Overlay Text Detection in Complex Video Background

Chun-Cheng Lin[1]

[1]Department of Electrical Engineering
National Chin-Yi University of Technology
Taichung, Taiwan
cclin@ncut.edu.tw

Bo-Min Yen[2]

[2]Department of Electrical Engineering
National Chin-Yi University of Technology
Taichung, Taiwan
s49712118@student.ncut.edu.tw

Abstract

The subtitles on video are very useful for us to understand video's contents. If we can extract text information from the subtitles, it will be very helpful to establish a database of video's content that includes annotations and indexes. To extract text from videos, most text detection and extraction methods use the text color, background contrast, and texture information. However, the limitations of the existing methods are various contrasts and complex backgrounds in the text processing. The purpose of this study is to develop a method to detect overlay text in the complex video background. This study utilized the characteristics that there are transient color changes existing between text and its adjacent background, to generate transition maps. And this work applied the connected components to smooth the connected rectangular shapes. The study results demonstrated that the text detection method based on the transition map can effectively detect the text regions under different text contrasts, fonts and background complexities.

Keywords: Transition map, Overlay text detection, Video subtitles, Local binary pattern, Texture analysis.

1. Introduction

Along with the development of the video text editing technology which is used to edit the texts in videos, the audience can better understand the content of the videos, such as the titles of TV programs or video content described below. So far in the video transmission aspect, the most popular media types are mostly the TV programs, Internet, or wireless network. In order to let viewers can quickly find interesting content from videos, many researchers have made many efforts in video indexing and summarizing [1] [2], especially for the text added to the video content. Three major reasons are the following. First, the added texts are closely associated with the video content. Second, it has a higher contrast ratio. Third, in the most advanced countries, the optical character recognition (OCR) technology is more robust than other existing speech analysis technologies. Therefore, almost all of the video indexing research starts with video text recognition.

The text recognition technique is usually divided into three steps: text detection, extraction and recognition. The text detection step is to differentiate between the text area and non-text area, and to determine the accurate text string's boundary by carefully considering of detecting area. The extraction step is to keep only text pixels in the text area and remove background pixels. The text detection and extraction steps will generate a binary text image. The following recognition step is to identify text via commercial OCR software. In the past years, many text recognition methods have been proposed and applied in the complex backgrounds, different font sizes, and multi-language texts.

The text detection methods can be further divided into three types of characteristics: color-based, edge-based and texture-based. The color-based method assumes that text color is formed of a uniform color. For example, Agnihotri *et al*. [3] adopted the red color component to obtain the highly contrast edges between the text and background. In [4], the blocks with the uniform color are chosen to extract the text area. The edge-based method is also considered as an effective way to detect text because the text area contains a lot of edge information. Lyu *et al*. [2] proposed a modified edge map based on the image strength to detect the text area. Their detection method used the projection of coarse-to-fine, and the way that they extract the text was based on the local threshold and inward filling. Liu *et al*. [5] used multi-stage edge detector to calculate the edge strength, density and direction variance to detect text areas. The texture-based method is based on that the characters have a specific arrangement way, and there has a color difference between text and background. Several methods based on Gabor filtering [6], the analysis of spatial variance [7] and wavelet transform [8], have been proposed to perform the calculation of the characteristics of texture blocks. They also apply different classification tools to identify the text areas and non-text areas, for example, neural networks [8] or support vector machine [9].

The text extraction methods can also be divided into two types. One is based on the color, and the other is based on the stroke of words. One of the color-based approaches is the polarity of the text

color which requires you to ensure whether it is bright or dark text first. The method proposed by Kim *et al*. [1] first tested whether the polarity of the text color needs to be changed, and then they applied local threshold values and inward filling to extract the texts. The stroke-based methods are based on the stroke direction filters for text extraction. A four direction character extraction filter [10] was developed to increase the stroke-like shapes and to suppress others. However, the cross of strokes may also be suppressed due to language-dependent some characters without obvious stripe shape.

