비디오에서 문자 검출을 위한 강인한 방법

Viet-Cuong Dinh, 전승수, 류한진, 설상훈
Department of Electronic and Computer Engineering, Korea University
{cuongdv, sschun, hanjin, sull}@mpeg.korea.ac.kr

A Robust Method for Text Detection in Video

Viet-Cuong Dinh, 전승수, 류한진, 설상훈
Department of Electronic and Computer Engineering, Korea University

Abstract

This paper proposes an effective method for text detection in video. First, we apply an edge detection method to the video frame with a relative low threshold to keep all possible text edge pixels. Second, a multi-frame integration method is applied to significantly remove background pixels which are not stationary in a specific period. Finally, text regions are extracted by using the coarse to fine projection method. Experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

With the growing number of digital multimedia libraries, the need to efficiently index, browse and retrieve multimedia information is increased. Extracting descriptive features and higher level entities, for example text [1] or human faces [2], has attracted more and more research interest recently. Text embedded in images and video, especially caption, provide brief and important content information, such as the name of players or speakers, the title, location and date of an event, etc. The text can be considered as a powerful feature (keyword) resource as are the information provided by speech recognizers for example.

Many efforts have been made for text detection in images and videos. Based on the way being used to locate text regions, text detection methods can be classified into two main approaches: connected component (CC) based methods [3], [4] and texture-based methods [5], [6]. The first approach is based on the analysis of geometrical arrangement of edge or homogeneous color that belong to character. Jan and Yu *et al.* [3] first applies multi-value image decomposition algorithm to decompose the input image into multiple foreground and background partitions. Then, text candidates are localized by performing CC analysis on each partition. However, this method is unsuitable for videos whose texts are embedded in complex background scenes with different colors since it can only extract the horizontal text of large size. The second one, texture-based method treats text region as a special type of texture. In this approach, we can use many methods to calculate the texture features, e.g. the Gabor filter, wavelet transform and edge detection. Julinda Gllavata *et al.* [6] proposed a wavelet transformation to detect text regions by claiming that DCT coefficients for text areas are dispersed and concentrated on a few discrete values. Therefore a clustering method can be applied to locate text regions.

However, most existing methods can not handle well with the background complexity trouble which varies from video sequence to video sequence and also from frame to frame in each sequence.

In this paper, we present an efficient method for text detection in video by sufficiently utilizing multiple frames to overcome the background complexity problem. Since text appearing in video is often stationary or linear moving for at least two seconds, our multiple frame integration method could remove the non-text pixels which do not last for a specific duration.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed method for text detection in video. To demonstrate the effectiveness of our method, experimental results are shown in Section 3. Finally, Section 4 is our conclusion remarks.

2. Proposed Method

In the proposed method, text regions in frame are detected by using three main steps. First, we apply an edge detection method with a relative low threshold to keep all possible text edge pixels. Second, multi-frame integration method is applied to remove the non-stationary pixels during a specific time. Finally, text regions are extracted by using the coarse-to-fine projection method. These steps are presented as follow.

2.1 Edge detection method

There are several edge detection methods producing different types of edge, such as Sobel edge detector, Canny edge detector and Laplacian of Gaussian Edge Detection. When connecting the edge in frames, only the thin edges with one pixel could be connected well. In our evaluation, Canny edge detector is easy to achieve consistency due to its thin characteristic. The comparison of Canny edge detector and Sobel edge detector is shown in Figure 1. In Figure 1.b, the edge obtained from Canny edge detector represents the actual detail edges. The edges obtained from Sobel edge detector in Figure 1.a is double hardly to connect. Thus, the Canny edge detector is applied in our proposed video detection method. The threshold for the edge detection method is set relative low to keep all possible text edge pixels remained.

(a) (b)

Figure 1. Comparison of edge detection method. (a) Sobel method. (b) Canny method.

2.2 Multi-frame Integration

The basic idea of multi-frame integration method is based on the fact that text appearing in video is often stationary for at least 2 or 3 second. Previous works, such as [7][8], utilized multi-frame for the purpose of verifying whether a text region candidate (which has localized in previous steps) is a true text region or not. The text region candidate is consider as true if the location of text region does not much change during a specific time. These methods work well when the text is embedded in clear background. However, when embedded in complex background, the text location is often imprecisely localized. This results to increasing the number of fault alarms of the text detection method. In this paper, we directly apply the multi-frame integration method to the whole video frame, which has passed the edge detection process. By applying directly the multi-frame integration to the whole video frame, we can overcome the incorrect text location caused by the background complexity problem mentioned above.

