VIDEO CAPTION DETECTION AND EXTRACTION USING TEMPORAL INFORMATION

Bo Luo[1], Xiaoou Tang[1], Jianzhuang Liu[1], and Hongjiang Zhang[2]

[1]Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong, China
(Email: bluo1, xtang, jzliu@ie.cuhk.edu.hk)

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

ABSTRACT

Video caption detection and extraction is an important step for information retrieval in video databases. In this paper, we extract text information in video by fully utilizing the temporal information contained in the video. First we create a binary abstract sequence from a video segment. By analyzing the statistical pixel changes in the sequence, we can effectively locate the (dis)appearing frames of captions. Finally we extract the captions to create a summary of the video segment.

1. INTRODUCTION

In recent years, content-based image and video indexing and retrieval has been an active research area that attracts attention of many researchers. There are a large number of researches on content-based video indexing using low-level features such as color, texture, shape, motion, etc [1][2][3][4]. Complement to the low level features, researchers are beginning to use such high level features as text in video for video indexing because of the rich content information contained in them.

There are two classes of text embedded in video frames: the scene text, which appears in the video scene as an integral part of the scene content, and the graphic text, which contains the mechanically embedded characters [5]. The later one, such as news video captions or movie subtitles, serves as an important supplement of the audio-visual content and provides abundant high-level semantic information. Efforts have been made to extract and recognize these characters automatically to enable access to the high-level content of video data. Current text detection and extraction schemes can be generally grouped into 3 categories [8] – connected component based [10][12], texture classification based [5] and edge detection based methods [9][13][15][16]. In these works, text in video frames is treated mainly the same way as that in still images. Although some of them compute the average, variance or maximal/minimal values of consecutive frames to enhance the text [5][12][14][15], the temporal information contained in video is not fully utilized.

In [8] we proposed a novel method to segment video text using *temporal feature vectors*, but the caption (dis)appearance detection is still based on the information in the spatial domain. In his paper, we present a method to achieve the entire caption detection and extraction processing by taking full advantage of temporal information.

2. ABSTRACT IMAGE SEQUENCE

By tracing the gray-level of each pixel in time over a sequence of consecutive frames, we use the brightness values of a pixel to form a vector, which is called the *temporal feature vector* (TFV), to describe the gray scale change of that pixel. Keeping on tracing the gray-level values through the whole segment, we observe that at a caption appearing frame, a number of background pixels turn to caption pixels; likewise, at a caption disappearing frame, a number of caption pixels turn to background pixels. Figure 1 shows some examples.

Based on the above observation, we design the following process to create a sequence of images that represents these collective actions more clearly. First, we pick up the first 30 frames (i.e. frames $[F_1, F_{30}]$), and apply a supervised classification of the TFVs to cluster the pixels into caption and background. When a stable caption is contained in these frames, the caption pixels will be segmented and shown in the resulting binary image; otherwise, shown in the image are only some noise pixels, whose TFVs act like caption during this 30-frame-period. Then we move forward at a step length of 5 frames (i.e. move from $[F_1, F_{30}]$ to $[F_6, F_{35}]$) to compute another segmented binary image. By repeating

0-7803-7750-8/03/$17.00 ©2003 IEEE

this process, a sequence of binary images of segmented captions is finally obtained. Each image I_i in the sequence represents abstract textual information of original frames $[F_{5i+1}, F_{5i+30}]$. We call it an *abstract image sequence*. Figure 2 gives a brief demonstration of this process and some examples of abstract images.

The selection of the number of frames in each period is based on the assumption that each caption is present at least 1 second, within which there are 30 frames as designed in many major video standards. Thus any caption appearing in at least 30 frames will produce at least one image in the abstract sequence.

3. CAPTION (DIS)APPEARANCE DETECTION

We call changes from background to caption *positive changes*, while changes from caption to background *negative changes*. The numbers of pixels taking these changes are calculated by

$$|PC|_i = |\, PositiveChanges \,|_i = |\, I_{i+1} \text{ AND NOT } I_i \,| \qquad (1)$$

$$|NC|_i = |\, NegativeChanges \,|_i = |\, I_i \text{ AND NOT } I_{i+1} \,|, \qquad (2)$$

where I_i and I_{i+1} denote two consecutive binary images in the abstract sequence. Since these images are binary, I_i and I_{i+1} are logical matrices in which a *true* value corresponds to a caption pixel. $|\cdot|$ denotes number of *true* values in the matrix. Computing |PC| and |NC| over the entire abstract sequence, we get two curves that describe statistically the state of pixels taking changes. Figure 3 shows these curves computed over a 5-minute movie segment.

