WUJNS
Wuhan University Journal of Natural Sciences

Vol.10 No.2 2005 363-367

Article ID: 1007-1202(2005)02-0363-05

A Method to Recognize Caption Area in MPEG Compressed Video

ZHENG Peng, LU Xiao-ping, ZHOU Dong-ru

School of Computer, Wuhan University, Wuhan 430072, Hubei, China

Abstract: An efficient method to recognize caption area in MPEG compressed video was presented, by making use of the contrast of I-frame to distinguish caption area with background. We define texture energy, intensity of boundary, distance of background, and texture correlation to recognize caption area and caption frame. The benefit of only analyzing I-frame is that we can make use of DCT coefficients directly without losing information. We have experimented with our algorithm, and the result of experiment indicates that the performance of the algorithm is efficient.

Key words: caption; texture; MPEG

CLC number: TP 311.5

Received date: 2004-05-06

Foundation item: Supported by the Natural Science Foundation of Hubei Province(2004ABA174)

Biography: ZHENG Peng (1965-), male, Associate professor, research direction: scientific visualization and video information processing, E-mail: pzheng51@163.com

0 Introduction

With the continued increasing of video data, we demand true content based indexing and retrieval system. In retrieving video, text is one of important clues. Li *et al*[1] pointed out that text can be divided into two classes, scene text and graphic text. Scene text appears within the scene and is captured by the camera. Graphic text is mechanically added to video frames to supplement the visual and audio content. Graphic text is often called caption. Since it is purposefully added, it is often more structured and closely related to the subject than scene text.

Considering the cost of transmitting and storing video data, most video data is transmitted and stored in compressed format. MPEG compressed video is the most popular, because of its efficient compressed ratio and quality of image.

In order to avoid cost of decompressing, we present a method to recognize caption area directly in MPEG compressed video. If we can recognize directly on compressed video, we can avoid the expensively decoding cost and accelerate the speed of indexing.

There are many scholars, such as Jain[2], Li[1,3], Lienhart[4,5], Mariano[6], Shim[7], have researched the methods of recognizing and extracting text in decompressed video, and the recognizing ratio that they have acquired is rather well. All above algorithms have high cost of computing because they need to decompress MPEG video at first. Lim[8], Zhang[9], and Zhong[10] present methods that detect caption area direct in MPEG compressed video. They all make use of DCT (Discrete Cosine Transform) coefficients, which provided by MPEG compressed video. In DCT, coefficient in location (0,0) is called DC(Discrete Cosine) coefficient and the other

values we call them AC (Alternating Current) coefficients. Lim restricts location which text emerges. Zhang makes use of DC coefficients to compute texture energy of block. Zhong only use horizontal intensity variation. Currently, many scholars suggest analyze AC coefficients of block. The main problems of the kind of method are that how to select AC coefficients, and how to deal with boundary, and how to deal with complicated background.

1 Texture Feature of Text Area

We only deal with I-frames. Caption usually will be lasted a rather long period, longer than 1s, in order to be seen clearly by audience. The text frame will be emerged in a GOP(groups of picture) at least, about 0.5 s. We can make use of I-frames in the shot to detect text frame. One advantage of using I-frame is that we can make use of AC coefficients directly.

Different text maybe has different color or luminance, and the case of background is similar. We can't distinguish text with background using feature of color or luminance. There is a high contrast between text and background in video sequence. The color and texture of text area is different from background. AC coefficients of MB(Macro block) represent texture feature. We use A_{ij} to represent AC coefficients.

