JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 22, 1145-1162 (2006)

Automatic Closed Caption Detection and Filtering in MPEG Videos for Video Structuring

DUAN-YU CHEN, MING-HO HSIAO AND SUH-YIN LEE
Department of Computer Science and Information Engineering
National Chiao Tung University
Hsinchu, 300 Taiwan
E-mail: {dychen; mhhsiao; sylee}@csie.nctu.edu.tw

Video structuring is the process of extracting temporal structural information of video sequences and is a crucial step in video content analysis especially for sports videos. It involves detecting temporal boundaries, identifying meaningful segments of a video and then building a compact representation of video content. Therefore, in this paper, we propose a novel mechanism to automatically parse sports videos in compressed domain and then to construct a concise table of video content employing the superimposed closed captions and the semantic classes of video shots. First of all, shot boundaries are efficiently examined using the approach of GOP-based video segmentation. Color-based shot identification is then exploited to automatically identify meaningful shots. The efficient approach of closed caption localization is proposed to first detect caption frames in meaningful shots. Then caption frames instead of every frame are selected as targets for detecting closed captions based on long-term consistency without size constraint. Besides, in order to support discriminate captions of interest automatically, a novel tool – font size detector is proposed to recognize the font size of closed captions using compressed data in MPEG videos. Experimental results show the effectiveness and the feasibility of the proposed mechanism.

Keywords: caption frame detection, closed caption detection, font size differentiation, video structuring, video segmentation

1. INTRODUCTION

With the increasing digital videos in education, entertainment and other multimedia applications, there is an urgent demand for tools that allow an efficient way for users to acquire desired video data. Content-based searching, browsing and retrieval is more natural, friendly and semantically meaningful to users. The need of content-based multimedia retrieval motivates the research of feature extractions of the information embedded in text, image, audio and video. With the technique of video compression getting mature, lots of videos are being stored in compressed form and accordingly more and more researches focus on the feature extractions in compressed videos especially in MPEG format. For instances, edge features are extracted directly from MPEG compressed videos to detect scene change [5] and captions are processed and inserted into compressed video frames [7]. Features, like chrominance, shape and texture are directly extracted from MPEG videos to detect face regions [1, 3]. Videos in compressed form are analyzed and parsed for supporting video browsing [11].

Received May 13, 2004; revised August 23, 2004; accepted September 2, 2004.
Communicated by Ming-Syan Chen.

However, textual information is semantically more meaningful and attracts increasing researches on closed caption detection in video frames [2, 4, 6, 12-17]. The researches [6, 12-14, 16, 17] detect closed captions in pixel domain. In [16, 17], they proposed to detect closed captions in specific areas. However, it is impractical to localize closed captions in specific areas of a frame since in different video sources closed captions normally do not appear in a fixed position.

A number of previous researches extract closed captions from still images and video frames [13-15, 27, 28] with a constraint that characters are bounded in size. Besides, these approaches usually require the property that text has a good contrast from the background. However, text region localization with size constraint is not practical especially for the cases that those captions are small in size but are very significant and meaningful. For example, in sports videos, the superimposed scoreboards show the intermediate results between competitors and present the match as clearly as possible without interference.

There has been very little effort to extract features in compressed domain to detect closed captions in videos. Zhong *et al*. [2] and Zhang and Chua [4] detect large closed captions frame-by-frame in MPEG videos using DCT AC coefficients to obtain texture information in I-frames without exploiting the temporal information in consecutive frames. However, it is impractical and inefficient to detect closed captions in each frame. Due to the temporal nature of long-term consistency of closed captions over continuous video frames, it would be more robust to detect the closed caption based on its spatial-temporal consistency. Gargi *et al*. [15] perform text detection by counting the number of intra-coded blocks in P and B frames based on the assumption that the background is static. Hence, it is vulnerable to abrupt and significant camera motion. Besides, this approach is only applied to the P and B frames and does not handle captions that appear in the I-frames.

In this paper, in order to detect closed captions efficiently and flexibly, we propose an approach for compressed videos to detect caption frames in meaningful shots. Then caption frames instead of every frame are selected as targets for localizing closed captions without size constraint while considering long-term consistency of closed captions over continuous caption frames for removing noise. Moreover, we propose a novel tool – font size detector to identify font size in compressed videos. Using this tool, after the targeted font size is indicated, we can allow users to automatically discriminate captions of interest instead of captions in the presumed position. It is worth noticing that font size recognition is a critical step in the process of video OCR since a bottleneck for recognizing characters is due to the variation of text font and size [22-25]. Therefore, this tool can be used as a prefilter to quickly signal the potential caption text and thus reduce the amount of data that needs to be processed.

The proposed system architecture is shown in Fig. 1. All the tasks are accomplished in compressed domain. GOP-based video segmentation [8] is exploited to efficiently segment video into shots. The color-based shot identification is proposed to automatically identify meaningful shots. Caption frames in these shots are detected by computing the variation of DCT AC energy both in the horizontal and vertical directions. In addition, we detect closed captions using the weighted horizontal-vertical DCT AC coefficients. To detect closed captions robustly, each candidate closed caption is verified further by computing its long-term consistency that is estimated over the backward shot, the

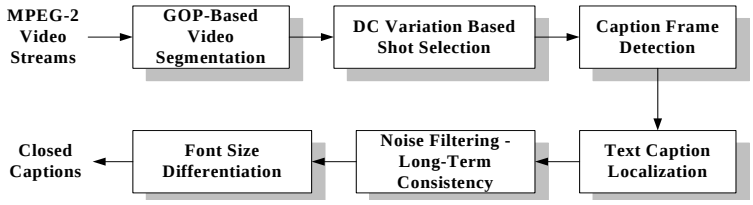


Fig. 1. Overview of the system architecture.

forward shot and the shot itself. After closed captions are obtained, we differentiate the font size of each closed caption based on horizontal projection profile of DCT AC energy in the vertical direction. Captions of interest can then be identified by the font size and size variance. Finally, captions of interest and the meaningful shots can be employed together to construct a high-level concise table of video content.

