Video Text Extraction Based on Stroke Width and Color

Xiaodong Huang,[1] Qin Wang, Kehua Liu, Lishang Zhu

Abstract. Video text can be used to infer important semantics information of video. Video text extraction is a crucial step to recognize video text. Most of previous works can not extract the text well when the text regions have complex background or text character stroke width is unknown. Therefore, we propose a text extraction method, which can accurately extract the video text in complex background and various text character sizes. Experimental results on TV news video show the encouraging performance of the proposed algorithm.

Keywords: Text extraction, K-means, High-frequency Emphasis, Text Stroke Width

1 Introduction

Text contains semantic information and thus can contribute significantly to video retrieval and understanding. Therefore, video text recognition is crucial to the research in all video indexing and summarization. Video text recognition is generally divided into four steps: detection, localization, extraction, and recognition [1]. The text extraction step removes background pixels in the text rows and the text pixels are left for the recognition. Therefore, the performance of text extraction will determine the final text recognition. In this paper, we mainly discuss the text extraction.

Video text is generally divided into two types: the superimposed text (added during the editing process) and the scene text (existing in the real-world objects and scenes). This paper focuses on the former type. The text extraction in video frames is difficult because of complex background, unknown text character color, and various stroke widths [1].

The text extraction methods are classified into three classes. The first class [2,3] uses threshold-based methods to retrieve text region. Graham *et al.* [2] com-

[1] Xiaodong Huang (✉)
Capital Normal University, Beijing 100048, China
e-mail: dawn_hxd@yeah.net

pared several single-stage thresholding algorithms which use either global or local thresholding techniques, and found that the performance can be improved by automatic selection or combination of appropriate algorithm(s) for the type of document image. The threshold-based method mainly processes the gray-level image. However, it is not be robust enough to handle the complex backgrounds.

The second class uses the stroke-based methods to retrieve text region. Chen *et al.* [4] proposed two groups of asymmetric Gabor filters which can efficiently extract the orientation and scale of the stripes in a video image. These features are used to enhance contrast at only those edges most likely to represent text. However, this method is sensitive to the text stroke width. As a result, the method is not suitable for retrieving the binary text images in video.

The third class is the color-based methods. Garcia *et al.* [5] obtain best non-quantified results with hue-saturation value (HSV) color space. Leydier *et al.* [6] use a serialization of the k-means algorithm for ancient documents with heavy defects and transparency. Two color spaces RGB and HSL (Hue, Saturation, Luminosity) are chosen to handle several degradations. However, this method needs a user interface to define the number of logical classes and select the multiple color samples for each logical class, which initialize the original centers of the clusters. Liu *et al.* [7] proposed an approach which extracts characters from image with complex background and other objects. Density-based clustering method and a new histogram segmentation is employed to find and segment characters.

Different from the above methods, Lyu *et al.*[1] proposed a language-independent text extraction method that consists of three steps: 1) adaptive thresholding; 2) dam point labeling; and 3) inward filling. The computation cost of this method is low and algorithm is fast, which is suitable for the quick Video OCR, however, in complex background, it will remove some text character pixels in its third step of inward filling.

According to the above analysis, we find that most of previous works can not extract the text well when the text regions have complex background or text character stroke width is unknown. Thus, we propose a new text extraction method using the stroke and density features. Fig. 1 illustrates the flowchart of our method.

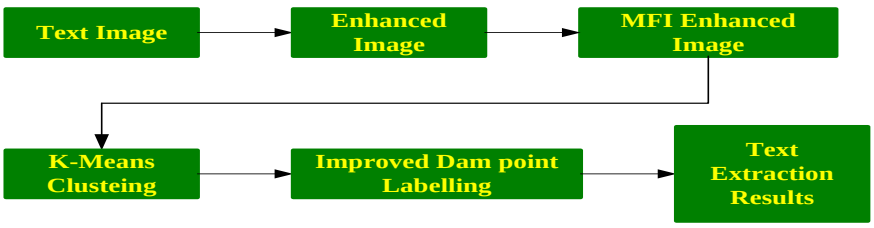


Fig.1 The Flowchart of the Method

First, we use the high frequency emphasis to enhance the text images in R,G,B channels respectively, which can enhance the text character contrast with the background. Then we use the multi-frame integration to synthesize the enhanced

RGB channels image, which can get the text images which have higher contrast with the background. Then we perform the K-means clustering to extract the text line. Finally, we use Boris *et al.*[10] method to get the character stroke width, which is robust to various character sizes. Then we perform the dam point labeling and inward filling based on the character stroke width to get the final text extraction results, which can extract text in any unknown character stroke width.

The rest of this paper is organized as follows. Section 2 describes to use the high-frequency emphasis to enhance image. Text line extraction based on color clustering is described in Section 3. In section 4, we describe how to filter the noise points based on the text stroke width. Experimental results are presented and discussed in Section 5. Finally, in Section 6, we draw conclusion.

