Improvement of Commercial Boundary Detection Using Audiovisual Features

Jun-Cheng Chen[1], Jen-Hao Yeh[1], Wei-Ta Chu[1], Jin-Hau Kuo[1], and Ja-Ling Wu[1,2]

[1] Department of Computer Science and Information Engineering,
National Taiwan University
[2] Graduate Institute of Networking and Multimedia, National Taiwan University,
No.1, Sec 4, Roosevelt Rd., Taipei, Taiwan 106, ROC
{pullpull, littleco, wtchu, david, wjl}@cmlab.csie.ntu.edu.tw

Abstract. Detection of commercials in TV videos is difficult because the diversity of them puts up a high barrier to construct an appropriate model. In this work, we try to deal with this problem through a top-down approach. We take account of the domain knowledge of commercial production and extract features that describe the characteristics of commercials. According to the clues from speech-music discrimination, video scene detection, and caption detection, a multi-modal commercial detection scheme is proposed. Experimental results show good performance of the proposed scheme on detecting commercials in news and talk show programs.

1 Introduction

The devices of digital video recording have become very common nowadays. Popularity of various devices for capturing and storage boosts the applications of time shifting recording on digital TV and broadcasting videos. Automatic detection and removal of commercials play an important role in digital home applications because commercials are often unwanted information when audiences browse recorded videos.

There are different ways to deal with the commercial detection problems. In the literature, for removing commercial segments in featured films, Lienhart et al. [1] use a combination of video features and commercial recognition scheme to achieve this goal; Duygulu et al. [3] mix audio features and detect video duplicate sequences to remove commercials in TV news programs; Juan et al. [4] use recognition of key frames on shots; Albiol et al. [6] apply template matching on TV station identification logo occurrence; Liu et al. [7] use adaboost with time constraint to detect commercial breaks.

The methods described above work well in conventional TV programs. However, we found that most of the prescribed approaches don't work well due to the advance of video technology. Recently, the traditional black frame insertion and the "always-on" assumption of TV station logo during commercials are not always valid. On the other hand, commercials have some intrinsic characteristics. For example, commercials often have music and have less caption ratio than programs. Hence, from clues of speech-music discrimination, video scene detection, and caption detection, an efficient three-level commercial detection scheme is proposed. It consists of detecting

Y.-S. Ho and H.J. Kim (Eds.): PCM 2005, Part I, LNCS 3767, pp. 776 – 786, 2005.

commercial break candidates, boundary refinement, and outlier (i.e. non-commercial parts) removing.

The organization of this paper is as follows. In Section 2, we address the problem and provide an overview of our system. Section 3 delineates how to find commercial-break candidates. Section 4 and Section 5, respectively, describe how to refine the resulting boundaries between commercials and regular TV programs and how to remove outliers. Section 6 presents the experimental results, and finally, Section 7 concludes this paper.

2 Problem Definition and System Overview

2.1 Problem Definition

TV commercial detection has been studied for commercial skipping or other applications for years. Actually, the problem is similar to partition the video data into consistent scenes semantically, which are then classified into programs or commercials.

A commercial break consists of one or more pieces of commercials and there are several commercial breaks in a program video. The problem of commercial detection is to label all possible commercial breaks and find the exact boundaries of them.

2.2 System Overview

The proposed system is illustrated in Fig.1. First of all, the recorded TV programs pass through the shot change detector. The results of shot change detector are used to calculate cuts and strong cuts (will be defined in the next section) per minute, which represent the frequency of shot changes. And then, we use these two features to label candidates of boundaries of commercial breaks. In the boundary refinement stage, the results of speech-music discriminator, video scene detection and caption text detection further characterize commercials and programs. The exact boundaries are

