Text Extraction in Video Images

Shwu-Huey Yen, Chun-Wei Wang, Jih-Pin Yeh*, Meng-Ju Lin, and Hwei-Jen Lin
Department of Computer Science and Information Engineering, Tamkang University
*Department of Information Management, Chinmin Institute of Technology**
shyen@cs.tku.edu.tw

Abstract

We propose a method to extract text information from video sequences. First the frequency of high horizontal energy in a video frame is examined to extract text blocks. Structural operations are then performed to remove the background so that the text can be extracted for later recognition. Experiments show that the method is efficient and effective for extracting text from various video documents.

1. Introduction

Text Information Extraction (TIE) has become an important task in many applications lately. A good TIE system can automatically add an annotation to the image and provide an image indexing mechanism with text information. A TIE system usually is composed of detection, localization, tracking, enhancement and recognition. However, in various image styles, the text may vary regarding font size, shape, orientation, as well as color variance, and these make it a challenging task.

Many methods for text localization have been proposed [1]-[4]. Most of these methods are only suitable for certain types of images. Our proposed method tends to provide a more general approach to the issue of text localization for a wider range of video images. First the horizontal energy of a sequence of video images is evaluated to extract candidate text blocks. The frequency of each of candidate text blocks keeping high energy in the sequence is examined to filter out the non-text blocks. Each of the reserved text blocks is then enhanced by background removal using the background image reconstruction and image subtraction. Finally, characters are extracted from the enhanced text blocks and then binarized for further usage.

2. Proposed Method

This section describes the proposed method, including text block localization in video images and background removal for text blocks.

2.1. Text Block Localization in Video Images

In [2], they divide each frame into 8x8 blocks and access the DCT coefficients from the MPEG files to detect textual information in videos. For each block, some selected horizontal DCT coefficients are accumulated to form its horizontal energy. As given in Eq. (1), $E_k(x, y)$ denotes the energy of block $B_k(x, y)$ in the kth frame of a video sequence and $C_{ij}(x, y)$ denotes the DCT coefficient in row i and column j of the block. Next, blocks containing high horizontal energy, i.e., $E_k(x, y)$ exceeds a given threshold T, are marked as candidate text blocks, as shown in Eq. (3). In this work, $T = 1.45 \times \overline{E}_k$, where $\overline{E}_k$ is the average energy as defined in Eq. (2), b is the number of blocks.

$$E_k(x,y) = \sum_{2\le r\le 6} \left|C_{0r}(x,y)\right| \tag{1}$$

$$\overline{E}_k = \frac{1}{b}\sum_{B_k(x,y)} E_k(x,y) \tag{2}$$

$$TB_k(x,y) = \begin{cases} 1 & \text{if } E_k(x,y) \ge T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

However, due to the complex background, there may be some non-text blocks remaining. To remove these non-text blocks, we calculate the horizontal energy of each block on every 5 frames in a video sequence, and count the number of occurrences of high energy blocks. As the temporal information of text region mentioned above, the more occurrences of a high energy block, the higher probability that it is a text block. As in Eq. (4), a high energy block is kept if its occurrences is more than $\alpha \times m$ in frames 1, 6, …, $5(m-1)+1$ (α = 0.9 and m = 5 in our experiments). Followed by operations of morphological closing,

978-0-7695-3266-0/08 $25.00 © 2008 IEEE
DOI 10.1109/SSIRI.2008.26

IEEE computer society

opening, and dilation with a 1×5 structuring element, most of non-text blocks are further eliminated. Finally, we apply connected component to form more compact rectangular text regions for later process. Fig. 1(a) shows an experimental result for this stage.

$$TB(x,y)=\begin{cases}\text{true} & \text{if } \sum_{0\le k\le m-1} TB_{1+5k}(x,y)\ge \alpha\, m \\ \text{false} & \text{otherwise}\end{cases} \qquad (4)$$

2.2. Background Removal for Text Blocks

To extract text, we remove the background in each text block by subtracting the reconstructed background [5]. As shown in Fig. 1, (a) is the image f containing text blocks, (b) R_f is the reconstructed background of (a) which uses the pixels connecting to block borders as the seeds and f as the mask, (c) shows the result of background removal by f-R_f. However, when part of text is connected to the border, it will be removed too. To do this, we extend the text blocks by four pixels in each side to obtain a larger block g and follow the same steps. Fig. 1(d) shows the result of the improved background removal. As can be seen, texts on the right (in vertical) and on the bottom (in horizontal) become clearer. Binarization [6] is then performed to obtain characters for later OCR process as shown in Fig. 2.

