A Spatio-temporal Approach for Video Caption Extraction

Liang-Hua Chen[1], Meng-Chen Hsieh[1] and Chih-Wen Su[2]
[1]*Department of Computer Science and Information Engineering, Fu Jen University, New Taipei, Taiwan*
[2]*Department of Information and Computer Engineering, Chung Yuan University, Chung Li, Taiwan*

Keywords: Video Content Analysis, Caption Detection, Spatio-temporal Slices.

Abstract: Captions in videos play an important role for video indexing and retrieval. In this paper, we propose a novel algorithm to extract multilingual captions from video. Our approach is based on the analysis of spatio-temporal slices of video. If the horizontal (or vertical) scan line contains some pixels of caption region then the corresponding spatio-temporal slice will have bar-code like patterns. By integrating the structure information of bar-code like patterns in horizontal and vertical slices, the spatial and temporal positions of video captions can be located accurately. Experimental results show that the proposed algorithm is effective and outperforms some existing techniques.

1 INTRODUCTION

The advances in low cost mass storage devices, higher transmission rates and improved compression techniques, have led to the widespread use and availability of digital video. Video data offers users of multimedia systems a wealth of information and also serves as a data source in many applications including digital libraries, publishing, entertainment, broadcasting and education. The usefulness of these applications depends largely on whether the video of interest can be retrieved accurately within a reasonable amount of time. Therefore, various video content analysis techniques have been proposed to index or retrieve the large amounts of video data. These techniques are based on the analysis of visual, audio and textual information in the videos. Among them, text can provide concise and direct description of video content. If textual part of the video can be extracted and recognized automatically, it will be a valuable source of high-level semantics for indexing and retrieval. On the other hand, the current optical character recognition (OCR) techniques are more robust than speech recognition and visual object recognition techniques. This facilitates the automatic annotation of video content using the extracted text.

In general, there are two types of text in video: scene text and caption text. The scene text is an integral part of the scene captured by the camera such as signpost, billboard, banner and so on. The caption text is artificially embedded in video frame during video editing. Usually, it is closely related to the content of the video. For example, the captions in newscast video summarize relevant names, locations and times of the reported events. In comparison to text segmentation for document image, the problem of extracting caption from video is more difficult. Some factors include (1) complex background, (2) lower resolution (small size) of text, (3) unknown text color and (4) degraded image quality caused by lossy compression method. Therefore, the quality of text in video is not suitable to be processed by conventional document image analysis technique.

Current techniques for video caption detection can be broadly classified into three categories (Tang et al., 2002). The first one is the connected component based methods (Mariano and Kasturi, 2000; Ye et al., 2005; Liu et al., 2010; Gonzalez et al., 2012). This category assumes that text regions have uniform colors and satisfy certain size, shape and spatial alignment. However, these methods are not effective when text touches other graphical objects or text is embedded in complex background. The second category treats text as a type of texture (Li et al., 2000; Zhong et al., 2000; Kim et al., 2001; Wang and Chen, 2006; Qian et al., 2007; Pan et al., 2011). Thus, the classic texture classification algorithms are applied to detect text regions. These methods are computationally expensive. Besides, it is hard to find accurate boundaries of text regions and false alarms often exist if the background

Chen, L-H., Hsieh, M-C. and Su, C-W.
A Spatio-temporal Approach for Video Caption Extraction.
DOI: 10.5220/0005939300830088
In *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications (ICETE 2016) - Volume 5: SIGMAP*, pages 83-88
ISBN: 978-989-758-196-0

contains texture that display the similar structure as text regions. The third category consists of edge based methods which assume text regions have high contrast against the background (Hua et al., 2001; Lienhart and Wernicke, 2002; Chen et al., 2003; Wang et al., 2004; Lyu et al., 2005; Tasi et al., 2006; Shivakumara et al., 2008; Gui et al., 2012; Huang et al., 2014). Therefore, those areas with dense edges are detected as text regions. These methods are less reliable for the detection of text with large font size. In most of the above works, text in video is treated mainly the same way as that in still image. Although some works use the short duration of temporal information (Tang et al., 2002; Wang et al., 2004; Wang and Chen, 2006), they do not fully exploit the spatio-temporal information in video. To utilize the long duration of temporal information, in this paper, we propose a caption extraction algorithm based on the analysis of spatio-temporal slices of video.