From DCT coefficients in MPEG, we know A_{0j} and A_{i0} represent density changing of image in horizontal and vertical direction respectively, and A_{ij} represent density changing of image in diagonal direction. The low frequency coefficients represent coarse texture, and more sensitivity for human eyes. The high frequency coefficients represent fine texture, and less sensitivity for human eyes. MB emerging text includes both kinds of texture. We select the first 9 AC coefficients of an MB to build the texture energy of the MB. The definition of E_T (texture energy) is as Eq. (1).

$$E_T(i,j) = \sum_{k=1}^{9} A_k(i,j) \tag{1}$$

i, j is the coordinate of MB, respectively. E(the average texture energy of a frame) is defined as:

$$E = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} E_T(i,j)}{M \times N} \tag{2}$$

M, N is the number of MBs in horizontal or in vertical , respectively. If the quotient that the texture energy of a MB divided by the average texture energy of a frame is greater than a threshold, the MB is called text MB. We can represent MB by binary digit. The definition of B_T(text block) is as follows.

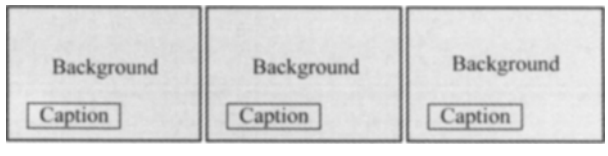
$$B_T(i,j) = \begin{cases} 1, & E_T(i,j)/E > T_H \\ 0, & E_T(i,j)/E \leqslant T_H \end{cases} \tag{3}$$

The texture energy of text MB is larger than the texture energy of other MB, so the T_H in (3) is larger than 1 at least. Based on the Shannon entropy[11], we can select T_H automatic, and T_H is about 1.5. In experiment, we find that the result is satisfied when the value of T_H between 1.2 and 1.5. The examples in the paper, we select 1.5 as T_H.

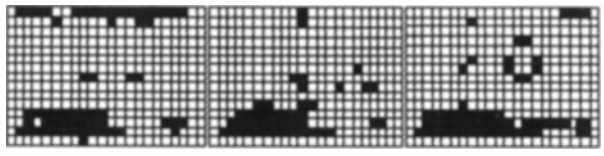
Fig. 1 is the example that I-frame and corresponding binary images using expression (1),(2),(3). In Fig. 1 (b), The black blocks represent text block, and the white blocks represent background block.

From Fig. 1, we find the MB that not emerging text maybe have large texture energy, because of the complexity of background texture. The large energy MBs of background emerge randomly and discretely. Caption is usually a sentence, so the large energy MBs of text emerge consecutively, and form a rectangle. In order to distinguish the two kinds large energy blocks, we take the method as follows.

Horizontal caption is more common than vertical caption. We emphasize to recognize horizontal caption. First, we scan binary image according to row. If the length of consecutive black blocks is less than 3, we take them as white blocks. Second, we scan binary image according to column. If the length of consecutive black block is less than 2, we take them as white. We call such operation as smoothness. Fig. 2 is the smoothing result of Fig. 1.



(a) 3 I-frames in moving shot



(b) Binarized image corresponding (a)

Fig. 1 Results of recognising moving shot

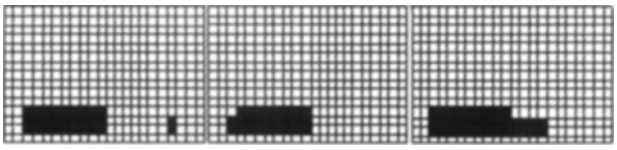


Fig. 2 The smoothing result about Fig. 1

In order to let audience watch caption clearly, we would not add caption at the brim of image. We can recognize text area without considering 4 brims. Because texture energy is independent of kinds of language, we can recognize other language text.

The precondition of smoothing is that a text area will include 3 MBs in horizontal direction and 2 MBs in vertical direction. The premise is suitable for most cases, but it is possible that the number of MB with high contrast will be decreased when there are several text areas in frame picture. If all text area is be smoothed, we will lost some information, and it is possible mistake text frame as non-text frame.