In order to detect overlay text from complex video background, this paper applied a new method based on the transition map for the detection of the subtitle area. This method used the feature that the transient color changes existed between the text and the adjacent background. Based on this feature, we calculate the saturation value and the image intensity to produce a transition map. Then the changes of the transition pixels in the transition map of text areas were analyzed to generate a link map based on connected components. Based on the assumption that text areas are mostly rectangular shape, we made a smooth link map into rectangular candidate text frame. This study also introduced a texture-based method to determine which candidate text areas are real ones. The study results indicated that the method based on the transition map can accurately detect blocks of text in the video with complex background. This paper will introduce the approach of overlay text detection from complex video background in the second section. The third section presents the study results of the method based on the transition map. Finally, the fourth section includes discussion and conclusions.

2. Overlay text detection from the video

The proposed method is based on the transient color changes existing between the text and adjacent background (such as the bright text on a dark background, or the dark text on a bright background), as shown in Fig. 1. Because the subtitles are embedded in the video background, they have a higher degree of saturation. Therefore, the image saturation and strength were then calculated to define the transition map that is necessary for the image localization.

Figure 1. Examples of video subtitles.

2.1 Transition map generation

In general, the texts in the subtitle usually have a consistent color. However, the image compression with loss may also cause a transient change in color between the text and background. In order to determine whether a pixel falls into a transition region, this study calculated the modified saturation value as a weight value of the transition map, as follows [1].

$$S(x,y) = 1 - \frac{3}{(\mathrm{R}+\mathrm{G}+\mathrm{B})[\min(\mathrm{R},\mathrm{G},\mathrm{B})]} \tag{1}$$

$$\tilde{S}(x,y) = \frac{S(x,y)}{\max(S(x,y))} \tag{2}$$

where

$$\max(S(x,y)) = \begin{cases} 2\times(0.5-\tilde{I}(x,y)), & \text{if } \tilde{I}(x,y) > 0.5 \\ 2\times\tilde{I}(x,y), & \text{otherwise.} \end{cases}$$

$S(x,y)$ and $\max(S(x,y))$ are the saturation value and the maximum value of saturation at the corresponding intensity level, respectively. $\tilde{I}(x,y)$ denotes the intensity value of the pixel (x,y), which is normalized to the interval of [0,1]. The maximum value of saturation in (2) is normalized according to $\tilde{I}(x,y)$ compared to 0.5. The combination of intensity change and the modified saturation can be used to describe the transition, defined as follows [1].

$$D_L(x,y) = (1+dS_L(x,y))\times\left|I(x-1,y)-I(x,y)\right| \tag{3}$$

$$D_R(x,y) = (1+dS_R(x,y))\times\left|I(x,y)-I(x+1,y)\right|$$

where

$$dS_L(x,y) = \left|\tilde{S}(x-1,y)-\tilde{S}(x,y)\right|$$

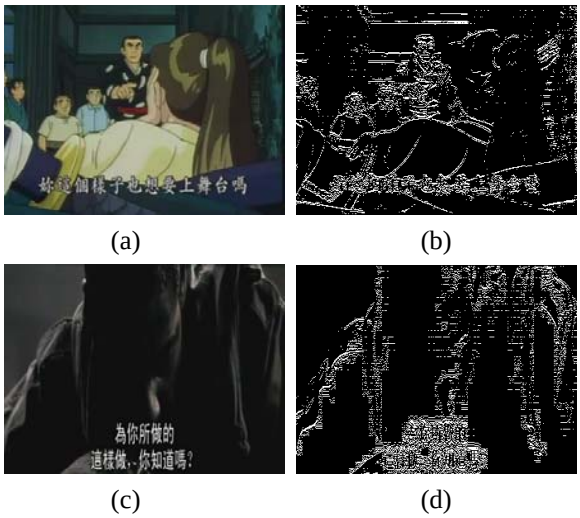
and

$$dS_R(x,y) = \left|\tilde{S}(x,y)-\tilde{S}(x+1,y)\right|$$

As the weight $dS_L(x,y)$ and $dS_R(x,y)$ may be 0 in the colorless text and background, a constant value of 1 was added in the weight. The transition map was then defined as

$$T(x,y) = \begin{cases} 1, & \text{if } D_R > D_L + T_h \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