The appearance of one caption implies a relatively large number of pixels taking positive changes at the same frame, which creates a peak in the |PC| curve. Likewise, disappearance of one caption corresponds to a peak in the |NC| curve. By detecting the peak values in the |PC| and |NC| curves, we can locate the (dis)appearance of captions. We develop the following scheme to detect the caption changes.

1. A global threshold α is set. For any $|PC|_i \geq \alpha$, F_{5i+1}, the first frame corresponding to I_i, is marked as the appearance frame of a caption. For any $|NC|_i \geq \alpha$, F_{5i+30}, the last frame corresponding to I_i, is marked as the disappearance frame of a caption.
2. If there are more than one consecutive caption appearance frames without any disappearance frame between them, only the one with larger $|PC|_i$ is marked.
3. If there are more than one consecutive caption disappearance frames without any appearance frame between them, only the one with larger $|NC|_i$ is marked.

Cases 2 and 3 are used to remove the errors caused by noise pixels that take positive or negative changes at the same time.

Figure 4 shows a small part of the |PC| and |NC| curves with the caption (dis)appearing marks found by our method, as well as the caption (dis)appearing frames manually labeled for comparison.

With the detected (dis)appearance, the whole video segment is divided into fractions, each containing the same caption text. Then each fraction is equidistantly re-sampled into 30 frames and sent to a final classification, to generate a summary image of the fraction. Detailed classification processing is described in [8]. Figure 5 shows some examples of the final results.

4. IMPLEMENTATION OF THE SYSTEM

To reduce the storage demand and time complexity, the implementation of our system is conducted as follows.

A few buffers and a flag are defined. They are: a FIFO (first-in-first-out) queue as frame buffer (FB), a previous abstract image buffer (PIB) which is for one binary image only, previous peak buffers, PC peak buffer (PCB), NC peak buffer (NCB), and an in-caption flag (ICF). At the beginning, the first 30 frames are read into FB and the segmentation result is stored in PIB. Then repeat the following procedure to compute final summaries directly.

1. When there is no caption, remove the first 5 frames in the FB and append 5 more from the video segment. Perform classification and calculate the current |PC| and |NC| values by operations relative to PIB (see Eq. (1) and (2)), then refresh the PIB with the new classification result.
2. If appearance of caption is detected by comparing the current |PC| with a preset threshold, set the ICF, store the frame number and |PC| value into PCB. Then stop removing frames from FB and keep on appending 5 frames at the end in each loop. This way, the FB length keeps increasing with the same caption in it. However we still only process the last 30 frames, until the disappearance of caption is found. The location of detected disappearance and the corresponding |NC| value is temporarily stored in NCB until it is confirmed by a following caption appearance. Otherwise, it might be replaced by another directly following disappearance frame with larger |NC|.
3. When the next appearance of caption is detected, frames between the locations indicated in PCB and NCB are then confirmed to have the same caption. They are re-sampled into 30 frames and are finally classified. A summary image of the corresponding caption is finally created and the redundant frames in FB are then cleared.

With the above algorithm, we do not store the abstract image sequence and only buffer a limited

number of frames of the video segment, so the storage requirement is small. The classification process is simple and there is no other time-consuming operation, so the computation is at an acceptable level. The total time complexity of our algorithm is $O(n)$ with n being the number of frames in the video segment. The low computation of the algorithm makes it applicable to extracting text from large segments, e.g. an entire movie.

5. SYSTEM PERFORMANCE

To test the performance of the proposed caption (dis)appearance detection algorithm, we extract 10000 consecutive frames from a US movie with Chinese captions, and manually select the appearance and disappearance frames as ground truth. There are totally 69 different captions in this segment, our system detected 65 of them, with only 1 false alarm. The detect and missing rates are calculated by

DetectRate = |DetectedCaptions| / GroundTruth = 94.2%
MissRate = |MissedCaptions| / GroundTruth = 5.8%
FalseAlarmRate = |FalseAlarms| / |DetectedCaptions| =1.5%

Another benchmark is the accuracy of the detected position of (dis)appearance boundary frames. The mean absolute error (MAE) and the mean square root error (MSRE) are calculated by

MAE = mean(|DetectedBoundary-ActualBoundary|)
MSRE = (mean(DetectedBoundary-ActualBoundary)^2)$^{1/2}$

are shown in the following table. As stated in section 2, the precision of our method is 5 frames (as we pick a step length of 5 frames). If the detected position of boundary frame is within 5 frames from the actual boundary frame, the detection is regarded accurate. Number of accurate detections (ADN) and the rate of accurate detections (ADR) are also shown in the table.

