Novel TV Commercial Detection in Cookery Program Videos

Venkatesh .N, Rajeev .B, and M. Girish Chandra

***Abstract*— In this paper, we present a simple but very effective method for detecting commercial boundaries in a cookery video program. Our approach to this problem is to first identify the best audio feature to detect commercial boundary for cookery videos with and without background music. Further, we boost the confidence level of detecting boundary using novel method of logo (program name) video frames detection indicating the start and end of the commercial at the boundaries detected by audio. Experiments show that combining audio features and video clues yields good recall and precision on a test set videos from different Indian television broadcast data.**

***Index Terms*— Commercial Detection, combining Audio and visual features, cookery video show.**

I. Introduction

The recent technologies such as set-top boxes coupled with the reduction of costs for digital media storage have led to the introduction of recording option for all kinds of television (TV) programs.

In this context, audiovisual analysis tools that help the user to manage the huge amount of data are very important in a very competitive market to introduce the novel recording devices. Among other analysis tools, detection of TV advertisements is a topic with many practical applications. For instance, from the point of view of a TV end-user, it could be useful to avoid commercials in personal recordings.

Existing commercial detection approaches can be generally divided into two categories: feature-based and recognition-based approaches. While the feature-based approaches use some inherent characteristics of TV commercials to distinguish commercials and other types of videos, the recognition-based methods attempt to identify commercials by searching a database that contains known commercials. The challenges faced by both approaches are the same: how to accurately detect commercial breaks, each of which consists of a series of commercials; and how to automatically perform fast commercial recognition in real time.

Manuscript received July 26, 2009.

Venkatesh. N is with Tata Consultancy Services Innovation Lab, Plot #96, EPIP Industrial area, White field main road, Bangalore-560066, India (phone: 080-67258614; e-mail: venk.n@tcs.com).

Rajeev.B is with Tata Consultancy Innovation Services Lab, Plot #96, EPIP Industrial area, White field main road, Bangalore-560066, India (e-mail: rajeev.b@tcs.com).

M. Girish Chandra is with Tata Consultancy Services Innovation Lab, Plot #96, EPIP Industrial area, White field main road, Bangalore-560066, India (e-mail: m.gchandra@tcs.com).

Commercial break detection [1] [2] have based their strategies in studying the relation between audio silences and black frames as an indicator of commercials boundaries. The analysis is performed in either compressed [1] or uncompressed [2] domains. In [3] also specific country regulations about commercials broadcast is used as a further clue. Another interesting approach is presented in [4], where the overlaid text trajectories are used to detect commercial breaks. The idea here is that overlaid text (if any) usually remains more stable during the program time than in the case of commercials. In [5] Hidden Markov Model (HMM) based commercial detection is presented with two different observations taken for each video shot: logo presence and shot duration.

Generally black frame [6] is used as the crucial information for detecting the commercial in video processing. But, in Indian TV stations black frames are absent and our methodology works without using black frames as video clues. Our simple and accurate approach for commercial detection uses audiovisual features. Initially, we find the commercial boundaries using a combination of different audio features and the confidence level of the boundary detected will be further enhanced or validated using logo (program name) video frames detection only at commercial boundaries detected by audio based feature and later we use video features combination of Prewitt edges[7] and Harris corner points[8] for image matching of logo frame appearing at the start of the commercial and end of the commercial logo frame for measuring the similarity in terms of simple pixel intensity based matching. The Block diagram of the system proposed in this paper is shown in figure 1.

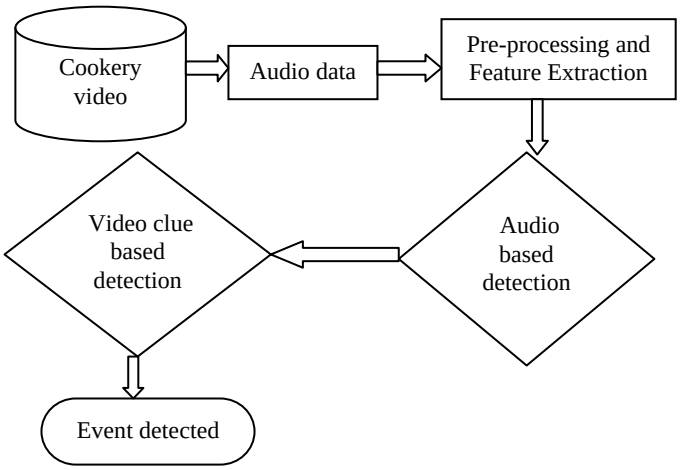


Fig.1 Block diagram of commercial detection

The rest of this paper is organized as follows. In Section 2 Preprocessing and Feature extraction is presented. Section 3 describes the proposed scheme used to detect commercial boundaries. Results and discussion are presented in Section 4. Finally conclusion and future work are presented in Section 6.