The proposed multi-frame integration method is presented as follow.

Call F_i is the frame at the time i after applying the edge detection method. For each of 3 consecutive video frame F_i, F_{i+1}, F_{i+2}, a combined frame F_i' is formed as follow:

$$F_i' = F_i \cap F_{i+1} \cap F_{i+2} \quad (1)$$

F_i' is formed by doing **and** operator with the 3 consecutive frames above. The pixel at location (x, y) of F_i' is only edge if pixel values at that location of F_i, F_{i+1}, F_{i+2} are also edges.

Then, we analysis the sequence of $3*n$ consecutive frames F_i, F_{i+1},..., F_{i+3n-1} . Corresponding, we have n combined frame F_i' , F_{i+1}' , ..., F_{i+n-1}' as shown in Figure 2.

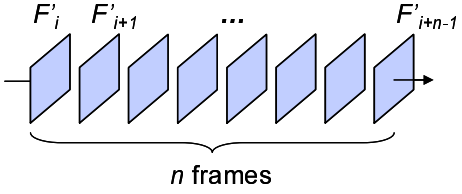

Figure 2. Multi-frame integration

From the sequence of F_i', we form the *stationary edge image* F which represents for the stationary property of pixels during a number of frames. The pixel at location (x, y) of F is determined as follows:

$$F(x,y) = \begin{cases} edge\ pixel, & if \sum_{k=0}^{n-1} P_i(x,y) > \theta \\ non\ edge\ pixel, & otherwise \end{cases} \quad (2)$$

where θ is a given threshold and $P_i(x, y)$ is:

$$P_i(x,y) = \begin{cases} 1, & if\ F_i'(x,y)\ is\ edge\ pixel \\ 0, & otherwise \end{cases} \quad (3)$$

The reason to apply and operator with 3 consecutive frames (formula (1)) is to reduce the effect of random noise which may occur in the video sequence. By doing this process, we can assure the correctness of formula (2).

Refer to (3), $F(x, y)$ is an edge pixel if at location (x, y), an edge pixel appears more than θ times, otherwise, $F(x, y)$ is a non-edge pixel.

By using our constraint on the stationary property, almost text edge pixels are still retained while background pixels are removed in F.

In order to recovery the missing text-edge pixels while still ignoring background edge pixels, we create a rule as follow:

A pixel is a text edge pixel in the text region candidate if and only if it satisfies two criteria:

1) It must be an edge pixel of at least one video frame: F_i' , F_{i+1}' , ..., F_{i+n-1}'.
2) It must be an edge pixel or have at lease one neighbor which is an edge pixel in image F.

Consequently, after applying this rule, F contains almost all text-edge pixels. The effectiveness of this improvement is shown in Figure 3. As shown in the figure, text pixels are still retained while almost background pixels are removed.

(a) (b)

Figure 3. Result of edge image after applying the multi-frame integration.

2.3 Text localization

After the multi-frame integration process, we employ a text localization process to localized text regions. In each created edge image, the high-density area indicates the text region. With the assumption that text orientation is horizontal, the combined horizontal and vertical projection method is an effective way to locate text strings. However, for some videos with complex text layouts, this method may not perform well since text strings can overlap the others horizontally or vertically. To overcome this problem, we employ a coarse-to-fine projection method which composes of several phases of projecting. In each phase, we do in turn horizontal and vertical projection. At the beginning phase, the whole frame is processed to coarsely separate text regions which may contain multiple overlapped text strings. Then, each text region is pushed to the next projection phase to be divided into softer sub-regions. This process is repeated until all text regions cannot be divided anymore. Then each region mostly contains only one text string. The number of edge pixels in each row and each column are used as the criteria to determine whether a region should be further partitioned. The effective of the coarse-to-fine projection method is shown in Figure 4 at which all of text regions are correctly detected.

(a) (b)

Figure 4. Coarse to fine projection. (a) Edge image after multi-frame integration. (b) Text detected regions.

