A Novel Method to detect the Static Subtitle using the Saliency Map

Min-Hoi Kim, Bo-Sang Kim, Seung Park, and Sung-Jea Ko, *Fellow*, IEEE
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
sjko@korea.ac.kr

***Abstract*—This paper presents a novel method to detect the static subtitle using the saliency map. In the proposed method, subtitle candidates are first extracted by using an unconstrained scene text detector algorithm, called FASText. Then, the saliency map is employed to select the final subtitle among the subtitle candidates. Experimental results show that the proposed method achieves more accurate subtitle detection performance as compared with the conventional methods.**

Keywords—Subtitle detection, Saliency map, Text detection, Pulse discrete cosine transform.

I. INTRODUCTION

A variety of subtitle detection methods have been proposed with applications in video analysis, indexing and retrieval [1]-[3].

Zafarifar *et al*. [3] proposed a method to detect the subtitles in video. In this method, the subtitle region is assumed as static regions where high horizontal gradients are densely located. However, since this method does not consider the form of stroke for subtitle detection, some non-subtitle regions with high horizontal gradients are regarded as the subtitle regions as shown in Fig. 1(a).

To alleviate this problem, we use the FASText algorithm [4] which utilizes the form of stroke for extracting the subtitle candidates. As shown in Fig. 1(b), the FASText algorithm detects the subtitle more accurately compared with the Zafarifar's method. Nevertheless, since some non-subtitle areas have the pattern similar to strokes, the FASText algorithm still yield some unexpected misdetection.

Therefore, based on the observation that the subtitle has a high saliency in an image, we utilize the saliency map to remove the false detection results obtained by the FASText algorithm.

This paper is organized as follows. First, we describe the pulsed discrete cosine transform (PCT) algorithm [5] for obtaining a saliency map. After then, the relationship between PCT and visual saliency of the subtitle is discussed. To refine the obtained saliency map, since the subtitles have the temporal consistency, an algorithm for accumulating the saliency values over frames is presented. Finally, we propose a method of utilizing the saliency map to classify the subtitle candidates detected by the FASText algorithm.

II. PROPOSED METHOD

We employed the FASText algorithm to obtain the subtitle candidates.

(a) (b)

Fig 1. The subtitle detection results obtained from (a) the Zafarifar's method and (b) the FASText algorithm.

A. Obtaining the saliency map

The PCT algorithm is utilized on the input image to obtain the saliency map. This algorithm is implemented by using the signum function after performing the discrete cosine transform (DCT), as follows:

$$P = \mathrm{sgn}(\hat{e}(M)) \tag{1}$$

where M denotes a gray level of the input image, and $\hat{e}$ and sgn(·) represent the DCT operator and signum function, respectively. Given the P, the result of applying PCT on the input image, the function to obtain the saliency map S is as follows:

$$S = \mathrm{abs}(\hat{e}^{-1}(P)) \tag{2}$$

where $\hat{e}^{-1}$ and abs(·) represent the inverse DCT operator and the absolute function, respectively.

Generally, the input image can be transformed into frequency domain by employing the DCT. Note that the large DCT coefficients contain the information of statistical homogeneity. Subsequently, since these DCT coefficients are flattened by applying the signum function, the DCT coefficients

(a) (b) (c)

Fig 2. Saliency map obtained by the PCT. (a) Input image, (b) Saliency map, (c) Accumulated saliency map.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

containing the information of image details are comparatively magnified. Thus, the subtitle region has high saliency values, as shown in Fig. 2(b).

B. Accumulated saliency map

Fig. 2(b) shows that the high saliency values are obtained not only in the subtitle region but also in the regions with the high frequency. To solve this problem, we accumulate the values in the saliency map over several frames, based on the observation that subtitles remain stationary for a period of time. Therefore, the accumulated saliency map up to the n-th frame S'_n can be obtained by

$$S'_n = (1-w)S_{n-1} + wS'_{n-1} \tag{3}$$

where S_n denotes the saliency map of the n-th frame, and w represents the weighting parameter. As shown in Fig. 2(c), by accumulating the saliency map, the saliency values in the non-subtitle regions are attenuated.