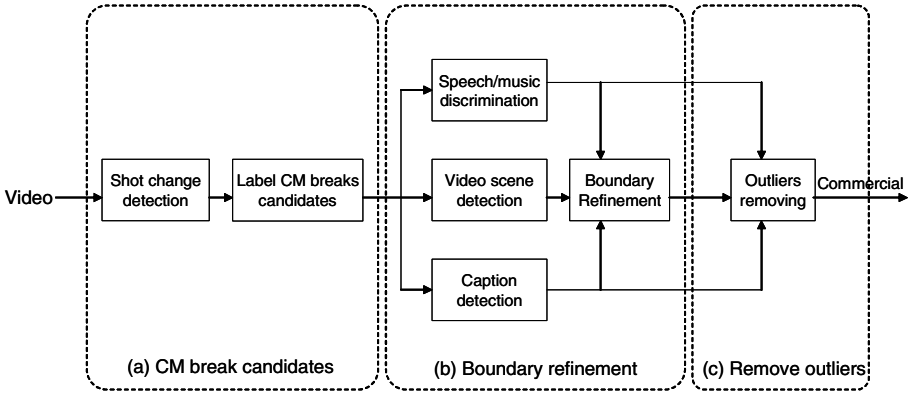

Fig. 1. The block diagram of the proposed commercial detection system

then refined based on the so-obtained characteristics. After boundary refinement, we remove outliers by some observations, such as low average caption ratio and high average music ratio during commercials.

3 Label the Commercial Breaks Candidates

3.1 Cuts and Strong Cuts

A "cut", shot change, shows discontinuity between two individual continuous camera shots. A shot change at frame index t is detected based on the difference of color histograms. In commercial production, directors setup dominant colors according to topics of commercials and make commercials prominent. Different commercials with different topics and styles, therefore, have drastic color difference. To capture this phenomenon, we introduce the feature, "strong cut", to indicate a cut with higher color-histogram difference. The definitions of cut and strong cut are formulated as follows:

$$Type\_of\_Cut(t) = \begin{cases} Strong\ cut, & D(t-1,t+1) > Th_{struct} \\ Cut, & (D(t-1,t+1) < Th_{struct}) \wedge (D(t-1,t+1) > Th_{cut}) \\ No\ shot\ change, & D(t-1,t+1) < Th_{cut} \end{cases} \quad (1)$$

and

$$D(t\text{-}1,t+1) = \frac{\sum_{i=1}^{N} |H_{t+1}(i) - H_{t-1}(i)|}{N} \quad (2)$$

Where H_t denotes the YUV color histogram of the t-th frame, N is the total number of bins (e.g., N=64). $D(t\text{-}1, t+1)$ indicates the average histogram difference between the $(t\text{-}1)$-th and the $(t+1)$-th frames. Two thresholds, Th_{struct} and Th_{cut} ($Th_{struct} > Th_{cut}$), are defined to classify shot changes into cut or strong cut.

3.2 Cuts Per Minute and Strong Cuts Per Minute

Usually, audiences are interested in TV programs rather than commercials. In order to catch audiences' attention, commercial producers must make their videos as interesting as possible. It's well known that videos with high motion and frequent shot changes are more attractive than static and smooth scenes. It's also true that almost all commercials have much more shot changes than that of programs. Hence, the idea of "cuts per minute", which is the number of cuts within a minute, is used to model the characteristics of commercials. We also adopt "strong cuts per minute" [1], which is the number of strong cuts within a minute, into account. The value of strong cuts per minute reflects the number of pieces of commercials in a commercial break, and also indicates the impression of dynamics of activity.

Cuts and strong cuts per minute are calculated from video clips through applying a sliding window of length W_s and with T_s overlapping to its neighbors. In our work, W_s is set to 30 second, and T_s is set to 10 seconds. We show examples of strong cuts

and cuts per minute in Fig.2. In the figure, we see the typical trend that commercials often have more shot changes than programs and so are strong cuts.