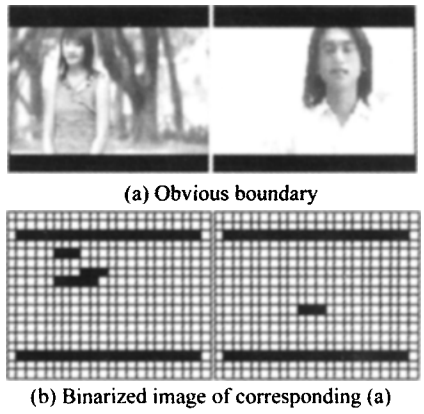
In order to avoid text area being smoothed, we modify the method of smoothing as follows.

After smooth in horizontal direction, if there is only one non-text block between two text area in every row, we consider the non-text block as text block, and merge the two text area as one text area. If the number of MB in the largest continuous text area takes the major, it is possible that emerge a large continuous text area, and we smooth the binary image in column. Otherwise, we don't smooth in column.

We make use of "and", "or" operation between two rows in binary matrix to compute the number of MB in a continuous text area. In binary matrix, 1 stands for text MB, and 0 stands for non-text MB. If the all result of "and" corresponding position between two rows is zero, the two rows is not continuous. Otherwise, the two rows is continuous and the number of MB in continuous text area is the sum of the value of "or" plus the value of "and".

2 Texture of Boundary and Special Background

From Fig. 3, we find some non-text block being considered as text block. If non-text block is considered as text block in a frame that emerges text, it doesn't matter, because we only decoded frames that emerge text. If non-text block is considered as text block in a frame that doesn't emerge text, it will increase the cost of unnecessary decoding.



(a) Obvious boundary

(b) Binarized image of corresponding (a)

Fig. 3 Mistake non-text area as text area

The reason that emerges in Fig. 3 is that we assume existing high contrast between text block and background block. When a boundary emerges in a frame, contrast of both side of boundary is also high. In order to distinguish text block with boundary block, we look into four blocks in MB. When boundary emerges in a MB, the blocks of both sides have large difference of luminance and the blocks of the same side have small difference of luminance. The luminance of 4 blocks in text block is uniformity and the difference of luminance is small. We can make use of luminance DC coefficients of block to distinguish boundary block with text block. The intensity of boundary is defined as follows.

$$\begin{aligned} L_1 &= |\,(D_0 + D_1) - (D_2 + D_3)\,| \\ L_2 &= |\,(D_0 + D_2) - (D_1 + D_3)\,| \\ L_3 &= |\,(D_0 + D_3) - (D_1 + D_2)\,| \\ B_I &= \frac{\frac{L_1 + L_2 + L_3}{3}}{1 + \sum_{i=0} |\,D_i\,| /4} \end{aligned} \tag{4}$$

D_i represent the DC coefficient of block i. In text MB, luminance of four blocks is uniformity, and the value of E is small. When boundary emerges in a MB, the value of BI is large.

$$B_{MB}(i,j) = \begin{cases} 1, & B_I < T_e \\ 0, & B_I \geqslant T_e \end{cases} \tag{5}$$

i and j represent position of high energy in matrix. 1 represents text block, and 0 represents boundary block. During experiment, we find the feature of boundary is obvious when the value of E is larger than 3, and so we select 3 as T_e. After building up binary matrix by Eq. (3), we compute boundary intensity and assign the non-text block as 0. Fig. 4 is the result of eliminating bounda-

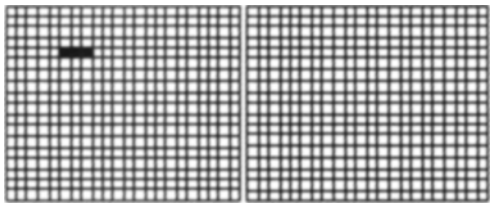


Fig. 4 Eliminate boundary MB in Fig. 3

ry of Fig. 3. In order to compare result, we don't smooth column in Fig. 3.