	MAE	MSRE	ADN	ADR
Appearance	2.3182	4.3641	64	98.46%
Disappearance	3.4242	5.6809	58	89.23%

6. SUMMARY

In this paper, we present a video caption detection and extraction method that takes full advantage of temporal information. We trace over the video segment to extract an abstract sequence with coarsely segmented caption text. Then we statistically analyze the pixels changing between adjacent abstract images and detect the (dis)appearance of captions. Refined caption text is then extracted and a summary of captions is finally created. The final indexing key frames give a summary of captions contained in the video segment. These frames are of high quality and can be sent to OCR recognition. With the implementation scheme described in section 4, the computational complexity of our system is low. Our algorithm does not make any assumptions on the shape of the caption, i.e. we do not need the captions to be horizontal, constant size, certain font or fixed location.

7. ACKNOWLEDGMENT

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region under Grant CUHK4378/99E, CUHK4357/02E, and Grant AoE/E-01/99.

8. REFERENCES

[1] S. W. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia*, pp. 62--72, Summer 1994.

[2] H. Zhang, J. Wang, and Y. Altunbasak. "Content-based video retrieval and compression: A unified solution." *In Proc. of IEEE International Conference on Image Processing*, 1997

[3] H. Zhang, J. Wu, D. Zhong, and S. Smoliar, "An integrated system for content-based video retrieval and browsing," Pattern Recognition, vol. 30, no. 4, pp. 643--658, April 1997.

[4] A. Yoshitaka and T. Ichikawa, "A survey on content-based retrieval for multimedia databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 81-93, Jan.-Feb. 1999.

[5] H. Li, D. Doermann, and O. Kia, "Automatic Text detection and tracking in digital video," *IEEE Transactions on Image Processing*, vol. 9, no.1, pp. 147-156, 2000.

[6] X. Tang, X. Gao, J. Liu and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, vol. 13, no. 4, July, 2002.

[7] X. Gao and X. Tang, "Unsupervised video shot segmentation and model-free anchorperson detection for news video story parsing," *IEEE Transaction on Circuits, Systems and Video Technology*, vol. 12, no. 9, Sept., 2002.

[8] X. Tang, B. Luo, X. Gao, E. Pissaloux, and H. Zhang, "Video text extraction using temporal feature vectors," *in Proc. of IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, Aug. 2002.

[9] 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.

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

[11] 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

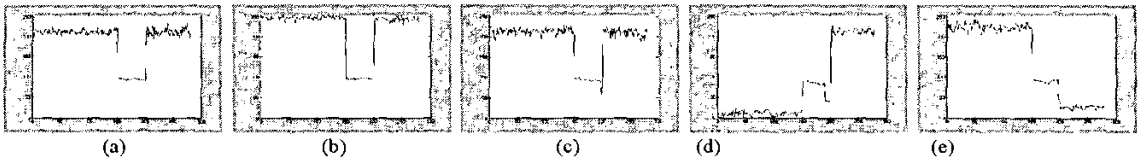
[12] R. Lienhart and F. Stuber, "Automatic text recognition in digital videos," *Proceedings of SPIE Image and Video Processing IV 2666*, pp.180-188, 1996.

[13] 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.

[14] J. C. Shim, C. Dorai and R. Bolle, "Automatic text extraction from video for content-based annotation and retrieval," *Proceedings of ICPR*, pp.618-620, 1998.

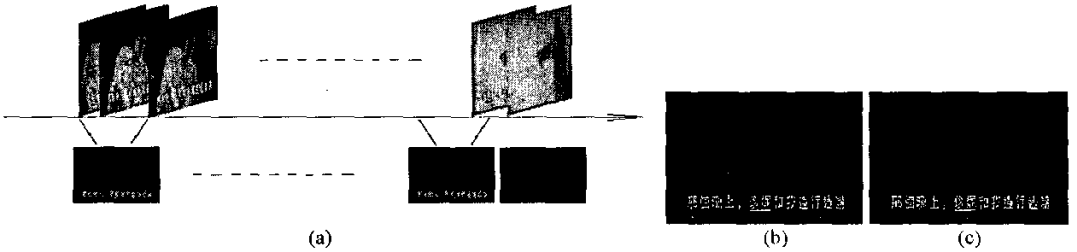
[15] A. Wernicke and R. Lienhart, "On the segmentation of text in videos," *Proceedings of ICME*, Vol. 3, pp. 1511-1514, 2000.

