ROBUST COMMERCIAL DETECTION SYSTEM

Liang Zhang, Zhenfeng Zhu, Yao Zhao
Institute of Information and Science, Beijing Jiaotong University, Beijing, 100044, China
zhangliangwin@hotmail.com

ABSTRACT

Automatic detection of commercials embedded in digital video materials is a challenging task. In this paper, a Three-phase system of commercial detection is constructed. Firstly, video shots and audio cuts are detected respectively and scene boundaries are refined by fusing the audio and video information simultaneously. Then, a SVM classifier is applied to classifying these shots as program shots or commercial ones. Finally, in order to make the commercial segments more accurate, the identity of time-spatial content of commercial shots is fully exploited to eliminate those falsely classified commercial ones, based on which the individual commercial shots can be integrated into the commercial sequences ultimately. The final test results on 30 clips of TV videos verify the practicality of the constructed system.

1. INTRODUCTION

Today, as one of main information conveying channels, TV commercials are playing an increasingly important role in our daily life. The automatic detection of commercials embedded in digital video materials has been a key focus for diversified purposes. From the view of advertisers, it will be great helpful for them if a fully automatic commercial detection technique can be sought, based on which when the specific commercials are broadcast can be identified and tracked in time to validate the contracts with broadcaster. Meanwhile, TV viewers will generally feel sick of being exposed to large number of commercials when a copy of program is created for viewing at a later time. Hence, how to filter out these unfavorable commercials in advance becomes necessary, which directly leads to the requirements for fully automatic commercial detection technique.

Currently, various strategies for commercial detection have been proposed. In the earlier years, people usually focused on logo-based detection [1] and black/silent frames-based detection [2]. However, TV stations now do not often hide the logos during commercials and black/silent frames are even inserted randomly for some editing purposes. Similar to the technology of video retrieval, the matching scheme can be taken against the TV stream being broadcasted with the premise that a database of previously known commercials is available. Once some commercials being broadcasted haven't appeared in the pre-constructed database, such method won't be workable.

In order to avoid aforementioned limitations, shot-based video processing and classification techniques have been probed and shown great potential [4][5]. But one of disadvantages for them is that less attention on the continuity of commercial contents has been taken into account to eliminate falsely classified commercial shots.

In this paper, a more robust solution is provided with a Three-phase detection system. Firstly, two visual region-based schemes are proposed for cut and dissolved shot detection respectively, and scene boundaries are refined by fusing the audio and video information simultaneously. In addition, some common statistical features are extracted for commercial characterization. Subsequently, the binary classifier SVM is taken for classifying the candidate shots as commercial ones or not. During the final phase, in order to make the commercial segments more accurate, the identity of time-spatial content of commercial shots is fully exploited to endeavor for eliminating those falsely classified commercial ones, based on which the commercial shots can be integrated into the commercial sequences ultimately.

2. SCENE BOUNDARIES DETERMINED

In order to determine the scene boundaries accurately, both visual and audio cuts are integrated to refine scene boundaries based on the co-occurrence of audio cuts with video shots.

2.1. Video Shots Detection

Generally, cut and dissolved shots are two common means of scene transition with high frequency. In the following part, two schemes based on salience of visual region are proposed to detect these two kinds of shot changes.

2.1.1. Cut Shots Detection

A hard cut is an effective characteristic used for video analysis. Difference between two consecutive frames in color or gray histogram space has been proved to work efficiently [6]. Nevertheless, the simply straight application of such method won't be expected to bear fruit for the task at hand with less consideration of the exhibition skills embroiled in commercial producing. That means more important information is generally covered in the middle region of commercial scenes. Thus, if information in middle regions between two consecutive frames has changed sharply with background unchanged, it should take high probability of being considered to be cut shot. Based on such observation, a region-based scheme is proposed.