3. Experimental Results

To evaluate the effectiveness of the proposed method, we have collected a number of videos containing newscast, live sports, and documents for a test database. The video frame format is 512×384 and 720×480 pixels. Totally, we have 170 distinct frames containing 450 text regions and the experimental results on this database are presented as follows.

For a quantitative evaluation, the detected text region is considered as the correct one if the intersection of the detected text region (DTR) and the ground-truth text region (GTR) covers more than 90% of this DTR and 90% of this GTR. The total number of GTRs used in our experiments is 450 and all of them are manually localized. The efficiency of our detection algorithm is assessed in terms of two measurements: *Detection Rate* and *Detection Accuracy*.

Detection Rate shows how many percents of all

ground-truth text regions are correctly detected, that is given by:

$$DetectionRate = \frac{NumberofCorrect(DTRs)}{Number(GTRs)} \quad (4)$$

Detection Accuracy represents how many percents of the detected text regions are correct:

$$DetectionAccuracy = \frac{NumberofCorrect(DTRs)}{NumberofAll(DTRs)} \quad (5)$$

In order to evaluate the effectiveness of the proposed method, we compare the performance with the typical edge-based method proposed in [8].

Table 1. Comparison of detection rate and accuracy.

	Correct DTRs	False DTRs	Detection Rate (%)	Detection Accuracy (%)
Cai *et al.* [9]	374	79	83.1	82.5
Proposed Method	421	37	**93.5**	**91.9**

Table 1 shows the experimental results. These results show that the proposed method achieves the highest accuracy compared to the other in term of both detection rate (93.5%) and detection accuracy (91.9%). By using the local adaptive threshold algorithm, our method can handle well the background complexity problem resulted from complex videos of many sources. Moreover, the time process for a frame of dimension 512×384 is in 0.21s, and 720×480 in 0.35s, respectively. It is fast enough to be used in real-time applications.

Figure 5 shows some examples of text detection results.

4. Conclusion

This paper presents a comprehensive method for text detection in video. To overcome the background complexity problem, we incorporate the edge detection method and the multi-frame integration method based on the observation that text appears in video often stationary for at least a number of frames. Not only the proposed method can reduce the affect of noise, but also it is invariant to the difference in size and language of text. Experimental results with a large set of videos demonstrate the efficiency of our method with the detection rate of 93.5% and

Figure 5. Examples of text detection results.

detection accuracy of 91.9%. In the future, we plan to continue researching about text tracking and recognition for real time text-based video indexing and retrieval system.

References

[1] D. Chen, H. Bourlard, J.-P. Thiran, "Text identification in compex background using SVM", *Int. Conf. on Computer Vision and Pattern Recognition*,2001, pp. 621-626.

[2] R.K. Srihari, Z. Zhang, A. Rao, Intelligent indexing and semantic retrieval of multimodal documents, Inform. Retrieval 2, pp. 245-275.

[3] A. K. Jain and B.Yu, "Automatic text location in images and video frames," *Pattern Recognition*, vol. 31, pp. 2055–2076, 1998.

[4] Y. Zhong, K. Karu, and A.K. Jain, "Locating text in complex color images," *Pattern Recognition,* vol. 28, pp. 1523–1535, Oct. 1995.

[5] Y. Liu, H. Lu, X. Xue, and Y.-P. Tan, "Effective video text detection using line features," *Proc. of Int. Conference on Control, Automation, Robotics, and Vision*, vol. 1, pp. 1528–1532, 2004.

[6] J. Gllavata, R. Ewerth, and B. Freisleben, "A text detection, localization and segmentation system for OCR in images," *Proc. of Int. Symposium on Multimedia Software Engineering*, pp. 310–317, Dec. 2004.

[7] H. Li et al., "Automatic text detection and tracking in digital video," IEEE Transactions on Image Processing, v.9, pp.147-156.

[8] Lim et al., "Text extraction in MPEG compressed video for content-based indexing," International Conferences on Pattern Recognition 2000.

[9] M.Cai, J. Song, and M. R. Lyu, "A New Approach for Video Text Detection," *Proc. of Int. Conference on Image Processing,* vol. 1, pp. 117–220, 2002.