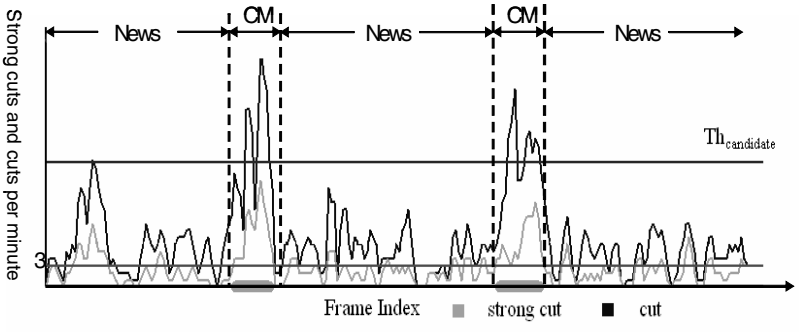

Fig. 2. Examples of strong cuts and cuts per minute

3.3 Mark Commercial Breaks Candidates

We modify the approach proposed in [1] to label commercial-break candidates. As shown in Fig. 2, the segments with large strong cuts and cuts per minute are selected as commercial breaks candidates. First, the segments with strong cuts per minute less than 3 are rejected. Then, the survived segments are further rejected if their cuts per minute do not exceed a threshold $Th_{candidate}$. Generally, the number of commercial breaks is restricted by the law (Enforcement Rules of the Radio and Television Act [11]). For instance, the law regulates that there should be 2 commercial breaks in a 40 minutes long news program. Thus, we can determine $Th_{candidate}$ from this domain knowledge. However, because we don't want to miss any possible cases, a loose threshold is set to allow more false alarms, which will be filtered out by later processes.

4 Boundary Refinement

After previous steps, we just get rough boundaries of commercial breaks. Some program segments may be included in the commercial breaks candidates. To refine the boundaries, we try to capture the characteristics of program-commercial and commercial-commercial changes through the clues of audio, video, and caption texts.

4.1 Audio Features

There are often volume changes between commercials and programs. For example, when commercials start or finish, there is often a very short duration of silence. Moreover, commercials often have background music. So we perform speech-music discrimination proposed in [8], which is constructed on the basis of RMS's (root mean square values) and ZCR's (zero crossing rates) of audio samples. It partitions audio data into segments and classifies each segment as speech or music. In addition to the label of each segment, normalized RMS Matusita distance, the intermediate

output of [8], can be used to detect relative significant volume change because it is calculated and normalized with neighbors within a local window.Thus, it's suitable for detecting transitions between commercials and programs.

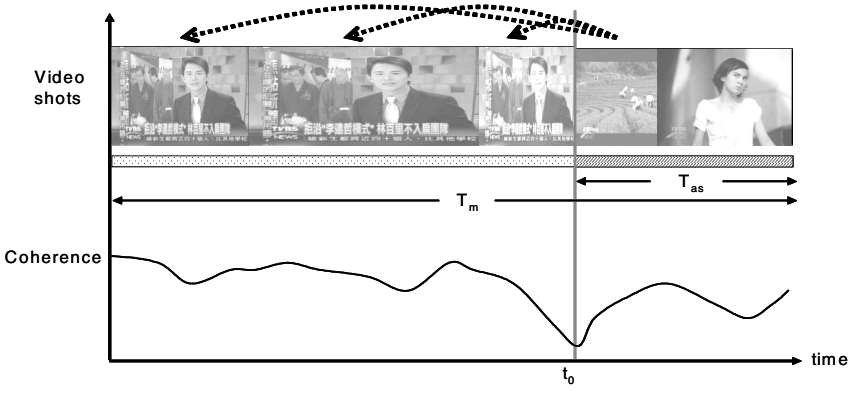

Fig. 3. The computation of video coherence between keyframes. If the shots in T_{as} are quite different to the shots in the rest of the memory, a video scene change is declared to occur at time t_0.

4.2 Video Scene Detection

We assume that the coherence of colors between commercials and programs and between different commercials is low. To discriminate styles of video scenes, we take the scheme of computable video scene with memory model addressed in [2] to simulate the "remembrance" of human. The coherence among shots in the attention span, T_{as}, is computed. Fig.3 illustrates the ideas of video coherence and memory model. We can see that the video scene boundaries are often occurred at the local minimums of the coherence curve. We setup T_{as} as 8 seconds and T_m as 24 seconds for later experiments.

If there is a local minimum in the curve of coherence at time t_0, a "video scene boundary" is declared. The middle frame in a shot is selected as the shot's keyframe in order to avoid the noises generated from special scene transition effects.