Referring to the illustration in Fig.1, multi regions can be first partitioned for each frame. Let $H(j, R_i, m)$ denotes the B bins gray histogram of the i th region in the m th frame, j is the corresponding bin index. Thus the identity FD_m of scene transition between the $m-1$ th and m th consecutive frames can be given as

$$FD_m = \sum_{i=1}^{n} \mathrm{w}_i \sum_{j=1}^{B} \frac{\left|H(j,R_i,m)-H(j,R_i,m-1)\right|^2}{Max\{H(j,R_i,m),H(j,R_i,m-1)\}} \qquad (1)$$

1-4244-1017-7/07/$25.00 ©2007 IEEE

where w_i s are some normalized weights to reflect the importance of the corresponding regions and are given as $w_n < w_{n-1} < \cdots < w_i < \cdots < w_1$. For the sake of consideration mentioned above, n denotes the number of partitioned regions and is set to be 3 in our case. In particularly, we set $w_1 = 0.5,\quad w_2 = 0.3, w_3 = 0.2$.

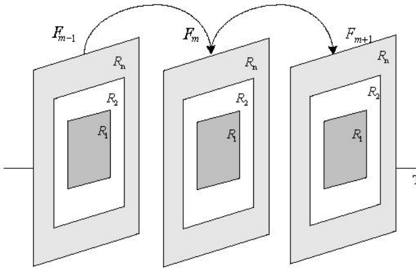

Fig. 1. Sketch map of cut shots detection based on salience of visual regions

With large number of statistic analyses on FD_m s, the identity FD_m can be felicitously characterized by a Rayleigh distribution. The confidence interval of the estimated Rayleigh distribution of FD_m is utilized to obtain the threshold t_{cut} adaptively. Thus a shot boundary can be declared for the i th frame once its FD_m exceeds t_{cut}.

2.1.2. Dissolved Shots Detection

Besides cut shots, another popularly adopted exhibition skill in commercial producing is dissolved shot. During dissolved shot, the change of average-gray keep consistent across some consecutive frames and this monotonous change usually go on for a period of time. So the cumulation of this change across multi frames can be applied to detecting dissolves. As shown in Fig.2, we use regions $R_{m,i}$ (the i th region of the m th frame) to compute gray histogram difference between two consecutive frames.

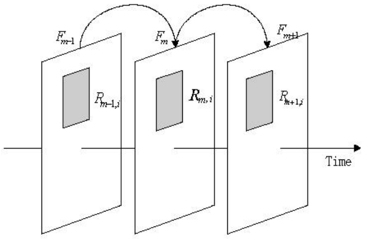

Fig. 2. Sketch map of dissolve shots detection based on local coherent temporal change

Let V_i^m denotes the average gray value of the i th region in the m th frame. For simplifying illustration, we denote S_i^m as an indicator of the change of the gray value of the i th region in the m th frame, T_i^m as the number of consecutive frames with monotonous gray changes of the i th region before the m th frames and N_h^m as the statistical value of all regions in the m th frame, i.e. the number of T_i^m in the range of $[n_1, n_2]$. The flowchart for the procedure of dissolved shot detection is given as follows:

1) Compute the average gray value of sample region V_i^m.
2) Initialize. $T_i^1 = 0$ and S_i^1 =+1.
3) Update S_i^m and T_i^m for each region of the n th frame as follows.

$$S_i^m = \begin{cases} 1 & V_i^m - V_i^{m-1} \geq 0 \\ -1 & V_i^m - V_i^{m-1} < 0 \end{cases} \tag{2}$$

$$T_i^m = T_i^{m-1} + \frac{1 + S_i^{m-1} \cdot S_i^m}{2} \tag{3}$$

Thus the value of N_h^m can be obtained,

$$N_h^m = \#\{T_i^m \mid T_i^m \in [n_1, n_2]\} \tag{4}$$

where # denotes the number of T_i^m.