3. Experimental Results

In this section, we compare our proposed method with the method proposed by R. Wang et al.. Because it selects 30 consecutive frames for text extraction, their method requires more computational time. In the experiments, our proposed method and the method of R. Wang et al. take 0.7s and 1.5s, respectively, on average. In addition, the results of R. Wang et al. are poor for images with a complex background. Even if the scan window is enlarged to a size of 20×10, there are still a lot of non-text blocks reserved with their method. The results of both methods are compared in Fig. 2.

4. Conclusions

In our proposed method, the DCT coefficients and temporal information of the video sequence are used to evaluate the horizontal energy, with which most of the non-text blocks can be filtered out. Some structural operations are performed to further remove the non-text blocks. For each detected text block, the background is removed by a reconstruction procedure. The text blocks are then further enhanced to enable each text instance for OCR. Compared to the method of R. Wang et al., our proposed method showed better results for rational text detection.

References

[1] M. Y. H. Yassin and L. J. Karam, ‘‘Morphological text extraction from images’’, *IEEE Transactions on Image Processing, November*, 2000, Vol. 9, No. 11, pp. 1978-1983.

[2] Y. Zhong, H. Zhang and A. K. Jain, ‘’Automatic caption localization in compressed video’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 22, No. 4, pp. 385-392.

[3] R. Wang, W. Jin and L. Wu, ‘‘A novel video caption detection approach using multiframe integration’’, *Pattern Recognition 17th International Conference on ICPR'04*, 2004, Vol. 1, pp. 449-452.

[4] D. Zhang, B. L. Tseng, C.-Y. Lin and S.-F. Chang, ‘‘Accurate Overlay Text Extraction for Digital Video Analysis’’, *Proc. of IEEE Int. Conf. on Information Technology: Research and Education, Newark*, 2003, pp. 233-237.

[5] Pierre Soille, ‘‘*Morphological image analysis: principles and applications*’’, Springer-Verlag, 1999, pp.163-164.

[6] H. J. Lin and F. W. Yang, ‘‘An intuitive threshold selection based on mountain clustering’’, *First International Workshop on Intelligent Multimedia Computing and Networking (IMMCN2000).*

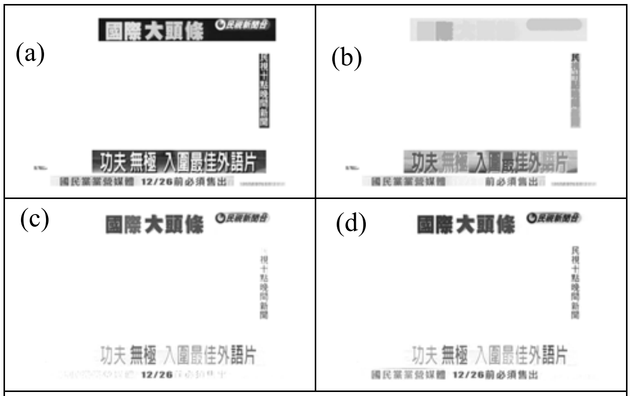

Figure 1 (a) Image f with text blocks, (b) R_f: result of the background reconstruction from f, (c) f-R_f: result of background removal, (d) g-R_g: improved result of background removal where g is the extended region of f.

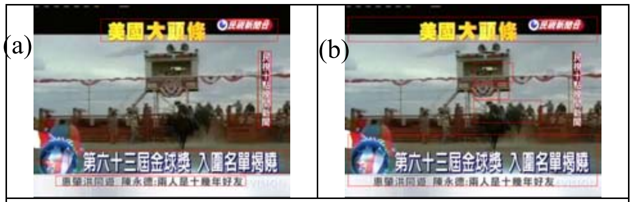

Figure 2 (a) Our method, (b) Method of R. Wang et al.