The rest of the paper is organized as follows. Section 2 describes the color-based shot identification. Section 3 presents the proposed approach of closed caption localization. Section 4 shows the experimental results and the prototype system of video content visualization. The conclusion and future works are given in section 5.

2. SHOT IDENTIFICATION

2.1 Video Segmentation

Video data is segmented into clips to serve as logical units called “shots” or “scenes”. In MPEG-2 format [9], GOP layer is a random accessed point and contains GOP header and a series of encoded pictures including I, P and B-frame. The size of a GOP is about 10 to 20 frames, which is less than the minimum duration of two consecutive scene changes (about 20 frames) [10]. Instead of checking frame-by-frame, we first detect possible occurrences of scene change GOP-by-GOP (inter-GOP). The difference between each consecutive GOP-pair is computed by comparing the corresponding I-frames. If the difference of DC coefficients between these two I-frames is larger than the threshold, then there might exist scene change in between these two GOPs. Hence, the GOP that might contain the scene change frames is located. In the second step – intra GOP scene change detection, we further use the ratio of forward and backward motion vectors to find out the actual frame of scene change within a GOP. By this approach, the experimental results [8] are encouraging and prove that the scene change detection is efficient for video segmentation.

2.2 Shot Identification

While the boundary of each shot is detected, the video sequence is segmented into shots consisting of the advertisement, close-up and court-view. Closed captions can then be detected in each video shot. However, it is impractical to detect closed captions in all video shots. In sports videos, the shots of court-view are our focus since the matches of the sports are primarily shown in the shots of court-view and the scoreboards are pre-

sented mostly in these kinds of shots. Therefore, scene identification approach is proposed to identify the shots of court-view.

To recognize the shots of court-view, it is worth noticing that the variation of the intensity in the court-view frames is very small through a whole clip and the value of intensity variation between consecutive frames is very similar. In contrast, the intensity of the advertisement and close-up varies significantly in each frame and the difference of the variance of intensity between two neighboring frames is relatively large. Therefore, the intensity variation within a video shot can be exploited to identify the shots of court view. In order to efficiently obtain the intensity variance of each frame and that of a video shot, DC-images of I-frames are extracted to compute the intensity variance. The frame variance $FVar_{s,i}^{DC}$ and the shot variance $SVar_s$ are defined by

$$FVar_{s,i}^{DC} = \sum_{j=1}^{N} DC_{i,j}^2 / N - \left(\sum_{j=1}^{N} DC_{i,j} / N \right)^2 , \tag{1}$$

and

$$SVar_s = \sum_{i=1}^{M} (FVar_{s,i}^{DC})^2 / M - \left(\sum_{i=1}^{M} FVar_{s,i}^{DC} / M \right)^2 , \tag{2}$$

where $DC_{i,j}$ denotes the DC coefficient of the jth block in the ith frame, N represents the total number of blocks in a frame, and M denotes the total number of frames in shot s.

Based on the fact that the intensity variance of a court-view frame is very small through a whole clip, shots are regarded as the type of court-view $Shot_{Court}$ by

$$Shot_{Court} = \left\{ Shot_s \mid FVar_{s,i}^{DC} < \delta_{frame} \text{ and } SVar_s < \delta_{shot}, \forall i \in [1,\ N] \right\} \tag{3}$$

where δ_{frame} and δ_{shot} are the predefined thresholds.

In order to demonstrate the applicability of the proposed shot identification, the variation of the intensity variance of each I-frame in sports videos including tennis, football and baseball is exhibited in Fig. 2. Fig. 2 (a) shows a tennis video composed of four tennis court shots, three close-up shots and a commercial shot. Fig. 2 (b) introduces a football sequence consisting of close-up shots and football field shots. A baseball sequence is presented in Fig. 2 (c) including pitching shots, baseball field shots and close-up shots. From Fig. 2, we can observe that the intensity variance of the type of court-view is very small and the value is very similar through a whole clip. Thus, the clips of court-view can be indicated and selected by the characteristic that the value of intensity variance $FVar_{s,i}^{DC}$ is small in each individual frame and is consistent over the whole shot. Therefore, the proposed approach of shot identification can be applied to identify court-view shots of sports videos, in which the view of a match consists of the intensity-consistent background of a court or athletic field.

3. CLOSED CAPTION LOCALIZATION

In this section, we shall elaborate how to detect caption frames and how to detect closed captions in caption frames. To avoid the time-consuming overhead of closed