2 BACKGROUND AND MOTIVATION

A spatio-temporal slice is a collection of scans in the same selected position of every frame of a video as a function of time (Ngo et al., 2001). Two common selection methods of scans are horizontal scan and vertical scan. Assuming the content of a video is represented by $f(x,y,t)$, a horizontal spatio-temporal slice is defined as

$$X-T(x,t)=f(x,y',t)$$

where y' is a constant, and a vertical spatio-temporal slice is defined as

$$Y-T(t,y)=f(x',y,t)$$

where x' is a constant.

Our approach to caption detection is based on the following observation: If the horizontal (or vertical) scan line contains some pixels of caption region then the corresponding spatio-temporal slice will have bar-code like patterns. Figure 1 illustrates this phenomenon where X-T slice and Y-T slice has bar-code like patterns in vertical direction and horizontal direction respectively. In each slice, there are several groups of patterns corresponding to different text captions. It is also noted that the structure of bar-code like patterns indicates the spatio-temporal information of caption as follows.

- The length of a bar-code like pattern corresponds to the duration time of a caption.
- In X-T slice, the length (in x-axis direction) of each group of patterns corresponds to the width of a caption.
- In Y-T slice, the length (in y-axis direction) of each group of patterns corresponds to the height of a caption.

Therefore, we may integrate the information in X-T and Y-T slices to locate the captions in video. The proposed algorithm is made up of two components: caption localization and character extraction. Each is described in the following two sections.

3 CAPTION LOCALIZATION

Since our caption localization algorithm is based on the analysis of spatio-temporal slices, the main issues are: (1) How to determine whether a spatiotemporal slice contains bar-code like patterns or not and (2) How to integrate the information in X-T and Y-T slices to locate video caption. These two issues are addressed as follows.

For a given X-T slice, we detect vertical edges by Sobel operator G_y and apply Hough transform to detect straight lines. If 90% of the detected lines are vertical (the orientation of line is between 85^0 and 95^0), then this slice is identified as the one containing bar-code patterns. In the similar way, we may use Sobel operator G_x and Hough transform to determine whether a Y-T slice contains bar-code like patterns or not.

$$G_x=\begin{vmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{vmatrix} \quad G_y=\begin{vmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{vmatrix}$$

Given a sequence of video frames, we sample horizontal spatio-temporal (X-T) slices every 5 pixels along the y-axis (from bottom to top) until detecting the one that contains bar-code like patterns. A line tracing process is performed on the edge image of this slice. Any line segment with length less than a threshold is identified as noise and removed from the edge image. Figure 2 shows the filtered edge image. Meanwhile the starting point and ending point of each line segment are recorded. Then all starting points are clustered based on their Y coordinates. Each ending point is assigned to the same cluster as its corresponding starting point. Figure 3 shows the clustering result of all starting and ending points. By analyzing the points of the same cluster, the width of its corresponding caption can be estimated. Likewise, we sample vertical spatio-temporal (Y-T) slices every 5 pixels along the x-axis (from left to right) until detecting the one that contains bar-code like patterns. Using the same

technique, the height of a caption can also be estimated.

The objective of caption localization is to determine a bounding box of each text caption. Assuming the X-T slice (sampled at $Y = y'$) estimates the width of a caption is W and the Y-T slice (sampled at $X = x'$) estimates the height of a caption is H, the lower left corner of bounding box is $(x'-\varepsilon, y'-\varepsilon)$ and the upper right corner of bounding box is $(x'+W, y'+H)$. Currently, ε is set to be 7.

While the bounding box may contain multiple text strings, a refinement process is needed. Sobel edge detector is applied to the original image frame inside the bounding box to obtain the contour map of characters. Then, we accumulate the horizontal projection of contour map. The valley of projection profile indicates the position of separate line. Finally, the vertical projection profiles of each separated region are also constructed to get more accurate bounding box. Figure 4 illustrates this process.

4 CHARACTER EXTRACTION

The goal of character extraction is to convert the color image inside the bounding box into the OCR-ready binary image, where all pixels of the characters are in black and others are in white. Our approach is based on the observation that characters embedded in video are mostly uniform in color. Using the k-mean clustering algorithm, all the pixels are classified based on their RGB values. In this step, we get binary images B_i, $i = 1,\cdots,k$ which indicates the location of each cluster. Then a certain binary image is selected as target image by integrating character contour information. A distance transform is applied to the edge image (resulting from Sobel edge detector) to get a distance map D. The value of distance map at each pixel is the distance from that pixel to the nearest character contour. Let

$$V_i = \frac{1}{N_i}\sum\sum B_i(x,y) \times D(x,y)$$

where N_i is the number of pixels belonging to cluster i. If $V_j = \min\{V_1,\cdots,V_k\}$, then binary image B_j is the target image for OCR.