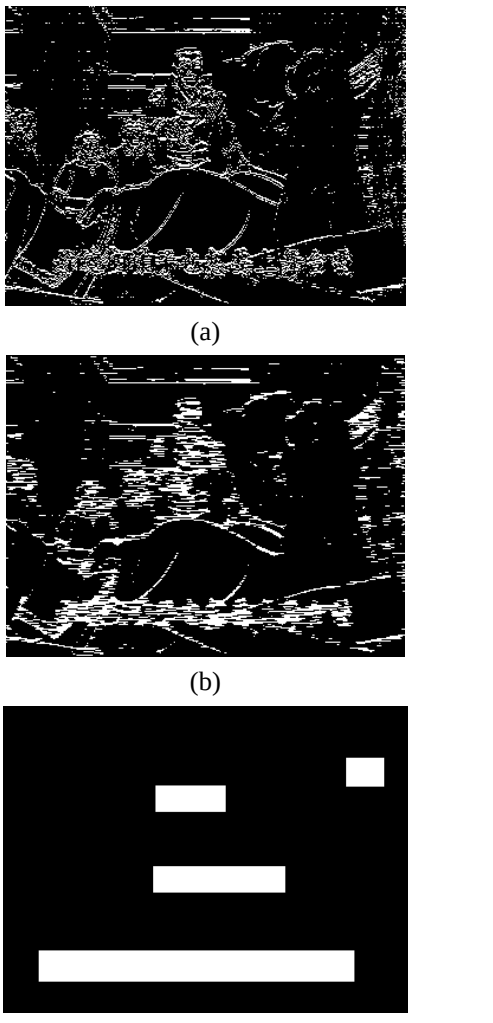
The threshold value T_h is set to 1 empirically in this study. Figure 2 shows an analytical result for the transition map generation. Figure 2(a) and (c) are two video images with overlay texts. The corresponding transition maps generated by equations (1) to (4) are shown in Fig. (c) and (d). It can be observed that the transition map can be created form the complex video backgrounds.



(a) (b)

(c) (d)

Figure 2. (a) and (c) are video images with overlay texts; (b) and (d) are the corresponding transition maps.



(a)

(b)

(c)

Figure 3. (a) an example of transition map; (b) the linked map through connected components building; (c) the candidate text regions established from linked map boundary filling.

2.2 Search for candidate text regions

Given the transition map as shown in Fig. 3(a), we further assume that near the text areas have more transition pixels. In order to create the connected components [11], this study searched all of the pixels in the transition map. If the distance between two non-zero pixels are shorter than 5% of the image width in the same row [1], they are filled with 1s (i.e. white color), as shown in Fig. 3(b). It is reasonable to assume that most of the candidate text areas have rectangular shapes. Therefore this study found out four corners of the rectangular regions, (min_x, min_y), (max_x, min_y) ,(max_x, max_y) ,and (min_x, max_y), as shown in Fig. 4. A threshold value was further used to remove unwanted areas according to practical minimum size of text area including the width and height of the text string. The candidate text regions were then filled with white colors. Figure 3(c) demonstrated the refined candidate regions.

2.3 Determination of overlay text regions

The determination of the real text regions from candidate text ones can be based on some simple clues; for example, the ratio of height and width of the subtitle string, or the location of the candidate text region on the video. Because most of the text strings are in a horizontal direction, the text string in the vertical direction can be easily removed. However a precise algorithm is needed to lower the false detection of the true text regions as far as possible. This study introduced a texture-based method based on local binary pattern (LBP) [1, 8] to determine the true text regions.