4.3 Caption Feature

We take the method developed in [9] to acquire the sizes and the locations of caption texts. In our observations, commercials usually use overlay texts to convey messages to audiences. But these overlay texts often appear only in a short duration at the same location. On the contrary, news or talk show programs often have headlines or captions on the bottom. According to the observation, we focus only on the video texts appeared at the bottom. We define the bottom area as shown in Fig.4 (a), with size frame-width $\times$(0.3 * frame-height). Caption ratio is defined as the number of 8 $\times$8-pixel blocks that display caption in the bottom area. We use it to indicate the caption characteristics of each segment (c.f. Fig.5). That is,

$$cap\_ratio(i) = \text{the number of } 8\times 8\text{-pixel blocks that display caption in the bottom area.}$$

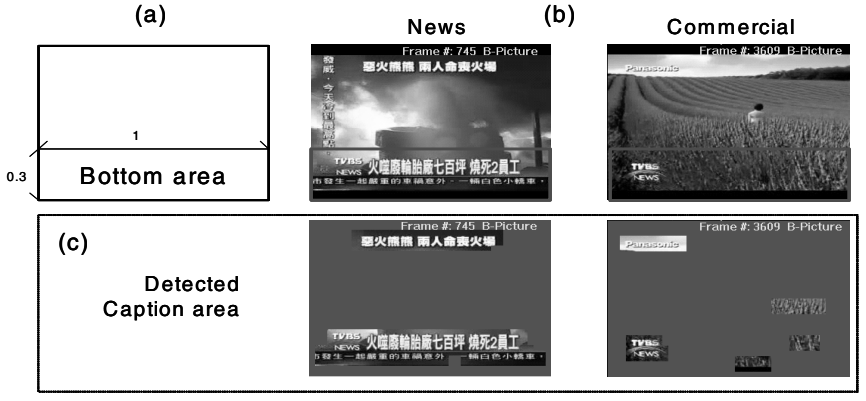

Fig. 4. (a) Definition of bottom area of each frame, (b) keyframes of news and commercials, and (c) the result of caption detection of news and commercial

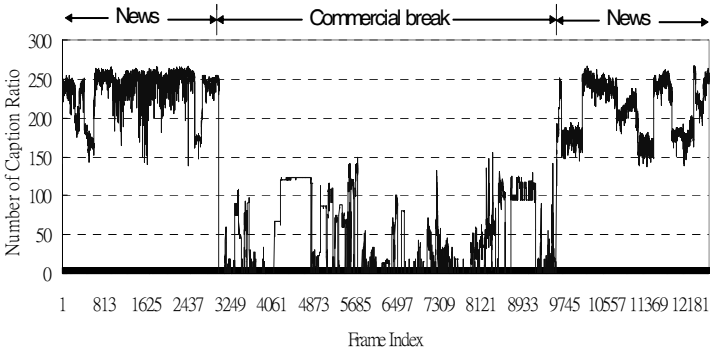

Fig. 5. The caption ratios for each of the frames of an MPEG encoded video (with resolution 320x240). It follows that the caption ratio is lower in commercials than that in news.

4.4 Mark the Exact Boundaries

Transitions of commercials and programs induce scene changes. Large volume changes and caption-ratio changes often occur with scene changes. Thus, we apply our algorithm to commercial-break candidates, and then several video scene boundaries in the suspicious segments are generated. Two types of clues are used to mark the exact boundaries. The first clue is the maximum difference of caption ratio around the video scene boundaries. The second one is volume change. If the *j*-th video scene boundary satisfies one of the following conditions, we'll retain it. Otherwise, we'll remove it from candidates. That is,

$$\text{Condition1: } Dcap(j) > Th_{diff\_of\_cap} \text{ and } \min_{i=L-w,..,L+w}\{cap\_ratio(i)\} < Th_{cap\_ratio}.$$

$$\text{Condition2: } Vc(j) > Th_{vc} \text{ and } \min_{i=L-w,...,L+w}\{cap\_ratio(i)\} < Th_{cap\_ratio}.$$

$$\text{and } Dcap(j) = \max_{i=L',...,L''}\{cap\_ratio(i)\} - \min_{i=L',...,L''}\{cap\_ratio(i)\}. \quad (3)$$