[16] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 12 no.4, April, 2002.



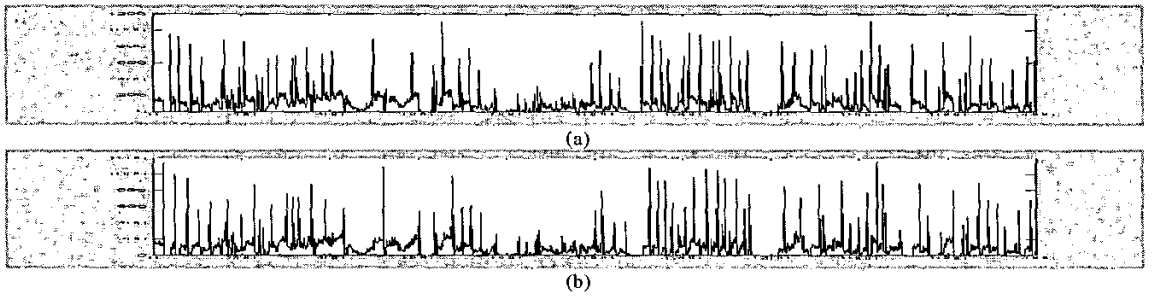
(a) (b) (c) (d) (e)

Figure 1: Tracing gray-scale level over two successive captions. (a), (b) and (c) are TFVs corresponding to caption pixels appearing in both captions. (d) and (e) correspond to pixels appearing only either caption 2 or caption 1 respectively.



(a) (b) (c)

Figure 2. (a). A brief demonstration of the process of extracting abstract image sequence. (b) and (c). Two examples of abstract images that contain the same text.



(a)

(b)

Figure 3. Examples of |PC| and |NC| curves. (a) shows a |PC| curve and (b) shows an |NC| curve. Part of both curves are enlarged and shown in figure 4 with more details.

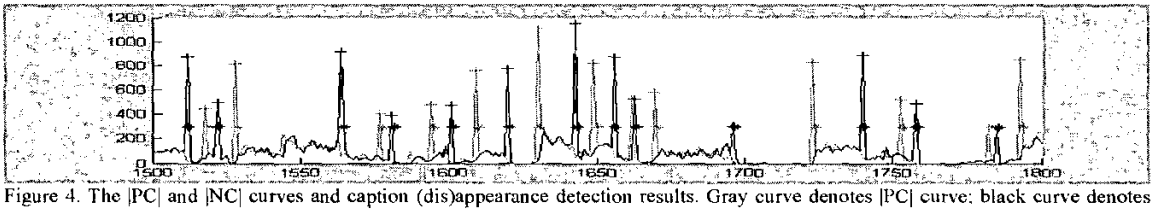


Figure 4. The |PC| and |NC| curves and caption (dis)appearance detection results. Gray curve denotes |PC| curve; black curve denotes |NC| curve. Gray marks correspond to appearance of captions and black marks correspond to caption disappearances. "+" marks caption (dis)appearance detected by our system; "*" marks caption (dis)appearance manually labeled as ground truth.

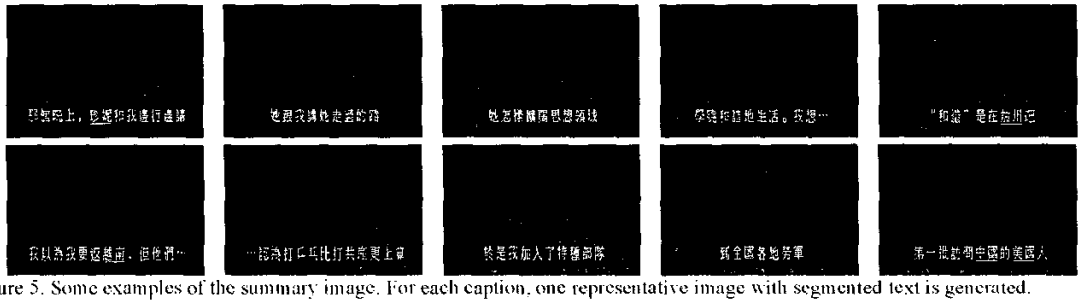


Figure 5. Some examples of the summary image. For each caption, one representative image with segmented text is generated.