C. Refinement of the subtitle candidates

In order to effectively remove the non-subtitle regions among the subtitle candidates, we utilize the accumulated saliency map as follows: As shown in Fig. 2(c), since the subtile has higher saliency values than non-subtitle region, we binarize the accumulated saliency map S'_n. The binarized saliency map can be obtained by

$$B(x,y) = \begin{cases} 1 & S'(x,y) \geq T \\ 0 & S'(x,y) < T \end{cases} \tag{4}$$

where $B(x,y)$ denotes the pixel value of the binarized saliency map at position (x,y) and T represents the threshold value. Experimentally, we set the T as 0.2. In the binarized saliency map, a subtitle candidate, where the number of pixels with the value of 1 is more than 60% of the total pixels in that region, is defined as the subtitle region.

III. Experimental results and Conclusion

In this section, the performance of the proposed method is compared with that of the conventional algorithms [3],[4]. On a desktop computer with a 3.5GHz quad-core CPU and 8GB RAM, the experiments were conducted on 10 HD resolution videos including subtitles.

As shown in Fig. 3(b), the Zafarifar' algorithm [3] detects not only the subtitle region but also the non-subtitle regions where high horizontal gradients are densely located like the window frame. As shown in Fig. 3(c), the FASText algorithm achieves more accurate detection results than the Zafarifar' algorithm. However, some non-subtitle regions having patterns similar to strokes are classified as the subtitle region. On the other hand, Fig. 3(d) shows that the proposed method detects the subtitle accurately without any false detection. As a result, the experimental results show that the proposed method outperforms the conventional algorithms.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2017-0-00250, Intelligent Defense Boundary Surveillance Technology Using Collaborative Reinforced Learning of Embedded Edge Camera and Image Analysis).

Reference

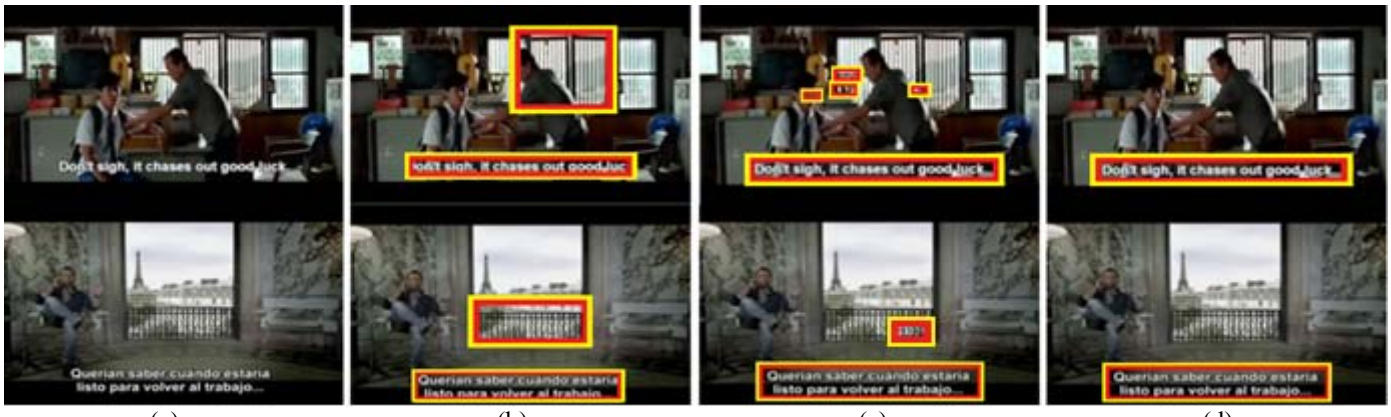
[1] E. Y. Kim, K. I. Kim, K. Jung, and H. J. Kim, "A video indexing system using character recognition," *IEEE Trans. on Consumer Electronics*, pp. 358-359, Aug. 2002.

[2] X. Qian, G. Liu, H. Wang, and R. Su, "Text detection, localization, and tracking in compressed video," *Signal Processing: Image Communication*, vol. 22, no. 9, pp. 752-768, 2007.

[3] Zafarifar, Bahman, J. Cao, and P. H. N. de With, "Instantaneously responsive subtitle localization and classification for TV applications," *IEEE Trans. on Consumer Electronics*, vol. 57, no. 1, 2011.

[4] M. Busta, L. Neumann, and J. Matas, "FASText: Efficient unconstrained scene text detector," *in Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1206–1214, Dec. 2015.

[5] Y. Yu, B. Wang, and L. M. Zhang, "Pulse discrete cosine transform for saliency-based visual attention," *in Proc. IEEE 8th Int. Conf. on Development and Learning*, pp.001-006, 2009.



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

Fig 3. Experiment results. (a) Input images. The subtitle detection results obtained from (b) the Zafarifar's method, (c) the FASText algorithm, and (d) the proposed algorithm (w=0.1).