5 EXPERIMENTAL RESULTS

The proposed algorithm is tested by five video sequences whose durations vary from 4 minutes to 30 minutes. The resolutions of video frames are 480× 360 and 640 ×360. The video types include movies, cartoons and newscasts. The languages are multilingual, i.e., the character set in the videos involves English, French, Italian, Chinese and Japanese. The total number of captions is 1596. Depending on the content of video, the duration time of each caption may last for 2-12 seconds. Figure 5 and 6 show some experimental results of caption detection. However, there are some false detections and missed detections. The missed detections are due to two factors. One is that the colors of caption and background are very similar. The other is that the filtering process on the edge image degrades the structure of bar-code like pattern. The false detections are mainly caused by some high contrast objects which remain in the same location of video frames for a long duration. Finally, some experimental results of character extraction are shown in Figure 7.

The performance of caption detection is usually measured by the following two metrics:

$$\text{Recall} = \frac{D}{D + MD} \quad \text{Precision} = \frac{D}{D + FD}$$

where D is the number of captions detected correctly, MD is the number of missed detection and FD is the number of false detection. For performance comparison, we also implement the connected components based algorithm (Gonzalez et al., 2012), texture based algorithm (Pan et al., 2011) and edge based algorithm (Huang et al., 2014). To compare four approaches fairly, the parameters of each algorithm are tuned to achieve the best performance. As shown in Table 1, our approach is, in overall, better than these conventional approaches in term of recall and precision.

6 CONCLUSIONS

We have presented a novel algorithm to extract captions from video. Our approach is based on the analysis of spatio-temporal slices of video. Unlike previous approaches which use the short duration of temporal information, our approach fully exploits the spatiotemporal information in video. Therefore, the location and duration of captions can be estimated accurately. Another advantage is that our algorithm is multilingual, i.e., it does not depend on the appearance and font of characters. Compared with related works, our approach is simple yet effective. Finally, our future work should also be

directed towards the detection of special effect captions which are moving texts or nonhorizontally aligned texts.

REFERENCES

Chen, D., Odobez, J., and Bourlard, H. (2003).Text detection and recognition in images and video frames.*Pattern Recognition*, 37:595–607.

Gonzalez, A., Bergasa, L., Yebes, J., and Bronte, S. (2012). Text location in complex images. In *Proceedings of International Conference on Pattern Recognition*, pages 617–620.

Gui, T., Sun, J., Naoi, S., and Katsuyama, Y. (2012). A fast caption detection method for low quality video images. In *Proceedings of IAPR International Workshop on Document Analysis Systems*, pages 302–306.

Hua, X., Chen, X., Liu, W., and Zhang, H. (2001).Automatic location of text in video frame. In *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 24–27.

Huang, H., Shi, P., and Yang, L. (2014).A method of caption location and segmentation in news video. In *Proceedings of International Congress on Image and Signal Processing*, pages 365–369.

Kim, K., Jung, K., Park, S., and Kim, H. (2001). Support vector machine-based text detection in digital video. *Pattern Recognition*, 34(2):527–529.

Li, H., Doermann, D., and Kia, O. (2000). Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing*, 9(1):147–156.

Lienhart, R. and Wernicke, A. (2002). Localizing and segmenting text in images and videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(4):256–257.

Liu, X., Wang, W., and Zhu, T. (2010). Extracting captions in complex background from video. In *Proceedings of International Conference on Pattern Recognition* , pages 3232–3235.

Lyu, M., Song, J., and Gai, M. (2005).A comprehensive method for multilingual video text detection, localization, and extraction.*IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):243–255.

Mariano, V. and Kasturi, R. (2000).Locating uniform colored text in video frames. In *Proceedings of International Conference on Pattern Recognition*, pages 539–542.

Ngo, C., Pong, T., and Chin, R. (2001). Video partitioning by temporal slice coherency.*IEEE Transactions on Circuits and Systems for Video Technology*, 11(8):941–953.

Pan, Y., Hou, X., and Liu, C. (2011).A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing*, 20(3):800–813.

Qian, X., Liu, G., Wang, H., and Su, R. (2007). Text detection, localization, and tracking in compressed video. *Signal Processing: Image Communication*, 22:752– 768.