Besides boundary MB has high texture energy, some special background MBs, such as branch of trees, network, and so on, also has high texture energy. Such case will lead to consider non-text frame as text frame.

In order to deal with special background, it is necessary to analyze high texture energy MB. We find that difference of color between high texture energy background and low texture energy background is small and the difference of color between text and background is large. So we make use of the difference of color to distinguish text with special background. When a high texture energy MB has adjacent low texture energy MB, we just estimate up, below, left, and right 4 neighbors at most. We define distance of background as follows.

$$\begin{aligned} N_1 &= | D_{TY} - D_{BY} | \\ N_2 &= | D_{TU} - D_{BU} | \\ N_3 &= | D_{TV} - D_{BV} | \\ D_{TB}(i,j) &= N_1 + N_2 + N_3 \end{aligned} \tag{6}$$

In Eq. (6), D_{TY}, D_{BY} represents the DC coefficients of luminance of high and low energy block respectively. D_{TU}, D_{BU} represents the DC coefficients of color of high and low energy block respectively. D_{TV}, D_{BV} represents the DC coefficients of another color of high and low energy block respectively.

When maximum $D_{TB}(i,j)$ of up, below, left, and right 4 neighbors is greater than T_t, $B_T(i,j)$ is 1, otherwise, $B_T(i,j)$ is 0.

i and j represent position of high energy in matrix. L,R,U,B represent left, right, up, and below four directions. If background distance between high energy MB and low energy MB is larger than a threshold, we consider the high energy MB as text MB. Otherwise, we consider the high energy MB as non-text MB. The threshold T_t is 300 based on the entropy of Shannon.

If caption emerges in vertical direction, we can take the similar method. The goal of the algorithm is recognizing text frame directly using DCT coefficients. In experiment, the algorithm can acquire good result not only recognizing caption text, but also recognizing scene text.

3 Recognize Caption Frame

In order to represent continuous of MB, We define the texture correlation between MBs in a row. If the correlation is large, the MBs belong to text MBs, otherwise, non-text MBs. The definition of texture correlation in a row is as follows:

$$H_C(i) = \frac{\sum_{j=2}^{N} T_B(i,j) \times T_B(i,j-1)}{N-1} \tag{7}$$

T_B represents the binary matrix of MB. i, j represents row and column respectively. The correlation between MBs in a column is defined as follows:

$$V_C(j) = \frac{\sum_{i=2}^{N} T_B(i,j) \times T_B(i-1,j)}{M-1} \tag{8}$$

If the direction of caption is horizontal, the value of is H_C large. If the direction of caption is vertical, the value of is V_C large. If the text area includes many MBs or the number of text area more than one, the correlation between rows or between columns is large. We make use of Eq. (7) and Eq. (8) to define correlation between rows and between columns as Eq. (9) and (10) respectively.

$$T_{HC} = \frac{\sum_{i=2}^{M} H_C(i) \times H_C(i-1)}{\left[(M-1) \times \sum_{i=1}^{M} \sum_{j=1}^{N} T_B(i,j)\right]/(M \times N)} \tag{9}$$

$$T_{VC} = \frac{\sum_{i=2}^{M} V_C(i) \times V_C(i-1)}{\left[(N-1) \times \sum_{i=1}^{M} \sum_{j=1}^{N} T_B(i,j)\right]/(M \times N)} \tag{10}$$

When text emerges in picture, T_{HC} or T_{VC} will increase obviously. In experiment, we find the special background can influence the value of T_{HC}, but it will not influence for recognizing text frame. The robust of the algorithm is good. Fig. 5 is an example of recognizing text frame.