Where we define *Dcap(j)* as the maximum difference of caption ratios around the *j*-th video scene boundary. The window size, *w,* is set to 15 frames and *L*, *L'*, and *L''* are respectively the frame index of *j*-th, *(j-1)*th, and *(j+1)*th video scene boundaries. *Vc(j)* is the normalized Matusita distance of RMS values within the duration (3 seconds), where we search volume changes around the *j*-th video scene boundary. The thresholds $Th_{diff\_of\_cap}$, Th_{cap_ratio}, and Th_{vc} are empirically defined.

Caption ratio is often higher in news programs. In addition to checking *Dcap(j)*, we have to ensure that caption ratio of the *j*-th video scene boundary is less than Th_{cap_ratio}. Boundaries of the first and the last survived video scenes will be chosen as the boundaries of commercial breaks. However, it's possible that the detected commercial break contains a tiny news story, such as weather reports at the end of news, as shown in Fig. 6(a). Thus, if the duration of the detected commercial break is greater than an empirical value (5 minutes), we focus on the detected duration and examine it again by applying condition 1 and condition 2, as shown in Fig. 6(b).

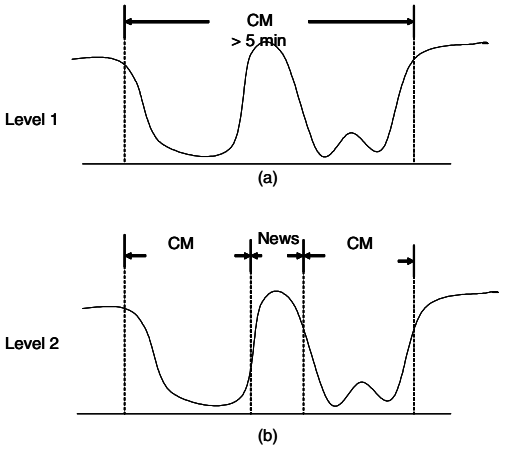

Fig. 6. Illustrations of boundaries refinement

5 Outliers Removing

Some parts of news programs with many shot changes or high motion may be regarded as commercial-boundary candidates at the steps described in Section 3. We observed that commercials often have background music and few long-lasting texts on the bottom. We calculate the average caption ratios and music ratios of segments through results of previous steps. If the average caption ratio of the segment is greater than an empirical threshold or the music ratio of the segment is less than another empirical threshold, this segment will be recognized as some parts of the news program, the so called outliers. These thresholds can be determined by checking the statistical characteristics of different video segments. Fig. 7 shows the histograms of average caption ratio and music ratio. We can see the significant difference between the distributions of two features of commercial breaks and programs, which include talk shows and news programs. We apply outlier removing at the final step because

we will get more accurate boundaries of commercial breaks after boundary refinement and this final step is helpful to avoid taking commercial breaks as program segments.

6 Experimental Results

Our experiments are conducted on news and talk show program videos. Our test videos are taken from five different channels in Taiwan. There are 8 news and 2 talk show program videos in our testing dataset and their frame rates are 29.98 frames per second. Their total length is about 7.5 hours. There are 33 different commercial breaks in this dataset. The ground truth consists of the precise timestamps of the commercial breaks for each video sequence.