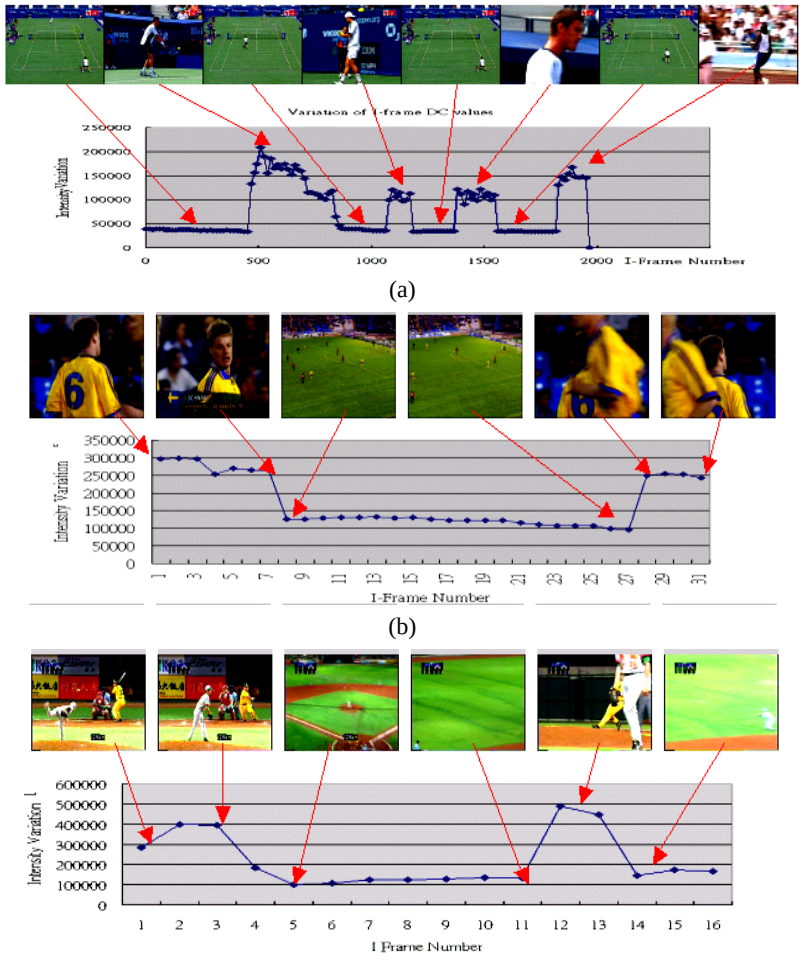


Fig. 2. Variation of I-frame DC value (a) tennis; (b) football; (c) baseball.

caption examination frame-by-frame, caption frames should be detected first. The details of caption frame detection are described in section 3.1 and the approach of closed caption localization is shown in section 3.2. Section 3.3 presents the approach of font size differentiation.

3.1 Caption Frame Detection

Caption frame detection is an essential step for closed caption localization because captions may disappear in some frames and then appear subsequently. Therefore, to avoid detecting closed captions frame-by-frame, we first identify the possible frames in which captions might be present. However, the caption size of closed captions in the shots of court-view is usually very small. Under this circumstance, the change of the AC

energy of the entire frame with the appearance or disappearance of the small caption would not result in significant variation. It means that the variance of the AC energy obtained from an entire frame cannot be used as a measurement of the possibility of the presence of a small caption.

In order to robustly detect closed captions without size constraint, each I-frame is divided into an appropriate number of regions (say R). However, the size of a region should be moderate to reflect the actual variation of appearance or disappearance of small captions. If the size of a divided region were too small, any slight change of color or texture would incur quite prominent variation of AC energy. Accordingly, in order to detect the appearance of super-imposed closed captions in four corner areas as well as in the middle of a frame, the number of regions R here can be set to six. Based on the frame division method, the variance $RVar_{s,i}^{r}$ of AC coefficients of each region r in the ith frame of shot s is computed by

$$RVar_{s,i}^{r} = \sum_{j=1}^{N}\sum_{h/v} AC_{h/v,j}^{2} / N - \left(\sum_{j=1}^{N}\sum_{h/v} AC_{h/v,j} / N \right)^{2}, \quad r = 1, 2, \ldots, R, \tag{4}$$

where $AC_{h/v,j}$ denotes the horizontal AC coefficients from $AC_{0,1}$ to $AC_{0,7}$ and the vertical AC coefficients from $AC_{1,0}$ to $AC_{7,0}$ in region r and N is the total number of blocks in region r.

Using the energy variance $RVar_{s,i}^{r}$ of each region, the method proposed to determine caption frames is illustrated as follows:

For each region r, (5)
If $\mathrm{Diff}(RVar_{s,i+1}^{r}, RVar_{s,i}^{r}) \le -\delta$, captions may disappear in frame $(i + 1)$
If $\mathrm{Diff}(RVar_{s,i+1}^{r}, RVar_{s,i}^{r}) \ge \delta$, captions may appear in frame $(i + 1)$

where $\mathrm{Diff}(RVar_{s,i+1}^{r}, RVar_{s,i}^{r}) = RVar_{s,i+1}^{r} - RVar_{s,i}^{r}$. In the method, $RVar_{s,i+1}^{r}$ of region-r in frame $i + 1$ is compared with $RVar_{s,i}^{r}$ of region-r of frame i. If the difference between $RVar_{s,i+1}^{r}$ and $RVar_{s,i}^{r}$ is larger than a threshold δ (3000), it means the texture of region-r in frame $i + 1$ is more complex than that of region-r in frame i, i.e., closed captions may be superimposed in frame $i + 1$. Similarly, if the difference between $RVar_{s,i+1}^{r}$ and $RVar_{s,i}^{r}$ is smaller than the threshold $-\delta$, the texture of region-r in frame $i + 1$ becomes less complex than that of region-r in frame i, *i.e.*, closed captions in frame i may disappear in frame $i + 1$.

An example of caption frame detection is demonstrated in Fig. 3. Fig. 3 shows the detection of caption frames with small closed captions presented. We can see that the curve of DCT AC variance $RVar_{s,i}^{r}$ of region-1 drops abruptly in the 18th I-frame and rises in 39th I-frame since the scoreboard disappears from the 18th I-frame to the 38th I-frame in the area of region-1 and then appears again in the 39th I-frame.

3.2 Closed Caption Localization

While the caption frames are identified, we then locate the potential caption regions in these frames by utilizing the gradient energy obtained from the horizontal and vertical

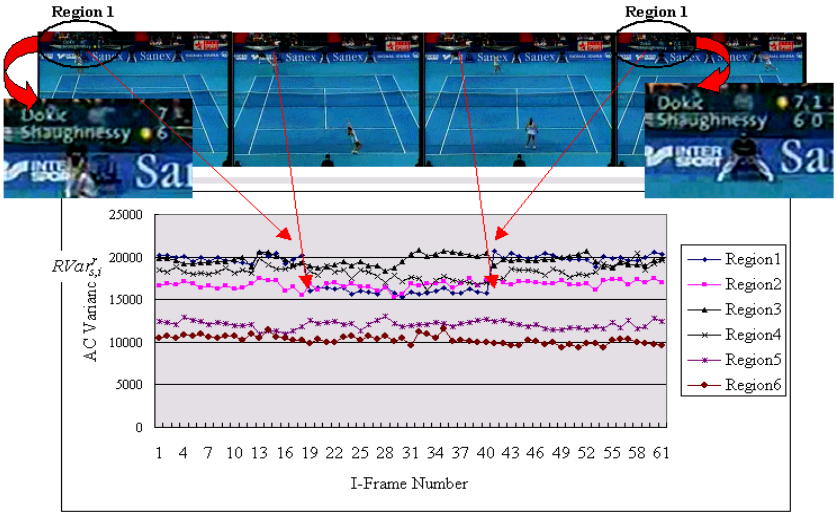


Fig. 3. Demonstration of caption frame detection.