Because LBP can be used to describe the consistency of texture, it is suitable to analysis the complex structure that can cause large intensity changes surrounding the transition pixels. LBP is a binary code created by the current pixel and its all circular neighbor pixels and can be converted into a decimal number. The LBP is defined as follows.

$$\mathrm{LBP}_{P,R} = \sum_{n=0}^{P-1} s(g_n - g_c)2^n \tag{5}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

P and R are the pre-selected number of circular neighbor pixels and the radius of circle, respectively. g_c and g_n denote the intensity of the current pixel and the circle's neighbor pixels, respectively. Figure 5 is an example of LBP calculation. The local binary pattern is $\mathrm{LBP}_{6,3} = 2^5 + 2^3 + 2^2 = 64$.

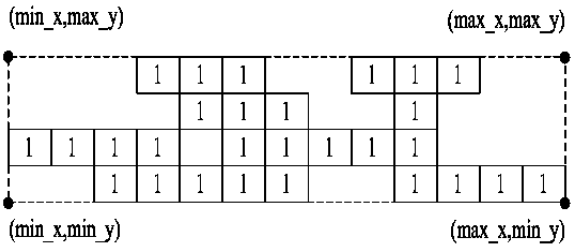


Figure 4. Search for four boundary points of a connected region.

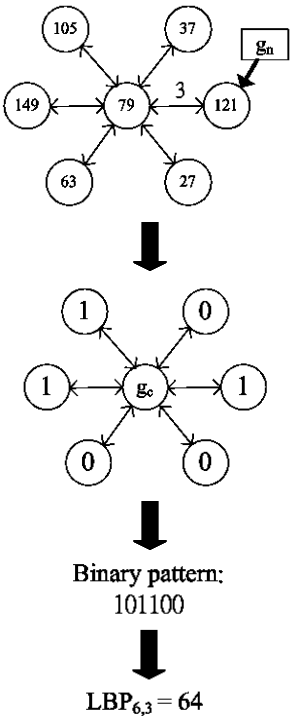


Figure 5. An example of LBP calculation.

Next we define the probability of texts (POT) [1] as follows.

$$\mathrm{POT}_i = \omega_i \times \mathrm{NOL}_i, \quad i = 1, \ldots, N \tag{6}$$

where ω_i is the density of the transition pixels in each candidate text area, and it can be calculated from dividing the number of transition pixels by the size of each candidate region. N is the number of the candidate text region. NOL_i is the number of different LBPs, which is normalized by the maximum of the number of different LBPs in each candidate area. A candidate text region is considered as an overlay text region, if its POT value is larger than a predefined value 0.3 in this study. Figure 6 demonstrated the result of an example for the determination of the overlay text region. Figure 6(a) showed the candidate text regions. The overly text region shown in Fig. 6(b) had a POT which is larger than 0.3. Figure 6(c) displayed the original video image corresponding to the overlay text region.

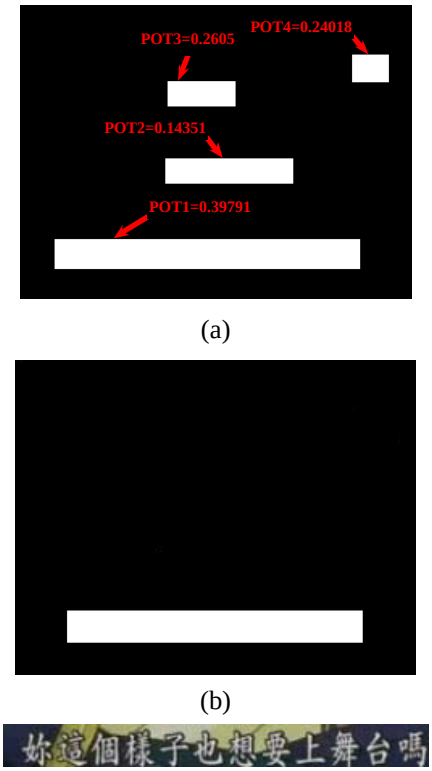


Figure 6. (a) an example of the candidate text region; (b) the overlay text region determined by the analysis of POT; (c) the original video image corresponding to the overly text region.