I. Pre-processing and Feature Extraction

A. Pre-processing

The audio signal of cookery video is extracted and is segmented into audio frames with 50% overlap and 23 msec duration.

B. Audio Feature Extraction

1) Spectral Rolloff (R): Spectrum rolloff [1] point is the boundary frequency f_r below which a 85% of the spectral energy for a given audio frame is concentrated. This feature is used to differentiate voiced from unvoiced speech. It is defied as the frequency below which an experimentally chosen percentage of the accumulated magnitudes of the spectrum is concentrated. Unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum.

$$\sum_{k=0}^{f_r} \left| R(k) \right| = 0.85 \times \sum_{k=0}^{K-1} \left| R(k) \right| \quad \text{..... (1)}$$

If f_r is the largest value of k for which equation 1 is satisfied then this frequency f_r is the rolloff.

2) Zero-Crossing Rate (ZCR): It is defined as the number of times the audio waveform changes its sign in the duration of the frame or in other words it is the number of time-domain zero crossings within a speech frame. It is called a measure of the dominant frequency in a signal. The zero crossing rate of the frame ending at time instant 'm' is defined by,

$$\text{ZCR} = \frac{1}{2} \sum_{n=m-N+1}^{m} \left| \text{sgn}[x(n)] - \text{sgn}[x(n-1)] \right| \text{..... (2)}$$

$$\text{where } \text{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$$

3) Short-Time Energy (E): The short-time energy of a frame is defined as the sum of squares of the signal samples normalized by the frame length.

$$\text{E} = \left(\frac{1}{\text{N}} \sum_{n-0}^{N-1} x\,(n) \right) \quad \text{..... (3)}$$

C. Video Feature Extraction

1) Prewitt Edge detection: This process detects outlines of an object and boundaries between objects and the background in the image [7].

The basic edge-detection operator is a matrix area gradient operation that determines the level of variance between different pixels. The edge-detection operator is calculated by forming a matrix centered on a pixel chosen as the center of the matrix area. If the value of this matrix area is above a given threshold, then the middle pixel is classified as an edge.

The Prewitt operator measures two components. The vertical edge component is calculated with kernel K_X and the horizontal edge component is calculated with kernel K_Y. The intensity of the gradient in the current pixel is given by $|K_X| + |K_Y|$.

$K_X =$

-1	0	1
-1	0	1
-1	0	1

$K_Y =$

1	1	1
0	0	0
-1	-1	-1

Fig. 2 Prewitt vertical and horizontal operators

2) Harris corner points: The Harris corner detector is a popular interest point detector due to its strong invariance to [8] rotation, scale, illumination variation and image noise. The Harris corner detector is based on the local auto-correlation function of a signal where the local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions.

Basic idea here is to find the points where edges meet that in other words strong brightness changes in orthogonal directions.

$$E(u,v) = \sum_{x,y} w(x,y)\,[I(x+u, y+v) - I(x,y)]^2 \text{ ... (4)}$$

.

where *w(x,y)* is a window function at point *(x,y)*, *I(x+u,y+v)* is shifted intensity and *I(x,y)* is intensity at point *(x,y)*.

For small shifts [*u,v*] we have a *bilinear* approximation:

$$E(u,v) \cong [u,v]\, M \begin{bmatrix} u \\ v \end{bmatrix} \text{……...…...……. (5)}$$

where *M* is a 2×2 matrix computed from image derivatives:

$$M = \sum w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \text{……...…...…. (6)}$$

Measure of corner response is given by:

$$\det M = \lambda_1 \lambda_2$$

$$\text{trace M} = \lambda_1 + \lambda_2 \quad \text{……..…..... (7)}$$

$$R = \det M - k\,(trace\,M)^2 \text{ ……...…...…… (8)}$$

where λ1, λ2 are eigenvalues of *M*.