DCT AC coefficients. We can observe the fact that closed captions generally appear in rectangular form and the AC energy in the horizontal direction would be larger than that in the vertical direction since distance between characters is fairly small and the distance between two rows of text is relatively large. Therefore, we assign higher weight to horizontal coefficients than that to vertical coefficients. The weighted gradient energy of an 8×8 block E used as a measurement for evaluating the possibility of a text block can be defined as follows:

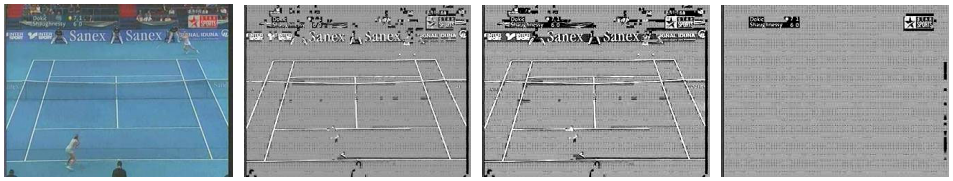
$$E = \sqrt{(w_h E_h)^2 + (w_v E_v)^2}, \tag{6}$$

$$E_h = \sum_{h1 \le h \le h2} \left| AC_{0,h} \right|, \quad h1 = 1,\ h2 = 7,$$

$$E_v = \sum_{v1 \le v \le v2} \left| AC_{v,0} \right|, \quad v1 = 1,\ v2 = 7.$$

If the energy E of a block is larger than a predefined threshold, this block is regarded as a potential text block. Otherwise, the block would be considered as a non-text block and be filtered out without further processing. Besides, in order to save computation cost, we select only 3 I-frames (first, middle and last) as representative frames in a shot for closed caption localization.

The result of closed caption localization is demonstrated in Fig. 4 with w_h set to 0.7 and w_v to 0.3. Although the scoreboard and the trademark in Fig. 4 (b) in the upper part of the frame are all located and indicated, caption regions are fragmentary and some noisy regions remain. Therefore, we adopt a morphological operator in the size of 1×5 blocks to merge fragmentary text regions and the result is demonstrated in Fig. 3 (c). Afterward, the merged text regions are further verified by computing the long-term consistency. For long-term consistency checking, we select another two I-frames as temporal



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

Fig. 4. Illustration of intermediate results of closed caption localization (a) Original frame (b) Closed caption detection; (c) Result after applying morphological operation (d) Result after long-term consistency verification.

reference, the last I-frame of the forward shot (P_F) and the first I-frame of the backward shot (F_B), where T_f, T_m and T_r are the first, middle and the last I-frames of the specific shot. One possible measurement of the long-term coherence of text blocks in potential regions is to check if the text blocks of a potential caption region appear more than half of the time in a shot. That is text blocks appear in more than or equal to three times among the five representative five I-frames.

Here, we exploit the position, intensity and texture information of potential text blocks among these representative I-frames (P_F, T_f, T_m, T_r and F_B) to measure the temporal coherence as defined by

$$C = \frac{\sum_{k=1}^{2}(DC_{Bk} - \overline{DC})(E_{Bk} - \overline{E})}{\sqrt{\sum_{k=1}^{2}(DC_{Bk} - \overline{DC})^2}\sqrt{\sum_{k=1}^{2}(E_{B_k} - \overline{E})^2}}, \quad -1 \le C \le 1 \tag{7}$$

where DC_{B_k} denotes the value of DC coefficient of B_k, $\overline{DC}$ is the average of DC_{B_k} and $DC_{B_{k+1}}$, E_{B_k} represents the weighted gradient energy E of B_k as defined in Eq. (6) and $\overline{E}$ is the average of E_{B_k} and $E_{B_{k+1}}$. A block is characterized by its intensity represented by the DC coefficient and also by its texture obtained from AC coefficients. We compute the correlation C to measure the similarity between two blocks B_k and B_{k+1}, which are in the same corresponding position in their respective frame i and frame $i + 1$. If a value C of a block pair is larger than δ_C, these two blocks are regarded as the same. To estimate the temporal coherence of potential text blocks, we need to compute the pair wise correlation C four times among the 5 representative I-frames. Therefore, a text block is long-term consistent in the specific video shot only when more than half of the times the pair wise I-frames correlation C is larger than δ_C.

The result of long-term consistency checking of text blocks is demonstrated in Fig. 4 (d). We can see that the scoreboard and the trademark are all successfully localized and most of the noise is removed. The proposed closed caption localization can also be applied to other kinds of videos such as baseball, news and volleyball as demonstrated in Fig. 5. In Fig. 5 (a), we can observe that the closed caption primarily composed of Chinese characters is also localized correctly.

(a) (b) (c)

Fig. 5. Examples of closed caption localization (a) baseball; (b) news; (c) volleyball.

3.3 Font Size Differentiation

From Fig. 4 (d), we can notice that the scoreboard in the left upper corner and the trademark in the right upper corner are all successfully detected. Since scoreboards can be used for the content structuring of sports videos, the issue of separating out the captions in the scoreboard is one of our concerns. Hence, the tool – font size detector is proposed to automatically discriminate the font size as a support in the discrimination of scoreboards. To detect the font size, the gradient energy of each text block is exploited. Since a block consisting of characters will have much larger gradient energy than that of a block consisting of blank space, the distance between two character blocks can thus be determined by evaluating the distance between peak gradient values among blocks in a row or column. It means that the font size can be evaluated by measuring the distance between blocks with peak gradient value (i.e., the periodicity of peak values). The gradient energy in the vertical direction instead of horizontal direction is exploited since the blank space in between two text rows is generally larger than that between two letters and hence the variation of gradient energy in the vertical direction would present in more regular pattern.