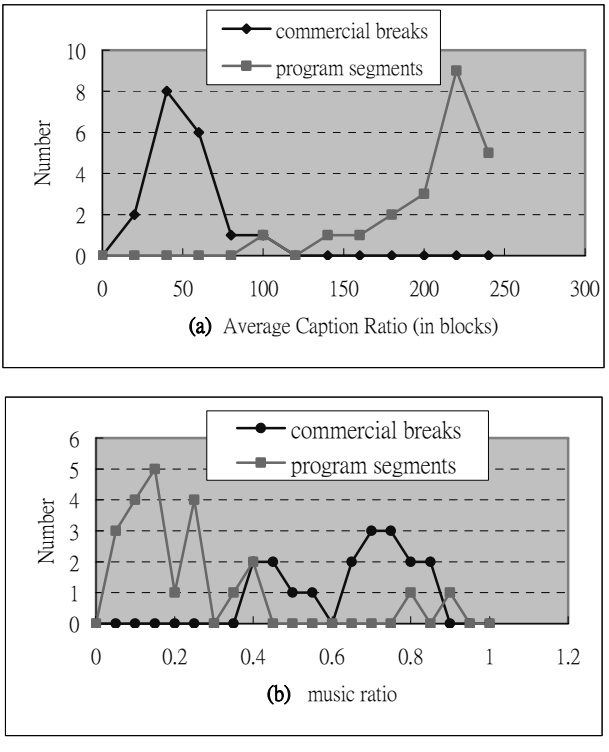

Fig. 7. The histograms of (a) average caption ratio, and (b) the music ratio. Both of them are generated by 18 commercial breaks and 23 program segments from a 5-hour news video.

To estimate the performance of our algorithm, as depicted in Fig. 8, we need to determine the number of frames of commercials correctly identified (true positive: TP), the number of frames of commercials missed (false negative: FN), and the number of frames of programs recognized as commercials (false positive: FP).

We use recall, precision, and F1 metric to show the ability of our algorithm. The definitions of recall, precision, and F1 are shown as follows:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{4}$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{5}$$

$$\text{F1} = 2 \times \text{Recall} \times \text{Precision} / (\text{Recall} + \text{Precision}) \tag{6}$$

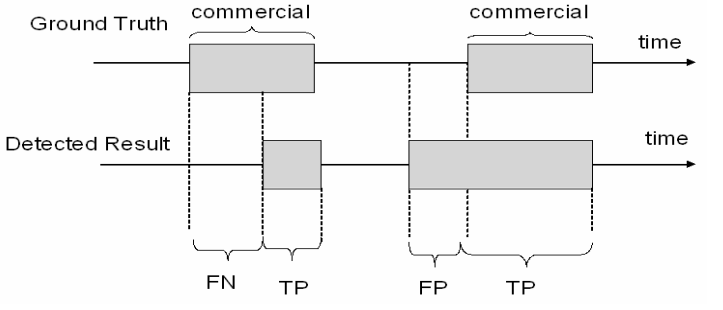

Fig. 8. An example of false negative(FN), true positive(TP), and false positive

Table 1. Experimental results with parameters W_s = 30(sec), T_s = 10(sec), T_m = 24(sec), and T_{as} = 8(sec)

Sequences	Recall	Precision	F1
News 1 (30 minutes)	95.70%	96.56%	0.9612807
News 2 (30 minutes)	95.60%	99.90%	0.9770271
News 3 (60 minutes)	76.50%	81.40%	0.788 7397
News 4 (40 minutes)	100%	96.00%	0.9795918
News 5 (40 minutes)	97.66%	98.70%	0.9817724
News 6 (30 minutes)	82.20%	79.6%	0.8084851
News 7 (45 minutes)	100%	82.50%	0.9041096
News 8 (60 minutes)	99.90%	94.60%	0.9717779
Talk show 1 (60 minutes)	99.4%	98.3%	0.9884693
Talk show 2 (60 minutes)	95.2%	97.6%	0.9638506
Average	94.22%	92.52%	0.948485

The recall is the percentage of commercials that our algorithm successfully detects, in terms of duration. The precision is the percentage of actual commercials within what the algorithm detects as commercials. F1 takes recall and precision into consideration

at the same time. The corresponding results are listed in Table 1. News 3 and News 6 are taken from the same channel. The results of them have worse performance than others because one of our assumptions, low caption ratio during commercial breaks, is not always valid in these channels. In most cases, our method performs well and averagely achieves 94.22% recall, 92.52% precision and 0.948 F1 metric.