Here, the effectiveness of the statistical values N_h^m on a test video is shown in Fig.3, from which we can find there are three dissolve shots, appearing during frames 2950--2980, 3320--3350 and 3390—3410 respectively, which are accordant to the real instances. Compared with those during the common program, the changes of N_h^m during dissolves can be distinguished distinctly.

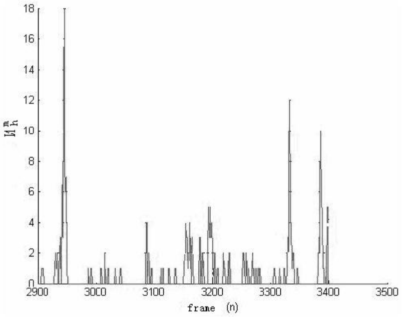

Fig. 3. Effectiveness of the statistical values N_h^m for characterizing the dissolve cuts

2.2. Audio Cut Detection

At present, there are numerous measures for audio analysis, such as pitch, short-time energy, averages zero crossing and so on. Segmenting an audio data means to detect the time indexes corresponding to changes of audio. Note that audio segments belonging to the same scene class may be segmented to different shots.

Here, we explore distances between feature vectors of adjacent audio intervals. Considering of short time property of audio, we compute audio distances by sliding 2s intervals with 100ms overlapping, as shown in Fig.4. Note that a 2s interval is composed of 100 audio clips limited into 20ms long.

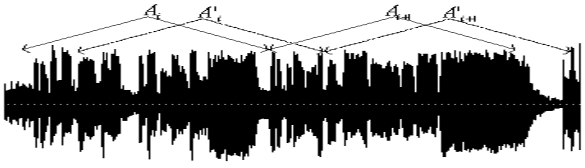

Fig. 4. Audio intervals with 100ms overlapping

For a clip to be declared as an audio change or not, it should be similar to all the neighboring future clips and different from all the neighboring previous clips. Audio is first converted into a sequence of feature vectors V_i based on each clip, so a feature vector of each audio interval can be depicted as $F_A(i) = (V_1, V_2, \cdots V_{100})$. After that, we can detect audio changes as follows,

$$Diff_A = \frac{\left\| \frac{1}{N}\sum_{i=-N}^{-1} F_A(i) - \frac{1}{N}\sum_{j=0}^{N-1} F_A(j) \right\|^2}{\sqrt{c + \mathrm{var}(F_A(-N), \cdots, F_A(-1)) \times \mathrm{var}(F_A(0), \cdots, F_A(N-1))}} \quad (5)$$

where $\| \cdot \|$ is the L-2 norm, $\mathbf{var}(\cdots)$ is the average of the squared Euclidean distances between each feature vector and the mean feature vector during the corresponding interval, and c is just used to prevent from division by zero. From Eq. (5), if $Diff_A$ is smaller than a threshold defined previously, two audio intervals can be declared to belong to the same audio cut; otherwise, when it is significantly larger than this threshold, the audio cut change is detected.

2.3. Scene Boundaries Refinement

In order to refine the scene boundaries, the concurrence of audio cuts and visual shots is exploited.

For briefness, $T_{VSC}(i)$ denotes the time of the i th video shot change and $T_{ASC}(j)$ denotes the time of the j th audio cut change. The refining process is as follows:

```
Step 1: If( $|T_{VSC}(i) - T_{ASC}(j)| \prec t$ ) then
            Candidate scene boundary is detected
        Else
            goto step2
Step 2: Detect $T_{ASC}(j')$and $T_{ASC}(j'+1)$ , between them
        only exist one shot $VSC(i)$
        compute: $Diff_T(j') = T_{VSC}(i) - T_{ASC}(j')$
                 $Diff_T(j'+1) = T_{ASC}(j'+1) - T_{VSC}(i)$
            goto step3
Step 3: If( $Diff_T(j') < Diff_T(j'+1)$ ) then
            Candidate scene boundary is adjusted as $T_{ASC}(j')$
            and merge video[ $T_{ASC}(j')$ --$T_{ASC}(j')$ ] into $VS_i$
        Else
            Candidate scene boundary is adjusted as $T_{ACS}(j'+1)$
            and merge video[ $T_{VSC}(i)$ -- $T_{ACS}(j'+1)$] into $VS_{i-1}$
```

3. SVM-BASED SHOT CLASSIFICATION

As an eminent binary classifier, the nature of SVM is to find a discriminate hyper-plane that optimally separates two classes of objects by using structural risk minimization [7]. The final discriminate hyper-plane can be given as [8].

