VIDEO TEXT EXTRACTION USING TEMPORAL FEATURE VECTORS

Xiaoou Tang[1], Bo Luo[1], Xinbo Gao[1], Edwige Pissaloux[2], Hongjiang Zhang[3]

[1]Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
Xtang@ie.cuhk.edu.hk

[2]Paris Robotics Laboratory
10-12 Av. de L'Europe
78140 Velizy-Villacoublay, France

[3]Microsoft Research Asia
49 Zhichun Road, Beijing 100080, China

ABSTRACT

A new caption text extraction algorithm that takes full advantage of the temporal information in a video sequence is developed. By detecting the (dis)appearance of caption text in a video stream, we first identify video segment that contains the same caption text. Then using the gray-level vector traced across the segment as the feature vector for a pixel point, we can clearly separate a caption pixel from a background pixel for the entire segment.

1. INTRODUCTION

Due to the rich information contained in caption text, video-caption based methods are increasingly used for efficient video content indexing and retrieval in recent years. Caption text routinely provides such valuable indexing information as scene locations, speaker names, program titles, sports scores, dates and time. Compared with other video features, information in caption text is highly compact and structured, thus is more suitable for video indexing.

However, extracting captions embedded in video frames is a difficult task. In comparison to OCR for document images, caption extraction and recognition in videos involves several new challenges [10]. First, captions in videos are often embedded in complex backgrounds, making caption detection much more difficult. Second, characters in captions tend to have a very low resolution since they are usually made small to avoid obstructing scene objects in a video frame. Therefore, the character quality in a video frame is too low to be processed by a conventional OCR system directly. In addition, popular lossy compression methods such as MPEG often lower the image quality even further.

In order to overcome these difficulties, new text detection and extraction methods have been developed recently. They are generally grouped into three categories -- connected component methods [5][6][8][15], texture classification methods [7][14], and edge detection methods [1][2][4][10][12]. The connected component methods detect text by extracting the connected components of monotonous color that obey certain size, shape, and spatial alignment constraints. The texture-based methods treat the text region as a special type of texture and employ conventional texture classification method to extract the text. Recently, edge detection methods have been increasingly used for caption extraction due to the rich edge concentration in characters.

Many of the existing works deal with text extraction in static images [4][6][9][14][15]. Even though some address text extraction in video frames, they usually treat each video frame as an independent image [1][5][8][13]. When temporal information are utilized, they are used only for text enhancement through multi-frame averaging or time-based minimum pixel search [7][8][10][11]. These approaches require text detection and localization for every frame of a video, and careful caption blocks tracing and matching are needed between each frame pair for multi-frame enhancement. After all the processing, temporal information is still not fully utilized.

In this paper, we propose a new text extraction method that takes full advantage of the temporal information in a video sequence. First, by detecting the appearance and

0-7803-7304-9/02/$17.00 ©2002 IEEE

disappearance of caption text [3], we identify video segment that contains the same caption text. Then using the gray-level vector traced across the video segment as the feature vector for a pixel point, we distinguish a caption pixel from a background pixel. For a caption pixel, the vector should have a stable intensity. For a background pixel, the vector should change significantly over the same period. Such an approach can utilize all the temporal information in the video segment without requiring the time-consuming text detection and tracing for every frame in the segment.

2. CAPTION (DIS)APPEARANCE DETECTION

In order to obtain the temporal feature vector, we need to segment the video sequence into segments that contain the same caption text. First, we can use a conventional shot boundary detection technique to segment a video sequence into camera shots. Since there is relatively small change in content between adjacent frames within each shot, it should be easier to detect the caption changes within a shot.

We use the metric, quantized spatial difference density (QSDD), to detect the caption transition frame [3]. We first compute the direct difference between two neighbor frames. We observe that a small movement of a scene between adjacent frames produces many residual edge pixels at object boundaries in the difference image. A direct summation of these edge pixels may result in a value higher than that caused by caption transition. Fortunately, most of these edge pixels are sparsely distributed, while the residual pixels produced by the caption are highly concentrated because of the dense stroke pattern of characters. So we compute a feature that can measure the residual pixel density distribution.

The QSDD metric is defined by a two-step thresholding of the difference image between a pair of adjacent frames. We first compute the difference between a pair of adjacent frames. For each pixel with a difference value higher than a binarization threshold, the value 1 is assigned to the same location in a binary difference map. Otherwise the value 0 is assigned. The binary map is then uniformly partitioned into a number of small blocks. A block is labeled as Significant Change (SC) block if the summation of binary values in the block exceeds a second threshold.

Then the QSDD metric simply counts the number of SC blocks. Since the caption residual pixels are closely distributed in the difference image, they tend to produce more blocks with significant changes. Therefore, the difference image at a caption transition tends to produce a high QSDD value, thus can be identified. Finally, to check whether there is a caption transition at a shot boundary, we compare the caption regions in the two frames right before and after the shot boundary. If they are the same, then no caption transition happens. Otherwise, we consider them as two different captions.