Choosing the points with large corner response function R (*R* > threshold) and considering the points of local maxima of R gives us the Harris corner points.

II. Commercial Detection Scheme

A. Audio Feature Based Detection

Having extracted audio features, we use heuristic rules found by experimentation for commercial detection.

1) *Zero Crossing Rate & Spectral Roll off:* With Zero crossing rates (*Z*) we set a threshold of 75% of maximum

number of zero crossings value across all the frames denoted by T_Z. We consider the frames which has the zero crossing rates value greater than experimentally found threshold T_Z as boundaries in commercial which is given in equation 9. Let N be the number of frames.

$$Z'(n) = \begin{cases} 1 & \text{if} \quad Z(n) > T_Z \\ 0 & \text{else} \end{cases} \quad \text{where, } n = 1 \ldots N$$

..... (9)

Similarly we identify the frames corresponding to boundary detection of commercials using Spectral roll off (R) by having an experimentally found threshold denoted by T_{SF} and it given in equation 10.

$$R'(n) = \begin{cases} 1 & \text{if} \quad R(n) > T_{SF} \\ 0 & \text{else} \end{cases} \quad \text{where, } n = 1 \ldots N$$

..... (10)

Having found frames corresponding to boundary detected by ZCR (Z') and SR(R') we perform a logical AND operation between Z' and R' given by equation 11 which results in reducing the false detection.

$$ZR(n) = Z'(n) \,\&\, R'(n) \quad \text{where, } n = 1 \ldots N$$

..... (11)

The above mentioned feature combination technique works only for cookery video containing the background music or else it fails miserably and hence we initially check for number of zero crossings detected frames N_Z and if it exceeds the threshold T_F, then we go for the other set of features as explained in section III.A.2. In this case threshold T_F is given by equation 12 where, N is the total number of frames.

$$T_F = \left(\frac{N}{2} \right) \quad . \quad \text{..... (12)}$$

2) Spectral Energy

We have observed in many cookery videos without background noise that during commercial interval, the spectral energy (E) is high we make use of this information for detecting commercial boundaries.

Frames exceeding the experimentally found threshold T_E are considered as the key frames for commercial boundary detection as given in equation 13.

$$E'(n) = \begin{cases} 1 & \text{if} \quad E(n) > T_E \\ 0 & \text{else} \end{cases} \quad \text{where, } n = 1 \ldots N$$

..... (13)

B. Video Feature Based Detection

Once we have detected the commercial boundaries using audio features we validate those boundaries using logo (program name) video frames matching at start and end of commercials. We identify video frame with logo at the boundaries detected by audio based features which reduces computational time for video processing.

Image matching algorithm using combination of edge and corner points as features are as given below:
Initially we convert colour image to gray scale image and extract edges and corner points of an image as a feature for image matching. We match the combined features which are in binary forms by simple pixel intensity matching as explained below:

Step 1: $M_P = 0$ where M_P is the number of pixels matched.

Step 2: $\sum_x \sum_y$ *if* $(R(x,y) = 1 \;\&\; T(x,y) = 1)$ then

$M_P = M_P + 1$

where R is the reference template image at point (x,y) and T(x,y) is test image at point (x,y).

Step 3: $P_M = \frac{M_P}{W_R} * 100$

where W_R is the number of white pixels in reference template image R.

Step 4: If P_M is greater than 80% then we can say that two images are almost similar.

III. Results and Discussion

In order to examine the proposed methodology and to carry out the relevant experimentation we have recorded four videos from three different Indian TV stations.

Videos are captured at 25 frames per second with 288×352 resolution and the audio was extracted from these videos and stored in uncompressed wav format at 44100 kHz with 16 bits per sample for further processing. Commercial details for the recorded data are provided in table 1. We have performed our experiments using MATLAB programming in windows operating system.