In section 2, to represent the segmented shots efficiently, some robust features, such as Shot Frequency, Average and Variance of Frame Difference, the percent of domain color of frames, shot-time average energy function, average zero crossing and energy distribution have been extracted for each shot, based on which the SVM classification model is trained from a pre-labeled training data-set. Hence, it can be applied to classifying video shots into commercial shots or program ones.

4. POST-PROCESSING ON COMMERCIAL SHOTS

As it is inevitable to have some falsely classified shots through the SVM classification, some post-processing mechanisms are necessary to refine the extent of commercials and merge the different commercial shots into commercial sequences. As it is known, shots of commercials are continuous and appear in group. With this characteristic, the following strategy is conducted to eliminate these misclassified shots and get the commercial sequences.

Let $shot_i$ ($1 \le i \le n$) represents the i th shot and C_i represents a score of $shot_i$ used to decide whether this shot is commercial shot or not. n is the number of shots.

1) Initialize C_i as.

$$C_i = \begin{cases} +1 & if\ the\ ith\ shot\ is\ commercial one \\ -1 & else \end{cases} \quad (6)$$

2) Use a sliding window including 5 shots to update C_i. Let $W\{w_j | w_j = 1,\ if\ \ -2 \le j \le +2\}$ work as a sliding window. Then update C_j ($i-2 \le j \le i+2$) as follows.

$$C_j = \begin{cases} C_j + 1 & if\ \sum_{k=-2}^{2} C_{i+k} w_k \ge 3 \\ C_j - 1 & else \end{cases} \quad (7)$$

3) Then if $C_i > 0$, $shot_i$ is considered to be a commercial shot. Otherwise, it is a program shot.

4) If there are more than two shots changing form commercial ones to program ones or vice versa, go to the first step. If there is less, the refinement is ended.

5. EXPERIMENTAL RESULTS

We have tested our system on about 30 clips of teleplays taken from different TV stations in China. The identity FD between two frames is shown in Fig.5 with x-axis as the number of frames. Since dense FDs have appeared during commercial broadcasting compared with formal program, it is feasible to take it as a reliable measure for discrimination of commercial from common program.

The results of cut detection on video 1 including 80 shots are presented in Table1 with adaptively selected thresholds. The method in [6] (P-algorithm for short) detects 70 shots with 65 shots right. Our algorithm (O-algorithm for short) detects 78 shots with 76 shots positive.

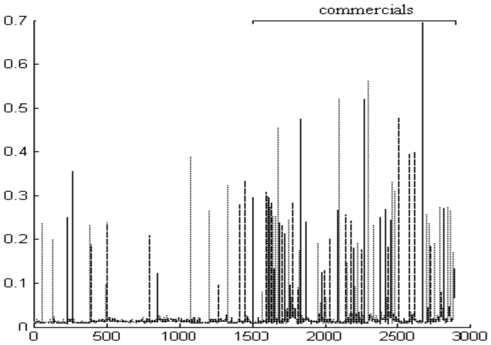

Fig. 5. dentity of FD between two frames

Table 1. Results of hard cut detection

	Threshold	Cut	Recall	Precision
P-algorithm	0.10	70	81.25%	92.85%
O-algorithm	0.18	78	95.00%	97.44%

In Table2, the first clip of teleplay contains 82 shots and the number of dissolve shots is 2. We detect 78 cut shots with 75 right. While the second one contains 49 shots and the number of dissolve shots is 5, 43 cut shots with 40 right are detected. Table 3 shows the classification results: 93.87% without refinement and 97.58%with it. The final refining results are presented in Table 4, from which we can find that the refined boundaries of these shots via audio-visual information are very consistent with the real ones. Thereby, it is helpful for detecting the commercial sequences.