3. TEMPORAL FEATURE VECTOR

Within a video segment that contains the same caption, if we trace the gray level of each pixel across the whole segment, we can obtain a vector describing the gray scale change of the pixel during the period. For a pixel on the caption, such a vector should have fairly constant values. For a pixel at the background, the vector may vary over a wide range of values. Such a vector can thus be used as feature vector to classify the two types of pixels.

Figure 1 shows some sample frames of a video segment containing the same caption text. As we can see, the background varies significantly over the period, thus produces very different clarity of the characters in different frames. A simple averaging may not always produce better results since the character can have similar averaging value to the background. If a minimum operation is used, some pixels on a character may be lost because of random noise thus reducing the already low quality of the character.

By using the temporal gray-scale vector as a feature vector, we retain all the information that can distinguish a caption pixel from a background pixel. To illustrate the vector difference between the caption and background, we use the principal component analysis method to compress the vector then show the first three principal components in Fig. 2. Figure 2 (a, b, c) show some sample frames from the original video sequence and (e, f, g) are the first three principal components computed from the feature vector. Figure 2 (d) illustrates the three principle components in a 3-D coordinates with each point in the 3-D space corresponds to a pixel in the video frame. We can clearly see that the caption pixels and background pixels do not overlap with each other. Using a simple classifier we can easily classify the two class of pixels. Some results are shown in Fig. 2 (h).

4. SUMMARY

In this paper, we describe a caption text extraction approach that takes full advantage of the temporal information contained in a video. Video captions are clearly separated from the background. The new method appears more complicated than some conventional approaches. However, instead of working on one frame at a time like the conventional methods, the new method addresses a whole segment of video at once. So the computational load is not high after averaging over each frame. One drawback of the algorithm is that it cannot deal with special effect moving captions. However, since these types of captions are used far less frequently than regular captions, we can afford to use more complicated caption

tracing techniques for them.

ACKNOWLEDGMENT

We thank Dr. Qiumei Yang for many constructive comments. The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region. (Project no. CUHK4378/99E).

REFERENCES

[1] L. Agnihotri and N. Dimitrova, "Text detection for video analysis," *Workshop on Content-based access to image and video libraries, in conjunction with CVPR*, Colorado, June, 1999.

[2] X. Gao and X. Tang, "Automatic news video caption extraction and recognition," in *Proc. of Intelligent Data Engineering and Automated Learning 2000*, pp. 425-430, Hong Kong, Dec. 2000.

[3] X. Tang, X. Gao, J. Liu, and H. J. Zhang "A spatial-temporal approach for video caption detection and recognition," *IEEE Trans. on Neural Networks*, special issue on intelligent multimedia processing, July, 2002.

[4] C. Garcia and X. Apostolidis, "Text detection and segmentation in complex color images," *Proc. of IEEE International Conf. on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 2326 -2329, 2000.

[5] A. K. Jain and B. Yu, "Automatic text location in images and video frames," *Pattern recognition*, Vol.31, No.12, pp.2055-2076, 1998.

[6] C. M. Lee and A. Kankanhalli, "Automatic extraction of characters in complex scene images," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 9, No. 1, pp. 67-82, 1995.

[7] H. P. Li, D. Doemann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. on Image Processing*, Vol.9, No.1, pp.147-156, 2000.

[8] R. Lienhart and F. Stuber, "Automatic text recognition in digital videos," *Proceedings of SPIE Image and Video Processing IV 2666*, pp.180-188, 1996.

[9] J. Ohya, A. Shio, and S. Akamatsu, "Recognizing characters in scene images," *IEEE Trans. On PAMI*, Vol.16, pp.214-220, 1994.

[10] T. Sato, T. Kanade, E. K. Kughes, M. A. Smith, and S. Satoh, "Video OCR: indexing digital news libraries by recognition of superimposed captions," *ACM Multimedia Systems*, 7(5), pp.385-395, 1999.

[11] J. C. Shim, C. Dorai and R. Bolle, "Automatic text extraction from video for content-based annotation and retrieval," *Proceedings of International Conference on Pattern Recognition*, pp.618-620, 1998.

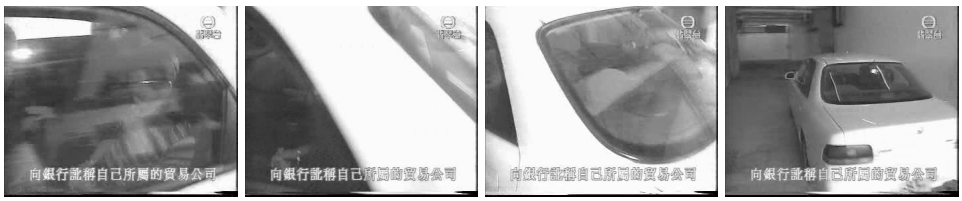
[12] A. Wernicke and R. Lienhart, "On the segmentation of text in videos," *Proceedings of IEEE International Conference on Multimedia and Expo*, Vol. 3, pp. 1511-1514, 2000.