4 Experiment Result and Analysis

In order to evaluate the algorithm, we use recognition ratio(R) and precision ratio(P) as standard. We define recognition ratio and accuracy ratio as follows:

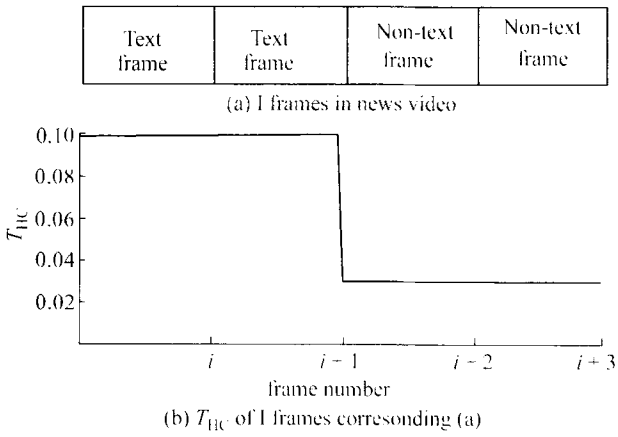


Fig. 5 The result of recognizing text frame

$$R = \frac{C}{C + M} \tag{11}$$

$$P = \frac{C}{C + F} \tag{12}$$

C stands for the correct number of text frames, M stands for the missed number of text frames, F stands for the false number of text frames.

To evaluate our method and compare with exist method, we have tested on about 5 000 I-frames MPEG compressed video sequences. Table 1 is the results of test. The algorithm is benefit for decreasing cost of recognizing text. The time complexity of the algorithm is $O(mnp)$. m represents the number of MB, n represents the number of I-frame in video sequences, p represents the number of high energy MBs.

Table 1 The results of test

Algorithm	R / %	P / %
Zhong *et al* [9]	92	83
Our method	95	90

5 Conclusion

In this paper, we propose a method to detect caption area and text frame. The algorithm bases on the DCT coefficients of MPEG. The computing cost is low, because we use the information directly, without decoding, and only analyze I-frames. If MPEG sequences are only made of I-frames, we will detect I-frame every other 15 frames. The next effort will be using A/V objects in MPEG-4 and the multimedia content description interface in MPEG-7 to help us detecting text area.

References

[1] Li H, Doermann D, Kia O. Automatic Text Detection and Tracking in Digital Video. *IEEE Trans on Image Processing*, 2000, **9**(1): 147-156.

[2] Jain A K, Yu B. Automatic Text Location in Images and Video Frames. *Proc of International Conf on Pattern Recognition*, 1998, **1**: 1497-1499.

[3] Li H, Doermann D. Automatic Identification of Text in Digital Key Frames. *Proc of the 11th International Conf on Pattern Recognition*, 2000, **1**: 618-620.

[4] Lienhart R, Stuber F. Automatic Text Recognition in Digital Video. *Proc of SPIE in Image and Video Processing*, 1997, **2666**: 180-188.

[5] Lienhart R. Comparison of Automatic Shot Boundary Detection Algorithms. *Proc of SPIE in Image and Video Processing*, 1999, **3656**: 29.

[6] Mariano V Y, Kasturi R. Locating Uniform-Colored Text in Video Frames. *Proc of the 15th International Conf on Pattern Recognition*, 2000, **4**: 539-542.

[7] Wernicke A, Lienhart R. On the Segmentation of Text in Videos. *IEEE International Conf on Multimedia and Expo*, 2000, **3**: 1511-1514.

[8] Lim Y K, Choi S H, Lee S W. Text Extraction in MPEG Compressed Video for Content-Based Indexing. *Proc of the 15th International Conf on Pattern Recognition*, 2000, **4**: 409-412.

[9] Zhang Y, Chua T S. Detection of Text Captions in Compressed Domain Video. *Proc of ACM International Conf on Multimedia*, 2000, **1**: 201-204.

[10] Zhong Y, Zhang H J, Jain A K. Automatic Caption Localization in Compressed Video. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2000, **22**(4): 385-392.

[11] Sahoo P K, Soltani S, Won A K C. A Survey of Thresholding Techniques. *Computer Vision, Graphics, and Image Processing*, 1988, **41**: 233-260.

□