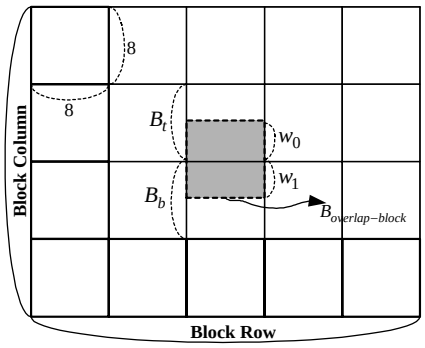


Fig. 6. Overlap-block is interpolated from its two neighboring blocks B_t and B_b.

In addition, to obtain robust periodicity, we compute the DCT coefficients of the 8×8 overlap-block between two neighboring blocks as defined in Eq. (8). A overlap-block $B_{overlap\text{-}block}$ shown in Fig. 6 comprises lower portion of the top neighboring 8×8 block B_t and upper portion of the bottom neighboring block B_b, where I_{w0} and I_{w1} are the identity matrix in the dimension of $w0 \times w0$ and $w1 \times w1$, respectively. More robust results would be achieved if more overlap-blocks are computed and exploited. For example, $w0$ and $w1$ can be respectively set to 1 and 7, 2 and 6, 3 and 5, etc. to acquire more overlap-blocks for more accurate estimation of font size.

$$B_{overlap-block} = \begin{pmatrix} 0 & 0 \\ 0 & I_{w0} \end{pmatrix} B_t + \begin{pmatrix} 0 & 0 \\ I_{w1} & 0 \end{pmatrix} B_b \tag{8}$$

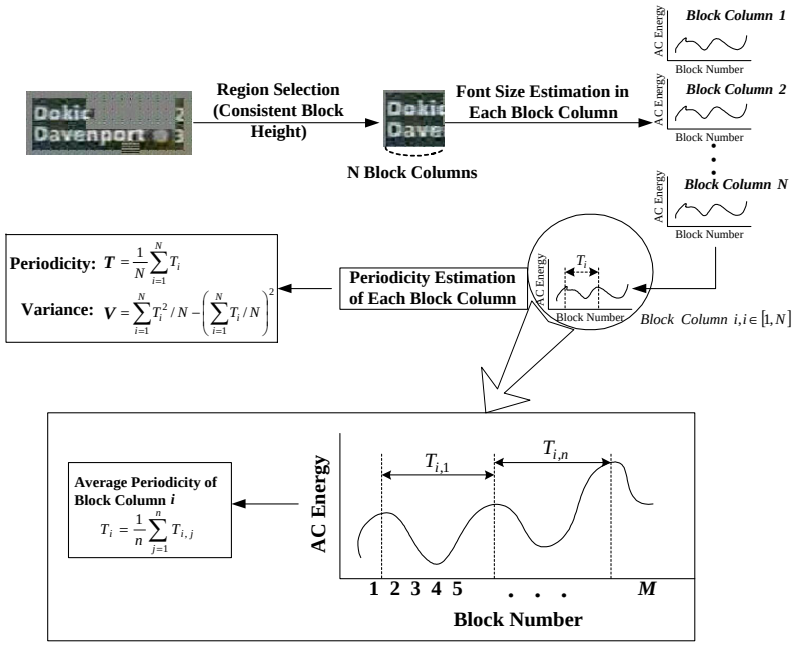


Fig. 7. The proposed approach of font size differentiation in compressed domain.

Fig. 7 shows the proposed approach of font size differentiation, in which the periodicity and variance are estimated for each block column. However, localized closed captions like the example in the top of Fig. 7 may not be complete in shape because some pieces with low gradient energy are filtered out. Therefore, to achieve robust font size differentiation, a region that forms a rectangular in the localized caption is determined for font size computation. Font size differentiation is performed on each block column in the selected region of the closed caption, where a block column depicted in Fig. 7 is defined as a whole column of blocks. While the AC energy of each block is extracted, the curve of the variation of AC energy for each block column is checked to locate each local maximum. We can observe that the region containing the boundary of closed captions would have conspicuous texture variation in the vertical direction and the value of the

gradient energy would be relatively high. Therefore the local maximum of the curve of vertical AC gradient energy is regarded as the boundary of closed captions. While all local maximums are recognized, we must filter out noise and select reliable curve peaks for further verification. Due to the fact that the first and the last local maximums usually reflect the boundary of closed captions, hence we select the first and the last peaks of the curve and compute the average of the value of these two peaks as the threshold adaptively for noise filtering. If the value of a peak is smaller than the threshold, the peak is filtered out. Otherwise, the peak is kept for font size computation. Therefore, the periodicity of each block column T_i is computed by averaging the distance between two peaks of the curve of AC energy. Finally, the average periodicity T and the periodicity variance V of the closed caption are obtained by

$$T = \frac{1}{N}\sum_{i=1}^{N} T_i \tag{9}$$

and

$$V = \sum_{i=1}^{N} T_i^2 / N - \left(\sum_{i=1}^{N} T_i / N\right)^2 \tag{10}$$

where N is the total number of block columns in the selected area of the closed caption.