In this work, we further devote our efforts to refine boundaries of commercial breaks. We compare the detection performance of the newly proposed method with our previous work [10], which doesn't integrate audio and caption information. In order to estimate the performance of boundary refinement, average frame difference is calculated as follows:

$$AvgD_j = \frac{1}{N}\sum_{i=1}^{N} fd_i \tag{7}$$

where fd_i is the frame difference between the i-th detected commercial break and the ground truth, N is the total number of detected commercial breaks in the sequence, and $AvgD_j$ is the average frame difference of the j-th test sequence.

Table 2 shows the average frame difference in 11 sequences. Overall, the proposed method has superior performance and achieves 1641 frame-difference gain in average. Experimental results in Table 2 reveal that our algorithm can determine more precise boundaries.

Table 2. Average frame difference of (a) our proposed algorithm and (b) the algorithm proposed in [10] with parameters W_s = 30(sec), T_s = 10(sec), T_m = 24(sec), and T_{as} = 8(sec)

Sequences	Average Frame Difference	
	(a)	(b)
News 1	482	5443
News 2	248	956
News 3	1494	2368
News 4	374	4927
News 5	244	1516
News 6	2035	3465
News 7	1140	563
News 8	504	770
Talk show 1	148	2172
Talk show 2	417	1314
Average	708	2349

7 Conclusion

In this paper, by combining speech-music discrimination, video scene segmentation and caption detection, an effective commercial detection scheme is proposed for news

and talk show program videos. From our experiments, the video scene segmentations with caption-change and volume-change detections do help for deciding the exact boundaries of commercial breaks in news videos. Outlier removing based on the average caption ratio and music ratio also performs well. The proposed framework is expected to be applied to other programs with notable caption presentation. However, due to the diversity of the television commercials, of course, the proposed method is not perfect in every situation. More features or model-based approaches should be investigated further in the future.

References

1. Rainer Lienhart, Christoph Kuhmünch and Wolfgang Effelsberg: On the Detection and Recognition of Television Commercials, Proceedings of IEEE International Conference Multimedia Computing and Systems, (1997) 509-516
2. Hari Sundaram and Shih-Fu Chang: Computable Scenes and Structures in Films, IEEE Transactions on Multimedia, Vol. 4, No.4, (2002) 482-491
3. P. Duygulu, M.-Y. Chen, and A. Hauptmann: Comparison and Combination of Two Novel Commercial Detection Methods, Proceedings of IEEE International Conference on Multimedia and Expo, Vol. 2, (2004) 1267 – 1270
4. Juan Maía Sánchez, Xavier Binefa, Jordi Vitrià, and Petia Radeva: Local Color analysis for Scene Break Detection Applied to TV Commercials Recognition, Proceedings of 3rd. International Conference on VISUAL'99, (1999) 237-244
5. B. Satterwhite and O. Marques: Automatic detection of TV commercials**,** IEEE Potentials, Vol. 23, No. 2, (2004) 9 – 12
6. A. Albiol, M.J. Ch, F.A. Albiol, L. Torres: Detection of TV commercials**,** Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, (2004) 541-544
7. Tie-Yan Liu, Tao Qin, Hong-Jiang Zhang: Time-constraint boost for TV commercials detection**,** IEEE International Conference on Image Processing, Vol. 3, (2004) 1617 – 1620
8. C. Panagiotakis and G. Tziritas: A speech/music discriminator based on RMS and Zerocrossings, IEEE Transactions on Multimedia, Vol. 7, No. 1, (2005) 155 – 166
9. Chin-Fu Tsao, Yu-Hao Chen, Jin-Hau Kuo, Chia-Wei Lin, and Ja-Ling Wu: Automatic Video Caption Detection and Extraction in the DCT Compression Domain, accepted by Visual Communications and Image Processing, (2005)
10. Jen-Hao Yeh, Jun-Cheng Chen, Jin-Hau Kuo, and Ja-Ling Wu: TV Commercial Detection in News Program Videos, accepted by IEEE International Symposium on Circuits and Systems, (2005) 4594-4597
11. Enforcement Rules of the Radio and Television Act, article 34, http://www.gio.gov.tw/taiwan-website/1-about_us/6-laws/ra8.htm