2 Image Enhancement Using High Frequency Emphasis

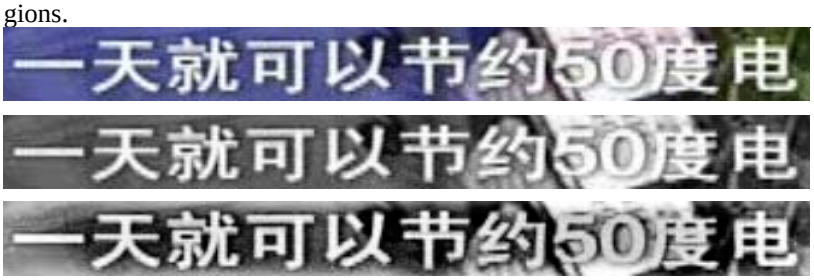
In [8], Gonzalez etc. propose a high-frequency emphasis, which has a filter transfer function given by

$$H_{hfe}(u,v) = a + b\ H_{hp}(u,v) \tag{1}$$

where $a \geq 0$ and $b > a$. Typical values of a are in the range 0.25 to 0.5 and typical values of b are in the range 1.5 to 2.0. H_{hp} is a high pass filter function.

Therefore, we first use the High-Frequency Emphasis [8] to enhance the text regions. We accentuate the high-frequency components of text regions to enhance the text regions contrast. The text region enhance process is shown in Fig.2.

Fig.2(a) is the original image. Fig.2(b) is the gray image. Fig.2(c) is the enhanced text image. Compared the Fig.2(c) with Fig.2(b), we can find that after the text regions enhancement by the high-frequency emphasis filtering, the text regions of the enhancement image own higher contrast than that of the gray image. Therefore, the text regions can be more easily segmented from the background regions.



(a)

(b)

(c)

Fig.2 The text region enhance process

3 Text Line Extraction Based on Color Clustering

Because the text character color is important clues to extract text in general, we perform the color clustering to retrieve several candidate binary images according to color continuity feature. However, in some complex background, because the text regions color is similar to the background colors, it is very difficult to extract the text regions completely. As a result, if we perform the color clustering in RGB space, we can not get the satisfied text extraction results.

According to the analysis of section 2, we can get the enhanced image based on frequency emphasis filtering. Therefore, we first use frequency emphasis filtering to enhance the color image in R, G, B channels respectively. Then we can get three enhanced RGB image.

In text line extraction, it is very difficult to extract the text lines when the text lines color is similar to the background. As a result, we enhanced the RGB channels images. However, the enhanced image still can not guarantee that the character can be segmented from the background. Therefore, we use the Multi-frame Integration to produce the synthesized image in which the text character contrasts highly with the background.

We can get I_{min}, I_{max} and I_{mean} from the three enhanced RGB image, which is similar to [9].

$$I_{\min}(x,y) = \min(p_i(x,y)) \tag{2}$$

$$I_{\max}(x,y) = \max(p_i(x,y)) \tag{3}$$

$$I_{mean}(x,y) = mean(p_i(x,y)) \tag{4}$$

i={R,G,B}. where $p_i(x,y)$ can be the R,G,B channel value in (x,y). We use the formula (2)(3)(4) to synthesize the three enhanced RGB images, and the results is shown in Fig.3. Then we use the I_{min}, I_{max} and I_{mean} as three channels to form a enhanced color image. The Fig.4(a) is the original text line. The Fig.4(b) is the enhanced color image. Compared Fig.4(a) with 4(b), we can find that the enhanced color image has higher contrast than that of original image.

Finally, we perform the K-means clustering with the combination of Euclidean distance in enhanced RGB space. In some complex background situation, text lines are composed of text characters, character contours, background and some noise points. As a result we set the k to 4 in our experiments. The clustering results are shown in Fig.4(c)(d)(e)(f).

(a) I_{min} (b) I_{max} (c) I_{mean}.

Fig. 3 enhanced image integration

(a) (b) (c) (d) (e) (f)

Fig. 4 Text Extraction Results Based on Color Clustering

(a) (b) (c) (d)

Fig. 5 Text Filtered Results Using Stroke Width

4 Text Region Filtering Using Stroke Width

We use Lyu's method [1] to extract the character binary image. Lyu etc. use the dam point labeling to extract the dam point in the binary image, and then they use inward filling to remove the noise points which surround the character.

Lyu define the dam point as follows:

$$\begin{aligned} Dam\ Point\ s = \{(x,y) \mid B(x,y) = "WHITE" \wedge MIN\_W \\ \le MIN\left[H\_LEN(x,y), V\_LEN(x,y)\right] < MAX\_W\} \end{aligned} \tag{5}$$

In [1], According to Lyu's observation, the stroke width (in pixels) in a normalized text image varies between *MIN_W*= 1and *MAX_W*= 3.

However, because the text character stroke width is various, MAX_W can not be a fixed value. Therefore, we should give different MAX_W values according to different character stroke width.

In [10], Boris *et al.* propose a fast method to get the character stroke width. As a result, we use the Boris's method to get the median character stroke width MAX_W. Then we use the inward filling to process the clustering results. The filtered results are shown in Fig.5. Then, we can get k different binary images.