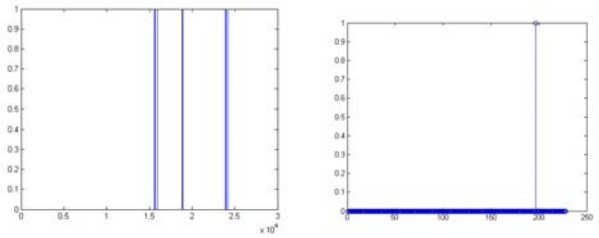
TABLE 1
Details of Data

	Video 1	Video 2	Video3	Video4
Background Music	Present	Present	Absent	Present
Video Length	3min 49sec	12 min 58 sec	18 min 49 sec	23 min 37 sec
Commercial Length in seconds	(i)197-end of the video	(ii)180-282	(i)367-456	(i)289-344 (ii)929-993 (iii) 1103 till end of the video

The first two videos (video 1 and video 2) were found to have background music and hence they are processed by non-energy features like ZCR and Spectral Roll off.

Non-Energy based features fail miserably for video without background music and hence we need to involve energy features. In the present case we have used short time energy feature.

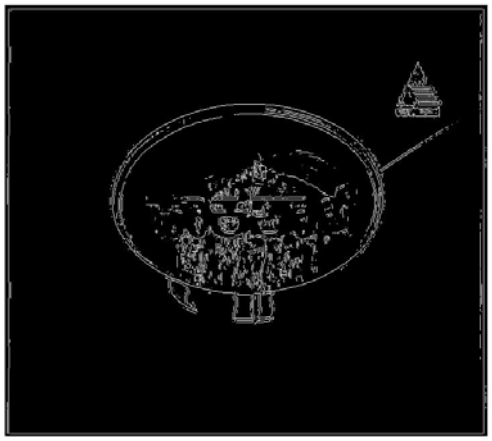
The audio based commercial boundary detection are shown in figure 2(a), 4(a) and further processing by using video features are as shown in fig 2(b), 4(b) for video 1 and 2 respectively. Logo (program name) video at the commercial boundaries is shown in figure 3(a). Correspondingly Prewitt edge detection and Harris point detection for logo video frame are shown in figure 3(b) and 3(c) respectively.



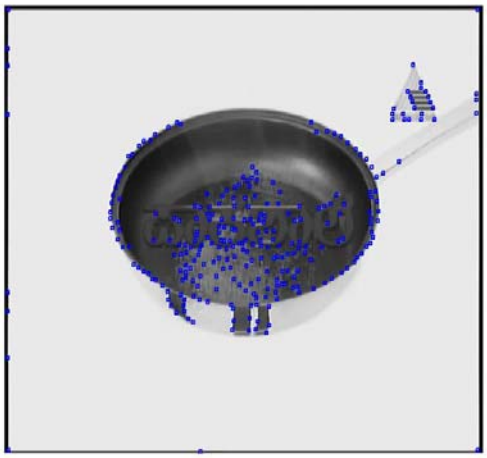
2(a) 2(b)

Fig.2 a) Audio feature based boundary detected for video1, b) Video feature based boundary detected for video

3(a)

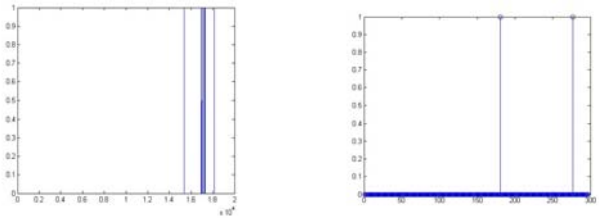


3(b)



3(c)

Fig.3 a) Logo (program name) video frame. b) Prewitt edge detection for logo video frame. c) Harris corner points for logo video frame.



4(a) 4(b)

Fig.4 a) Audio feature based boundary detected for video2, b) Video feature based boundary detected for video 2.

Table 2 gives a detailed split up of all the commercials detected by audio.

TABLE 2
RESULTS USING ONLY AUDIO FEATURES

	Total No of events detected	**No of false events detected**	**No of missed event detected**	**No of true events detected**
Video 1	1	0	0	1
Video 2	2	0	0	2
Video 3	2	1	0	1
Video 4	6	3	1	3

We can observe from figure 1(a) and figure 3(a) that using audio features lot of frames are highlighted during the commercial. From observation and experiment we consider only the first and last frame as a boundary for the commercial if and only if the time interval between first highlighted frame and last highlighted frame exceeds at least 1 minute because commercials usually lasts for more than one minute.