In order to estimate periodicity of font size more efficiently, we exploit the concept of the projection analysis of a print line [18, 19]. Since it can serve for the detection of blank space between successive letters, we thus compute the horizontal projection profile P_H of each block row P_y by summing up the vertical AC coefficients of the blocks. P_H is defined as follows:

$$P_H = \left\{ P_y \,\middle|\, P_y = \sum_{x=0}^{W-1}\sum_{v=1}^{7} \left|AC_{v,0}\right|,\ \ 0 \le y < H_T - 1,\ \ AC_{v,0} \in B_{x,y} \right\} \tag{11}$$

where H_T is the summation of the number of original blocks (H) and the number of overlap-blocks ($H - 1$) of a block column in an $H \times W$ caption region, and $B_{x,y}$ is a block of coordinate (x, y). By this method, we compute the periodicity T of each localized closed caption once instead of inspection of the periodicity T and of the variance V in each block column. The horizontal projection profile of the scoreboard and the trademark is demonstrated in Fig. 8, where the average periodicity T of the scoreboard and the trademark is about 2 and 3, respectively. Using horizontal projection profile, font size can be detected more efficiently since one curve of AC energy variation needs to be computed for a closed caption.

4. EXPERIMENTAL RESULTS AND VISUALIZATION SYSTEM

4.1 Experimental Results

In the experiment, testing dataset consisted of four kinds of videos including tennis,

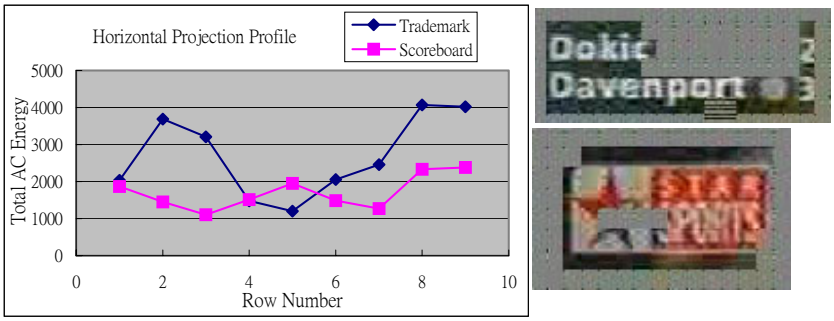


Fig. 8. Horizontal projection profile of DCT AC energy of the scoreboard and the trademark.

baseball, volleyball and news. Two tennis videos selected from US Open and Australia Open, respectively were recorded from the Star-Sport TV channel. A volleyball video was recorded from ESPN TV channel and a baseball game was recorded from VL-Sport TV channel. A news video was selected from MPEG-7 testing dataset. The testing sequences were encoded in MPEG-2 format with the GOP structure IBBPBBPBBPBBPBB at 30 fps. The length of the first tennis video and the news video was about 50 minutes, and the length of the second tennis video was about 30 minutes. The length of the volleyball video and the baseball video was about 40 minutes and 60 minutes, respectively.

The ground truth of the number of caption I-frames of tennis videos shown in Table 1 was 903 and 414, respectively. In Table 2, there were totally 42183 text blocks in the representative frames of tennis video 1 and totally 25680 text blocks of tennis video 2. The number of text blocks in baseball was larger than other videos due to the large superimposed captions. The results of caption frame detection and closed caption localization were evaluated by estimating the precision and recall. The experimental result of caption frame detection was shown in Table 1, and the best performance was achieved in the first tennis video. In tennis video 1, the recall was up to 100% and the precision was about 97%. There were 26 frames of false detection due to the factor that the scoreboard was not presented but some high-texture billboards appear with significant camera movement. In this case, we would detect large variation in the region where billboards were presented. In tennis video 2, the precision of caption frame detection was up to 98% and the recall was about 93%. The number of frames of miss detection was 31 because of the low intensity of the scoreboard in this video sequence. Besides, the color of the scoreboard and that of the tennis court were quite similar and hence it would be more difficult for caption detection in the case of low contrast between closed captions and the background. The worst case in detecting caption frames was presented in the baseball video since the background of several shot types was highly textured, such as the pitching shots and the audience shots. Therefore, when the camera moved, high-textured regions would be considered as the presence of captions. However, recall rate in detecting caption frames in the baseball video remained more than 80%.

The results of closed caption localization were shown in Table 2. In tennis video 1 41030 text-blocks were correctly detected, 347 blocks were falsely detected and 395 text blocks were missed. The precision was about 99% and the recall was about 97%. In tennis video 2, 24624 text blocks were detected, 732 blocks were falsely detected and totally

Table 1. Performance of caption frame detection.

Ground Truth of Caption Frames	Frames of Correct Detection	Frames of False Detection	Frames of Miss Detection	Miss Rate	Precision	Recall
Tennis 1	903	26	0	0%	97%	100%
903						
Tennis 2	383	8	31	7%	98%	93%
414						
Volleyball	578	57	24	4%	91%	96%
602						
Baseball	1290	407	264	24%	76%	83%
1554						
News	873	113	87	9%	88%	91%
960						
Average					90%	93%

Table 2. Performance of closed caption localization after caption frame detected.

Ground Truth of Text Block	Blocks of Correct Detection	Blocks of False Detection	Blocks of Miss Detection	Miss Rate	Precision	Recall
Tennis 1	41030	347	395	1%	99%	97%
42183						
Tennis 2	24624	732	1056	4%	97%	95%
25680						
Volleyball	27353	4087	2250	8%	87%	92%
28122						
Baseball	269792	63270	26676	9%	81%	91%
296405						
News	192016	23732	10080	5%	89%	95%
201616						
Average					91%	94%

1056 text blocks were missed. Hence, the precision and recall of tennis video 2 was 97% and 95%, respectively. Some text blocks were missed since the background of the closed caption was transparent and would change with the background while camera moved. In this case, if the texture of the background was similar to the closed caption, the letters of captions cannot reflect the large variation in gradient energy and some text blocks would be missed. The precision rate of the baseball video in detecting text blocks was 81% due to the highly textured background. However, the recall rate was up to 92% since the temporal consistency was exploited to filter noise. Most of the blocks, which appeared for a short duration and their spatial positions were not consistent, were regarded as noise and were thus eliminated. The good performance was due to the reason that the weighted horizontal-vertical AC coefficients were exploited and the long-term consistency of the closed caption over consecutive frames was considered.