However, we should find a way to select one clustering results as the final text extraction results.

We define the character ratio CR (formula (6)) to select one of the clustering results as the final text extraction results. The CR is used to evaluate the ratio between the character area and whole text regions.

$$CR = \frac{Character\ Area}{h * w} \tag{6}$$

where Character Area is the numbers of white pixels in every clustering results. h and w is the height and width of text regions.

If CR meets the formula(7), we select the binary image as the final text extraction results.

$$0.5 * \frac{MAX\_W}{h} < CR < 0.8 * \frac{MAX\_W}{h} \tag{7}$$

Based on the above computation, we can select Fig.5 (a) as the final text extraction result from k clustering results.

5 Experiments and Discussion

We compare our text extraction method with Lyu's method [1] and Otsu's method [11] in our experiment. We tested a lot of broadcast TV videos based on the three text extraction methods. Our test data contain different types, including Chinese and English video, which are captured from TV. The video resolution is 640×480 and 720×576. A total of 4210 text rows in 3500 frames of video are used to test the proposed algorithm. The text rows which contain the text character are generated by the previous text detection and localization experiments. Fig.6 illustrates experiment results of three text extraction methods in different background complexities. In Fig. 6, we find that our method performs well for images with complex backgrounds.

We use the character error rates (CER) [1] to evaluate the performance of the three methods by OCR software TH-OCR XP which can recognize English and Simplified Chinese character. For the total of 63271 English and Chinese characters in the 4210 text rows, the CERs of three methods are shown in Table 1.

Table 1 CERS Evaluations of Three Text Extraction Methods

CER / Methods	Simplified Chinese CER	English CER	Overall CER
Our Method	*0.215*	*0.162*	*0.189*
Lyu's Method	*0.293*	*0.198*	*0.246*
Otsu's Method	*0.613*	*0.375*	*0.494*

Fig. 6 Comparison of Three Text Extraction Methods. The first row of (a)(b)(c)(d) is the original text lines. The second row of (a)(b)(c)(d) is our extraction results. The third row of (a)(b)(c)(d) is the results of Lyu's method. The fourth row of (a)(b)(c)(d) is the results of Otsu's method.

6 Conclusions

In this paper, we proposed a new video text extraction approach. Because of the complex background and unknown character stroke width, video text extraction is difficult to extract. We propose a text extraction method, which can accurately extract the video text in complex background and various text character sizes. First,

we use the high frequency emphasis to enhance the text images. Then we use the multi-frame integration to synthesize the enhanced RGB channels image. Then we perform the K-means clustering to extract the text line. Finally, we use Boris's method to get the character stroke width, which is robust to various character sizes. Then we perform the dam point labeling and inward filling based on the character stroke width to get the final text extraction results.

Our experimental results and the comparisons with other methods are listed, which show that our method is capable of extracting English and Simplified Chinese text character accurately and is robust to extract text character with complex background and various character sizes.

References

1. Michael R. Lyu, Jiqiang Song, Min Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction", IEEE Trans. on Circuits and Systems for Video Technology. 15(2): 243-255, 2005.
2. L. Graham, Y. Chen, T. Kalyan, J. H. N. Tan and M. Li, "Comparison of Some Thresholding Algorithms for Text /Background Segmentation in Difficult Document Images", ICDAR, Vol 2, pp. 859-865, 2003.
3. S. Wu and A. Amin, "Automatic Thresholding of Gray-level Using Multi-stage Approach", ICDAR, pp. 493-497, 2003.
4. D. Chen, K. Shearer, and H. Bourlard, "Text Enhancement with Asymmetric Filter for Video OCR", ICIAP, pp.192–197, 2001.
5. C. Garcia and X. Apostolidis, "Text Detection and Segmentation in Complex Color Images", Proc. of ICASSP 2000, vol.4, pp. 2326-2330, 2000.
6. Y. Leydier, F. Le Bourgeois, and H. Emptoz, "Serialized Unsupervised Classifier for Adaptive Color Image Segmentation: Application to Digitized Ancient Manuscripts", Proc. of ICPR, pp. 494-497, 2004.
7. Fang Liu, Xiang Peng, Tianjiang Wang, Songfeng Lu, "A Density-based Approach for Text Extraction in Images", ICPR 2008, Page(s): 1-4.
8. Rafael C. Gonzalez, Richard E. Woods and Steven L. Eddins (2004). Digital Image Processing using MATLAB. Pearson Education. ISBN 978-81-7758-898-9.
9. Rongrong Wang, Wanjun Jin, Lide Wu, "A Novel Video Caption Detection Approach Using Multi-Frame Integration", ICPR 2004, Vol.1, pp: 449-452.
10. Boris Epshtein, Eyal Ofek, Yonatan Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform", CVPR 2010, pp: 2963-2970.
11. N. Otsu, "A Threshold Selection Method from Gray-level Histograms," IEEE Trans. Syst., Man, Cybernet., vol. SMC-9, no. 1, pp. 62–66, Jan.1979.