[13] E. K. Wong and M. Chen, "A robust algorithm for text extraction in color video," *Proc. of IEEE Int. Conf. on Multimedia and Expo*, Vol. 2 , pp. 797-800, 2000.

[14] V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: an automatic system to detect and recognize text in images," *IEEE Trans. PAMI*, Vol. 21, pp.1224-1229, Nov. 1999.

[15] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," *Pattern Recognition*, Vol.28, No.10, pp1523-1535, 1995.

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



(e) (f) (g) (h)

Figure	a	b	c	d	e	f	g	h
Frame Index	1	16	31	39	46	54	61	76

Figure 1: Sample frames of a video segmentation with the same caption text.

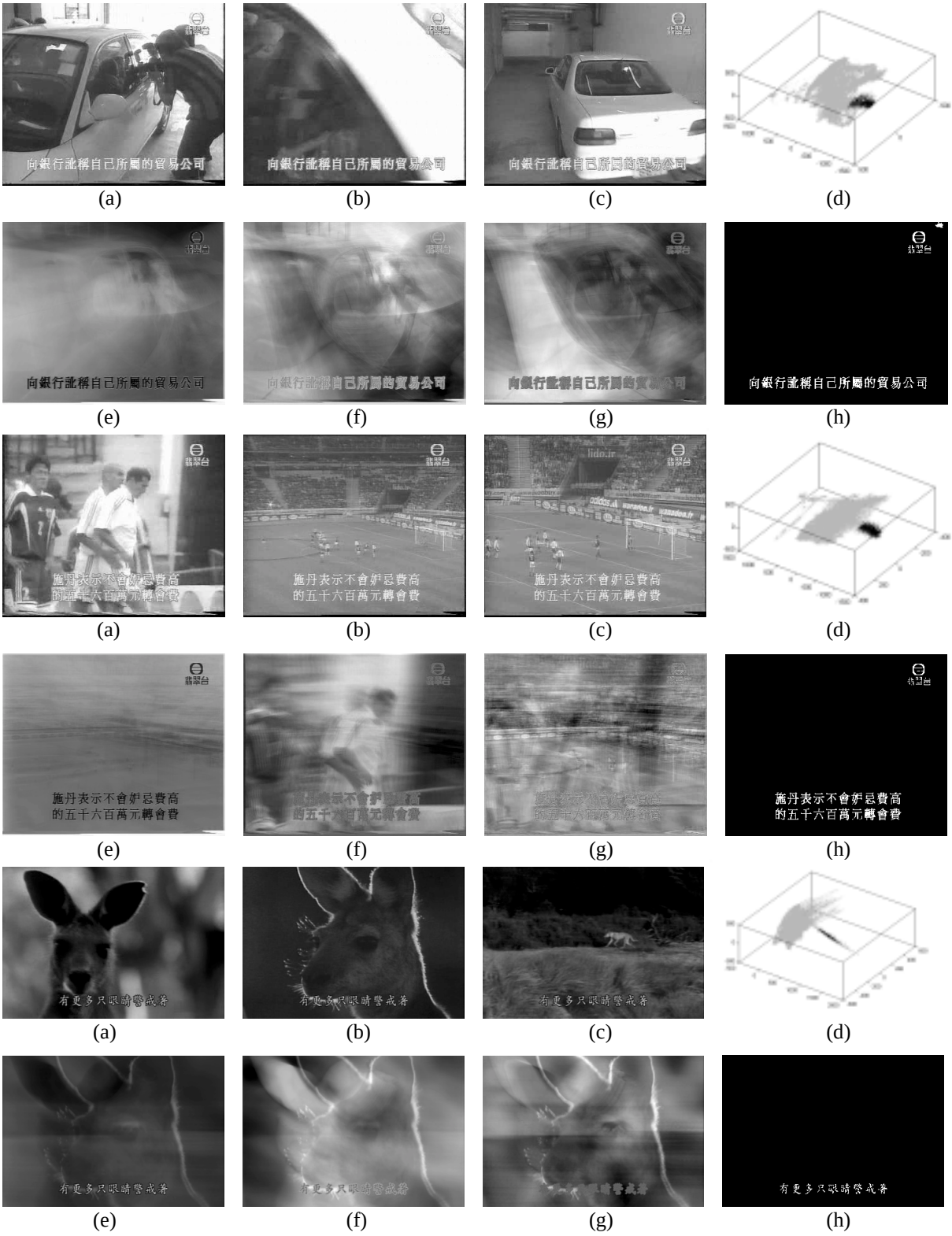


Figure 2: Sample results: (a,b,c), sample video frames; (d), first three principal components plotted in one 3-D coordinate with black points corresponding to caption pixels and gray points corresponding to background pixels; (e,f,g), first three principal components shown through three images; (h), segmentation results.