Table 2. Results of shot detection of cut and dissolve

	Dissolve	Cut	Recall	Precision
clip 1	2	78	93.90%	96.25%
clip 2	5	43	91..84%	93.75

Table 3. Classification results of commercial shots

Refinement	All Shots	Accuracy
Without	305	93.87%
With	305	97.58%

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present a more robust solution with a Three-phase detection system. Firstly, two visual region-based schemes are proposed for cut and disso0lved shot detection respectively, and scene boundaries are refined by fusing the audio and video information simultaneously. Subsequently, the binary classifier SVM is taken for classifying the candidate shots as commercial shots or not. During the final phase, in order to make the commercial segmentation more accurate, the identity of time-spatial content of commercial shots is fully exploited to endeavor for eliminating those falsely classified commercial ones, based on which the commercial shots can be integrated into the commercial sequences ultimately. In our future work, some more robust features, such as text information, will be exploited. In addition, the content- based segmentation techiniquecan also be considered for commercial boundary refining.

Table 4 The refining results based on audio-visual information

Video shots \ Commercial shots		1	2	3	4	5	6	7	8	9
Detection of Commercial shots	Real boundary	1164 — 1183	1184 — 1204	1205 — 1126	1127 — 1292	1293 — 1337	1338 — 1388	1389— 1463	1537— 1588	1589—- 1612
	Visual based	1160 — 1187	1188 — 1210	1211 — 1126	1127 — 1289	1290 — 1340	1341 — 1395	1397 — 1470	1530 — 1585	1589 — 1612
	Audio-visual based	1163 — 1182	1182 — 1207	1208 — 1126	1126— 1290	1291 — 1335	1336 — 1390	1391 — 1465	1533 — 1589	1590 — 1612

7. ACKNOWLEDGEMENTS

This work was funded in part by National Natural Science Foundation of China (Nos. 60602030, 90604032), 973 Program (No. 2006CB303104), Program for New Century Excellent Talents in University, Specialized Research Foundation of BJTU (No. 2005SZ005), Research Foundation of Beijing Jiaotong University (No. 2005SM013). Open Foundation of National Laboratory of pattern Recognition, CAS.

REFERENCES

[1] A.Hauptmann, M.Witbrock, “Story Segmentation and Detection of Commercials in Broadcast News Video”. Advances in Digital Libraries Conf., Santa Barbara, CA, April 22-24, 1998

[2] David A. Sadlier, et al, “Automatic TV Advertisement Detection from MPEG Bitstream”. Intl Conf on Enterprise Information System, Setubal, Portugal, 7-10 July 2001

[3] Rainer Lienhart, et al, “On the Detection and Recognition of Television Commercials”. Multimedia Computing and Systems '97. Proceedings, IEEE International Conference on 3-6 June 1997

[4] Xian-Sheng Hua, Lie Lu; Hong-Jiang Zhang, “Robust Learning-Based TV Commercial Detection”. IEEE International Conference on Multimedia and Expo, July, 2005

[5] Pinar Duygulu, et al, “Comparison and Combination of Two Novel Commercial Detection Methods”. The 2004 International Conference on Multimedia and Expo (ICME’04), Taipei, Taiwan, June 27-30, 2004

[6] A. Miene, A. Dammeyer, Th.Hermes, et al, “Advanced and Adaptive Shot Boundary Detection”. Proc. Of ECDL WS Generalized Documents, pp.39-43

[7] V.Vapnik, “The Nature of Statistical Learning Theory”. Springer, New York, 1995

[8] Changsheng Xu, Maddage. N.C et al, “Musical genre classification using support vector machines”. Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on Volume 5, 2003