3. Results

Because the overlay texts are embedded in the video with complex background, this causes the difficulties of the text detection. Therefore, this paper introduced the method based on the transition map to detect the overlay texts in video. This study adopted several video images with different background complexities to test the performance of the method described in this study, as shown in Fig. 7(a), (c) and (e). The image sizes were fixed at 340×255 pixels. The detection results of overlay text regions are shown in Fig. 7(b), (d) and (f). It is obvious that the proposed method can accurately detect the overlay text regions even in different complexities of video background.

4. Discussion and Conclusions

The overly text detection method described in this study is mainly based on the transient color changes existing between text strings and the adjacent backgrounds. The creation of the transition map is according to a modified saturation defined in eqs. (1) and (2), and the combination of intensity change and the modified saturation defined in eqs. (3) and (4). The connected components were then used to

produce the linked map and generate the smooth candidate text regions with rectangular shapes. Further this study calculated the density of the transition pixels and the number of different LBPs in each candidate region to determine the real overlay text region. The study results can confirm that the method based on the transition map can accurately detect the overly text regions in complex backgrounds. The future work of this study will be focused on the text extraction. If the binary texts can be extracted from the overly text region, we can further combine the commercial OCR software to perform the recognition of the subtitles in videos.

References

[1] W. Kim, and C. Kim, "A new approach for overlay text detection and extraction from complex video scene", *IEEE Trans. Image Proc.*, Vol. 18, pp. 401-411, Feb. 2009.

[2] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction", *IEEE Cir. and Sys. for Video Tec.* Vol. 15, pp. 243-255, Feb. 2005.

[3] L. Agnihotri, and N. Dimitrova, "Text detection for video analysis", *IEEE Content-Based Access of Image and Video Libraries*, pp. 109-113, Jun. 1999.

[4] X. S. Hua, P. Yin, and H. J. Zhang, "Efficient video text recognition using multiple frame integration", *IEEE Conf. Image Processing. 2002.* Vol. 2, pp. 397-400, Dec. 2002.

[5] X. Liu, and J. Samarabandu, "Multiscale edge-based text extraction from complex images", *IEEE Conf. Multimedia and Expo, 2006.* pp. 1721-1724, July 2006.

[6] A. Jain, and S. Bhattacharjee, "Text segmentation using gabor filters for automatic document processing", *Machine Vision and Applications*, Vol. 5, pp. 169-184, June 1992.

[7] V. Wu, R. Manmatha, and E.M. Riseman, "TextFinder: an automatic system to detect and recognize text in images", *IEEE Trans. Pattern Anal. Mach.* Vol. 21, pp. 1224-1229, Nov. 1999.

[8] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video", *IEEE Trans. Image Processing*, Vol. 9, pp. 147-156, Jan 2000.

[9] K. C. Jung, J. H. Han, K. I. Kim, and S. H. Park, "Support vector machines for text location in news video images", *IEEE Region 10 conf. Syst. Technology Next Millennium*, Vol. 2, pp. 176-180, Sep. 2000.

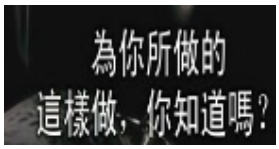
[10] T. Sato, T. Kanade, E.K. Hughes., and M.A. Smith, "Video OCR for digital news archive", *IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 52-60, Jan. 1998.

[11] Gonzalez, and Woods, "Digital Image Processing", 3rd Ed, Publisher: Prentice Hall, 2008.

(a)

(b)

(c)



(d)

(e)

(f)

Figure 7. (a), (c) and (e) are original video images; (b), (d) and (f) are the detection results for the overly text regions.