Shiva kumara, P., Huang, W., and Tan, C. (2008). Efficient video text detection using edge features. In *Proceedings of International Conference on Pattern Recognition*.

Tang, X., Gao, X., Liu, J., and Zhang, H. (2002). A spatial temporal approach for video caption detection and recognition.*IEEE Transactions on Neural Network*, 13(4):961–971.

Tasi, T., Chen, Y., and Fang, C. (2006).A comprehensive motion video text detection localization and extraction method. In *Proceedings of International Conference on Communications, Circuits and Systems*, pages 515–519.

Wang, R., Jin, W., and Wu, L. (2004). A novel video caption detection approach using multi-frame integration. In *Proceedings of International Conference on Pattern Recognition*, pages 449–452.

Wang, Y. and Chen, J. (2006). Detecting video text using spatio-temporal wavelet transform. In *Proceedings of International Conference on Pattern Recognition*, pages 754–757.

Ye, Q., Gao, Q. H. W., and Zhao, D. (2005). Fast and robust text detection in images and video frames. *Image and Vision Computing*, 23:565–576.

Zhong, Y., Zhang, H., and Jain, A. (2000).Automatic caption localization in compressed video.*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):385–392.

APPENDIX

Table 1: Performance comparison for caption detection.

Test Video ID	The Proposed Approach		Component Based Approach	
	Recall	Precision	Recall	Precision
(1)	93.07%	89.56%	86.35%	81.43%
(2)	94.92%	91.30%	88.53%	84.10%
(3)	88.24%	78.94%	83.89%	70.67%
(4)	90.44%	87.88%	83.07%	77.69%
(5)	91.38%	85.74%	87.42%	80.57%
Test Video ID	Texture Based Approach		Edge Based Approach	
	Recall	Precision	Recall	Precision
(1)	88.42%	84.71%	91.43%	86.36%
(2)	90.77%	88.42%	92.23%	87.10%
(3)	85.36%	75.31%	82.35%	77.06%
(4)	86.21%	81.35%	88.07%	83.64%
(5)	83.46%	77.74%	89.57%	83.33%

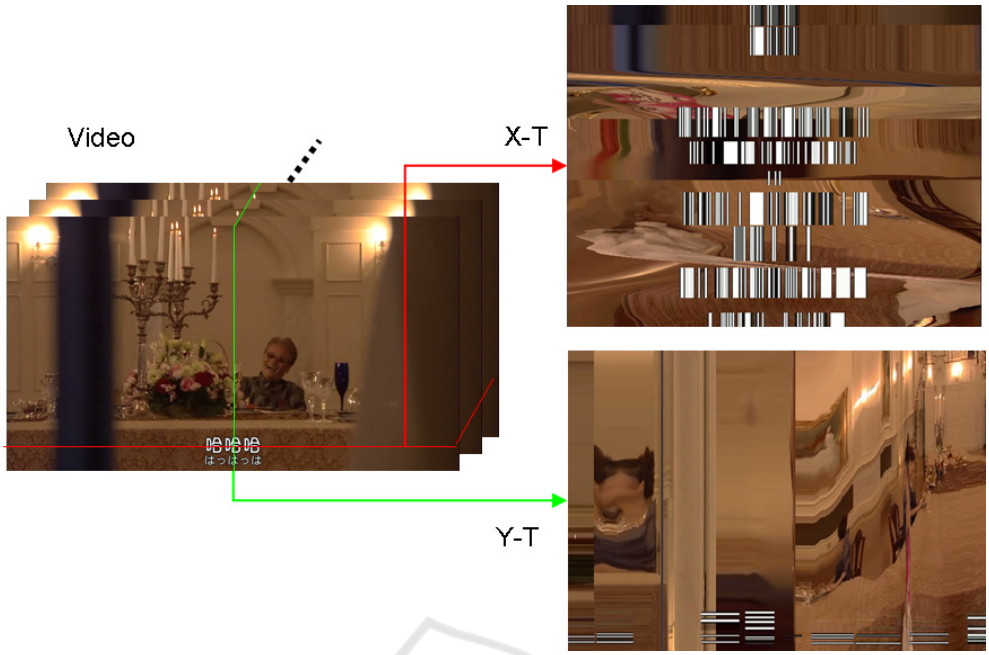

Figure 1: The horizonal and vertical slices containing captions.

Figure 2: The filtered edge map.