We measure the system performance using Recall and Precision given by equation 14 and 15 and tabulated the results accordingly in table 3.

$$\text{Recall} = \frac{\text{The number of true events detected}}{\text{The actual number of events}} \quad \text{..... (14)}$$

$$\text{Precision} = \frac{\text{The number of true events detected}}{\text{The total number of events detected}} \quad \text{..... (15)}$$

From table 2 we can see that using only audio based features there are some false events detected in video 3 and video 4 because it corresponds to prize distribution for the quiz winner of last episode. Here both audio and visual changes are very similar to commercial. To overcome these problems we have made use of video clues.

TABLE 3
RESULTS USING AUDIOVISUAL FEATURES

Game	Recall	Precision
Video 1	1	1
Video 2	1	1
Video 3	1	0.5
Video 4	0.75	0.5

TABLE 4
RESULTS USING BOTH AUDIO AND VIDEO FEATURES

	Total No of events detected	**No of false events detected**	**No of missed event detected**	**No of true events detected**
Video 1	1	0	0	1
Video 2	2	0	0	2
Video 3	1	0	0	1
Video 4	4	0	1	3

From table 4 we observe that using our simple video clues of matching logo video frames at commercial boundaries for refining the boundaries detected by audio we are able to get rid of the false events detected by audio.
One more observation is that the number of missed events remain same in video 4 because we try to locate the logo video frames only at the boundary instances detected by audio and if the events are missed by the audio then it cannot be recovered by video processing but instead it is used only to validate and increase the confidence level of commercial boundaries detected by audio.

TABLE 5
RESULTS USING BOTH AUDIO AND VISUAL FEATURES

Game	Recall	Precision
Video 1	1	1
Video 2	1	1
Video 3	1	1
Video 4	1	0.75

IV. CONCLUSION AND FUTURE WORK

In this paper we have presented a commercial detection scheme based on both audio and visual features. In the case of audio detection we have two sets of features based on the type of video data. Energy based features (spectral energy entropy) for videos without background music and features like (ZCR and Spectral Roll Off) for distinguishing between speech and non voiced speech segments for videos with background music are considered, hence covering all kind of possible variations in video data. Confidence level of commercial boundary has been improved and we were able to get rid of the false detections by using logo (program name) video frame matching at commercial boundaries.

The above presented system can be very much used for real-time embedded system as it is time efficient, simple and amenable for easy implementation.

Our future work will be focused on using more robust audio features for detecting commercial boundaries without missing any events and further we are trying to experiment on some more cookery videos and also on different TV shows like news, serials and reality shows.

REFERENCES

[1] D. A. Sadlier et al., "Automatic TV advertisement detection from mpeg bitstream," *Journal of the Patt. Rec. Society*, vol. 35, no. 12, pp. 2–15, Dec. 2002.
[2] Y. Li and C.-C. Jay Kuo, "Detecting commercial breaks in real TV program based on audiovisual information," in *SPIE Proc. on IMMS, vol 421*, Nov. 2000.
[3] R. Lienhart, C. Kuhmnch, and W. Effelsberg, "On the detection and recognition of television commercials," in *Proc. IEEE Conf. on MCS*, Otawa, Canada, 1996.
[4] N. Dimitrova, "Multimedia content analysis: the next wave," in *Proc. of the 2nd CIVR*, Illinois, USA, Aug.2003.
[5] Alberto Albiol, Mar´ıa Jos´e Ch. Full`a, Antonio Albiol, Luis Torres, "Commercials Detection using HMMs",2004.
[6] Alberto Albiol Mar , María José , Ch. Fullà , Antonio Albiol, "Detection of TV Commercials" *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*,2004. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
[7] Hong Shan Neoh, Asher Hazanchuk, "Adaptive Edge Detection for Real-Time Video Processing using FPGAs", *International Signal Processing Conference (ISPC), Santa Clara, California*, September 27-30, 2004.
[8] C. Harris, M. Stephens. "A combined corner and edge detector" *Proceedings of the 4th Alvey Vision Conference: pp 147—151*,1988.