By applying the proposed approach of font size differentiation, we can automatically discriminate the font size either in a closed caption or in different ones. Therefore, this designed tool can be used as the closed caption filter to recognize and select those of interest, once the user indicates the targeted font size of closed captions. Moreover, researches [22-25] focusing on video OCR indicate that a bottleneck for recognizing characters was due to the variation of text font and size. In addition, to make learning data for the filter of character extraction, the size of the filter, which was defined to include a line element of characters, should be determined. Since the size of the line element strongly depends on the font size, it was possible to design a filter that can enhance the line elements dynamically with widely varying font sizes when the font size in the localized captions was known. Consequently, the tool – font size differentiation can be exploited to be a pre-processing tool for video OCR.

4.2 The Prototype System of Video Content Visualization

With the successful localization of the super-imposed scoreboard in sports videos, video content can be visualized in a compact form by constructing the hierarchical structure. Taking tennis as an example, the structured contents composed of scoreboards and the related can be combined with the detected tennis semantic events [20], such as baseline rally, serve and volley and passing shot. Each competition shots can be annotated using the type of corresponding event and can be labeled exploiting the scoreboard. Consequently, the information of the type of events, the boundary of events, the key frame of events and the result of the event – the scoreboard can be used in the Highlight Level Description Scheme shown in Fig. 9 to support users to efficiently browse videos by viewing the images of scoreboards and the important text information of semantic events. The name of highlight corresponded to the type of tennis event, the descriptor of video segment locator was described by the event boundary and the position of the key frame in the video sequence was used for the key image locator. The key image locator for scoreboard indicates the time point in the video sequence.

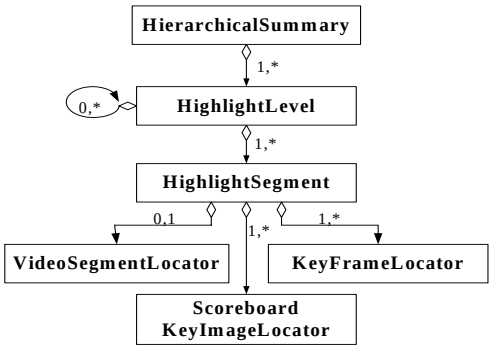


Fig. 9. Hierarchical summary description scheme [21].

The table of video content was composed of the original video sequence in the top level, the scoreboard of a set, the scoreboard of a game and the key frame of one point.

The user interface of the prototype system was shown in Fig. 10 and two areas of "Playback" and "Visualization" were present in the left and the right side, respectively. Initially, the key frame of the original video sequence and the scoreboards of sets were exhibited. While users can click the symbol "[+]" as the arrow lines indicated, the system would show the scoreboards of the corresponding games. Users can select which game they want to watch according to the scoreboards of the games and click the symbol "[+]" for more detail. Each point of the game was represented by its key frame. Users can view the point by clicking the corresponding key frame and the shot of the point would be displayed in the "Playback Area". By exploiting the system of video content visualization, users can efficiently browse video sequences. Since the length of a sport video was up to one or two hours generally, the system thus provided a compact and brief overall view of the match for users by exhibiting the textual information of the scoreboards hierarchically.

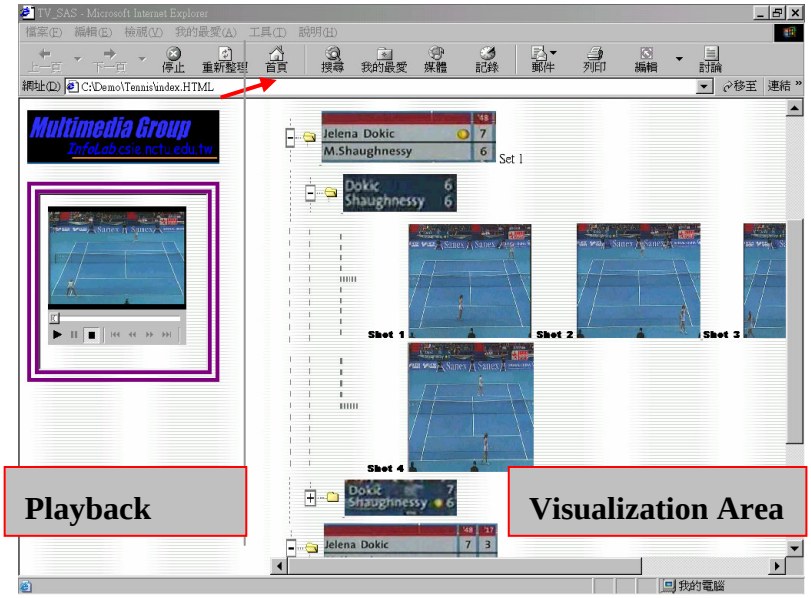


Fig. 10. Video content visualization system was composed of two areas – "Playback" and "Visualization". The hierarchical structure of the scoreboards was shown while the user clicks the symbol "[+]".

5. CONCLUSION

In the paper, we have proposed a novel mechanism to detect temporal boundaries, identify meaningful shots and then build a compact table of video content. GOP-based video segmentation was used to efficiently segment videos into shots. To efficiently detect closed captions, color-based shot identification was proposed to identify shots of interest, especially for sports videos. Caption frames were detected in the shots of interest using the compressed data in MPEG videos. Then caption frames instead of every frame

were selected as targets for detecting closed captions based on the long-term consistency without size constraint. While closed captions were localized, we differentiate the font size of closed captions based on the horizontal projection profile of AC gradient energy obtained from both the original blocks and the interpolated sub-blocks. The proposed tool – font size detector can thus be used as a prefilter to effectively eliminate uninterested closed captions and avoid most of the extremely time consuming post-processing of localized captions. Finally, having the proposed mechanism of high-level video structuring, one can browse videos in an efficient way through a compact table of content.

