filling

Authors assume that the background pixels in TR are generally connected to ER in terms of intensity. They use filling from ER to the connected pixels in TR to remove the background pixels. However, since the text pixels might be connected to the background pixels in TR, the unwanted removal can occur when the filling task is conducted. Performance measurement is done on the basis of precision and recall, probability of error parameters. The recall and precision is used to detect overlay text with higher efficiency. Recall is used to detect pixels other than overlay text pixels. Precision is used to detect the overlay text region[23].

IV. RESULT ANALYSIS

A. Comparative Result between Change in Intensity and Edge based Approach

Here we are comparing the results of overlay text detection and extraction using change in intensity and edge based approach in figure 2. It shows from transition map generation it generates the candidate regions. In change in intensity approach due to LBP false rate is negligible whereas in edge based approach it is nearly 7% as shown in above figure.

B. False Alarm

False alarm rate depends upon one phase thresholding. If edge gradient is higher than false alarm rate is lower and vice versa[25]. So edge gradient is inversely proportional to false alarm rate. Figure 3 shows the accuracy of overlay text detection based on intensity based approach.

To detect text in edge based approach pixels based and box method is used the results are as shown in Table 1 and Table 2 respectively.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Detection rate (%)</th>
<th>False Alarm Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box based</td>
<td>92.5</td>
<td>7.8</td>
</tr>
<tr>
<td>Pixels based</td>
<td>94.8</td>
<td>7.1</td>
</tr>
</tbody>
</table>

V. CONCLUSION

A novel method for overlay text detection and extraction from complex videos is implemented in this approach. The
TABLE II
FALSE ALARM RATE AND PROCESSING TIME USING EDGE BASED APPROACH

<table>
<thead>
<tr>
<th>Method</th>
<th>Detection rate (%)</th>
<th>False Alarm rate (%)</th>
<th>Processing Time (CPU/MBD)</th>
<th>Processing Time (CPU/TMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y. Liu</td>
<td>98.5</td>
<td>7.1</td>
<td>0.51</td>
<td>8.82</td>
</tr>
<tr>
<td>MGD (Max. Gradient Diff)</td>
<td>90.7</td>
<td>13.9</td>
<td>0.79</td>
<td>3.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REFERENCES</th>
</tr>
</thead>
</table>

### Detection Method

The detection method is based on the observation that there exists a transient color between inserted text and its adjacent background. The method is useful for real-time application. An experimental result shows that change in intensity based approach gives 99% hit rate and 1% false rate. Whereas in edge approach implemented an effective video text detection method based on line features. This method exploits an improved canny edge detector to detect text pixels. By considering spatial distribution of edge pixels stroke information is incorporated in text region generation and filtering.

This approach represents the distinguishing features of edge based. To detect the text canny edge detector is used because it acts as an optimal edge detector. From experimental results it is observed that edge based approach is having 95% edge detection where is change in intensity approach is having 99% correct overlay text detection rate. The 7% false rate is present in edge based approach where as in change in intensity approach only 1 to 2% false rate is present.

Then line map generated to detect continuous lines and remove isolated lines which are placed horizontally. In this way we detect overlay text using edge based approach. An edge based method increases the computational cost whereas change in an intensity based approach reduces the computational cost.

### References