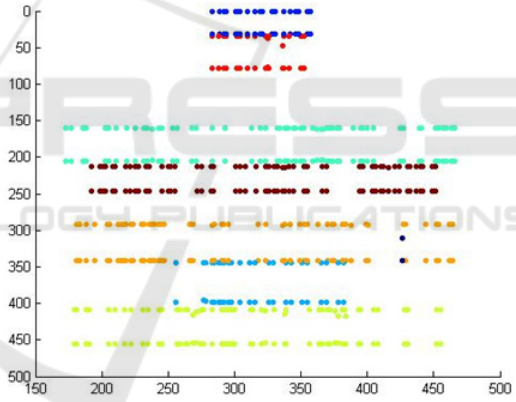

Figure 3: The clustering result.

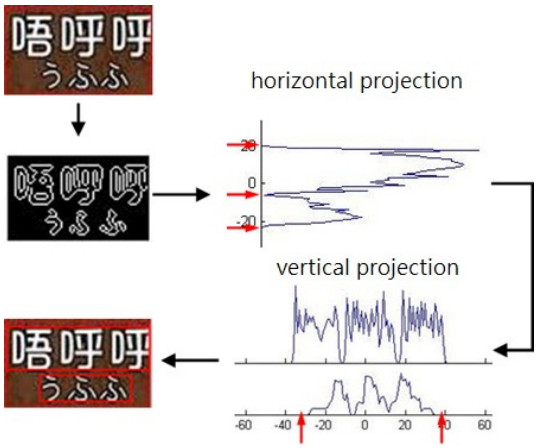

Figure 4: The process of partitioning caption that contains multiple text lines.

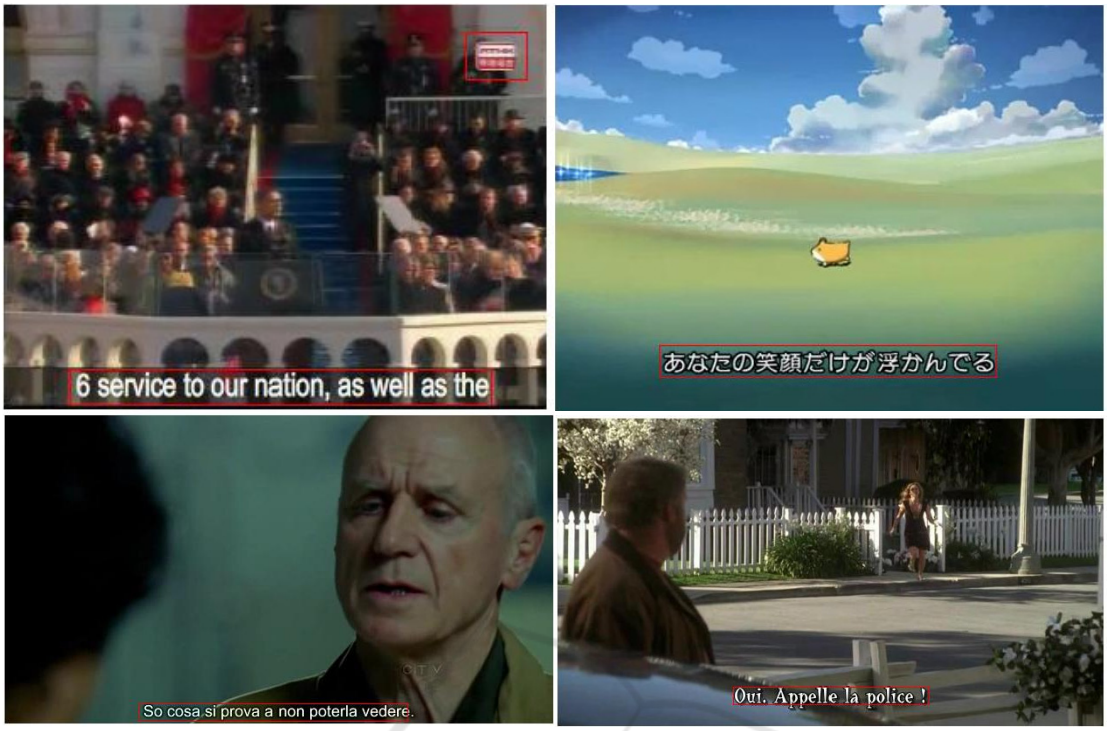

Figure 5: Caption localization for different languages.

Figure 6: Caption localization for multiple text lines.

Figure 7: Some experimental results of character extraction.