REFERENCES

1. H. Wang and S. F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, 1997, pp. 615-628.
2. Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 385-392.
3. H. Luo and A. Eleftheriadis, "On face detection in the compressed domain," in *Proceedings of the ACM Multimedia*, 2000, pp. 285-294.
4. Y. Zhang and T. S. Chua, "Detection of text captions in compressed domain video," in *Proceedings of the ACM Multimedia Workshop*, 2000, pp. 201-204.
5. S. W. Lee, Y. M. Kim, and S. W. Choi, "Fast scene change detection using direct feature extraction from MPEG compressed videos," *IEEE Transactions on Multimedia*, Vol. 2, 2000, pp. 240-254.
6. X. Chen and H. Zhang, "Text area detection from video frames," in *Proceedings of the 2nd IEEE Pacific Rim Conference on Multimedia*, 2001, pp. 222-228.
7. J. Nang, O. Kwon, and S. Hong, "Caption processing for MPEG video in MC-DCT compressed domain," in *Proceedings of ACM Multimedia Workshop*, 2000, pp. 211-214.
8. S. Y. Lee, J. L. Lian, and D. Y. Chen, "Video summary and browsing based on story-unit for video-on-demand service," in *Proceedings of the International Conference on Information and Communications Security*, 2001.
9. J. L. Mitchell, W. B. Pennebaker, C. E. Fogg, and D. J. LeGall, *MPEG Video Compression Standard*, Chapman and Hall, New York, 1997.
10. J. Meng, Y. Juan, and S. F. Chang, "Scene change detection in a MPEG compressed video sequence," in *Proceedings of the IS & T/SPIE Symposium Proceedings on Electronic Imaging: Science & Technology*, Vol. 2419, 1995, pp. 14-25.
11. H. J. Zhang, C. Y. Low, S. W. Smoliar, and J. H. Wu, "Video parsing and browsing using compressed data," *Multimedia Tools and Applications*, Vol. 1, 1995, pp. 89-111.
12. H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Transactions on Image Processing*, Vol. 9, 2000, pp. 147-156.
13. J. C. Shim, C. Dorai, and R. Bollee, "Automatic text extraction from video for content-based annotation and retrieval," in *Proceedings of the 14th International Conference on Pattern Recognition*, 1998, pp. 618-620.

14. J. Ohya, A. Shio, and S. Akamastsu, "Recognizing characters in scene images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, 1994, pp. 214-220.
15. U. Gargi, S. Antani, and R. Kasturi, "Indexing text events in digital video databases," in *Proceedings of the 14th International Conference on Pattern Recognition*, 1998, pp. 916-918.
16. S. Kannangara, E. Asbun, R. X. Browning, and E. J. Delp, "The use of nonlinear filtering in automatic video title capture," in *Proceedings of the IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing*, 1997.
17. V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: an automatic system to detect and recognize text in images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, 1999, pp. 1224-1229.
18. S. W. Lee and D. S. Ryu, "Parameter-free geometric document layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 1240-1256.
19. R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, 1996, pp. 690-706.
20. D. Y. Chen and S. Y. Lee, "Motion-based semantic event detection for video content descriptions in MPEG-7," in *Proceedings of the 2nd IEEE Pacific Rim Conference on Multimedia*, 2001, pp. 110-117.
21. ISO/IEC JTC1/SC29/WG11/N3964, "MPEG-7 multimedia description schemes XM (v7.0)," Singapore, March 2001.
22. W. Qi, L. Gu, H. Jiang, X. R. Chen, and H. J. Zhang, "Integrating visual, audio and text analysis for news video," in *Proceedings of the International Conference on Image Processing*, Vol. 3, 2000, pp. 520-523.
23. D. Chen, K. Shearer, and H. Bourlard, "Text enhancement with asymmetric filter for video OCR," in *Proceedings of the 11th International Conference on Image Analysis and Processing*, 2001, pp. 192-197.
24. T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for digital news archive," in *Proceedings of the IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998, pp. 52-60.
25. Y. Ariki and K. Matsuura, "Automatic classification of TV news articles based on Telop character recognition," in *Proceedings of the IEEE International Conference on Multimedia Systems*, 1999, pp. 148-152.
26. D. Y. Chen, S. J. Lin, and S. Y. Lee, "Motion activity based shot identification and closed caption detection for video structuring," in *Proceedings of the 5th International Conference on Visual Information System*, 2002, pp. 288-301.
27. A. K. Jain and B. Yu, "Automatic text location in images and video frames," *Patter Recognition*, Vol. 31, 1998, pp. 2055-2076.
28. S. W. Lee, D. J. Lee, and H. S. Park, "A new methodology for grayscale character segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, 1996, pp. 1045-1050.

Duan-Yu Chen (陳敦裕) received the B.S. degree in Computer Science and Information Engineering from National Chaio Tung University, Taiwan in 1996, the M.S. degree in Computer Science from National Sun Yat-sen University, Taiwan in 1998, and the Ph.D. degree in Computer Science and Information Engineering from National Chiao Tung University, Taiwan in 2004. His research interests include computer vision, video signal processing, content-based video indexing and retrieval and multimedia information system.

Ming-Ho Hsiao (蕭銘和) received the received the B.S. degrees in Computer Sciences and Information Engineering from Fu Jen Catholic University, Taiwan in 2000. He received the M.S. degree in Computer Sciences and Information Engineering from National Chiao Tung University, Taiwan, where he is currently pursuing the Ph.D. degree. His research interests are content-based indexing and retrieval and distributed multimedia system, in particular, media server architecture and peer-to-peer system.

Suh-Yin Lee (李素瑛) received the B.S. degree in Electrical Engineering from National Chiao Tung University, Taiwan in 1972, the M.S. degree in Computer Science from University of Washington, U.S.A., in 1975, and the Ph.D. degree in Computer Science from Institute of Electronics, National Chiao Tung University. Her research interests include content-based indexing and retrieval, distributed multimedia information system, mobile computing, and data mining.