COMMERCIAL DETECTION IN HETEROGENEOUS VIDEO STREAMS USING FUSED MULTI-MODAL AND TEMPORAL FEATURES

*[1,2]*Masami Mizutani*
mizutani.masami@jp.fujitsu.com

*[2]*Shahram Ebadollahi*
shahram@ee.columbia.edu

*[2]*Shih-Fu Chang*
sfchang@ee.columbia.edu

*[1]Fujitsu Laboratories Limited, Kawasaki 211-8588, Japan
*[2]Electrical Engineering Department, Columbia University, New York, NY 10027, USA

ABSTRACT

We provide an integrated approach for detecting commercial segments in video streams. This approach systematically fuses the "local" multi-modal characteristics of commercials in the context of their "global" temporal behavior throughout the video stream.

Discriminative classifiers are employed to distinguish between commercial and program segments based on their local multi-modal features. The decisions made by different discriminators are fused using a Support Vector Machine. The fusion results are then used as the probabilistic outcomes of a generative model describing the transitions between the commercial and program segments, with explicit models for the inter-arrival times of the commercial segments throughout the video.

This approach aims to enhance upon the simple, yet effective blank frames, which usually indicate the start of commercials. It also provides acceptable performance when such indicators do not exist in the program stream.

The results of comprehensive experiments on a heterogeneous data set of 36 hours of video taken from 6 different sources are reported. Our method provides almost 92% correct detection of the commercial segments and 8% enhancement over just using the blank indicators. For the case when blank indicators do not exist, our approach results in almost 85% correct detection.

1. INTRODUCTION

There are two main reasons for the interest in automatic detection of commercial segments in video streams: 1) adding commercial detection/skipping capability to video set-top boxes; 2) focusing the content analysis algorithms to the program segments for more efficiency.

Various algorithms have been proposed for detecting commercials in video streams. The work reported in [2] uses blank and silence detectors in a heuristic manner and shows very good performance on detecting commercials. However, the drawback of this method is that blank frames which flag the start of the commercial segments are not consistent in video streams (Figure 3). In [4], an algorithm is proposed that does not rely on blank frames. This method fuses the decisions obtained from classifiers that classify programs and commercials based on their audio and color patterns with the ones obtained from repetitious video segments which could potentially be commercials. The drawback of this algorithm is that commercials do not necessarily repeat themselves when one deals with heterogeneous data sets obtained from different video streams at different times. In addition, the method was tested only on videos comprised of news and commercials, and therefore does not necessarily scale to heterogeneous data sets.

We aim to develop an algorithm for detecting commercial segments in video streams obtained from heterogeneous sources, which not only improves upon simple yet effective blank detection but also provide reliable detection performance when blank indicators are not available.

To achieve our goal, we systematically fuse audio/visual/temporal local features of commercials in the context of their global temporal characteristics. Discriminative classifiers are employed to distinguish between commercial and program segments based on their local multi-modal features. The decisions made by different discriminators are fused using a Support Vector Machine (SVM) classifier. The results of the fusion of the decisions are then used as the probabilistic outcomes of a generative process modeling the global temporal characteristic of the commercials. A section of the video stream is declared to be a commercial, if its location in the program stream resembles the pattern of occurrence of commercial segments, and its audio/visual/temporal local characteristics resemble those of commercial rather than program segments in a probabilistic sense.

We report the results of comprehensive experiments on a heterogeneous data set of 36 hours of video taken from 6 different general interest channels. We show that the most effective results are obtained when one uses the combination of local features and blank frames in the context of the global temporal behavior of commercials. This provides on average 8% improvement over the case that only uses blanks for commercial detection. We also show that the global temporal characteristic significantly improve the results compared to the results using just local classifiers.

Most importantly, the fusion of the local and global characteristics of commercials provides acceptable results

when there are no blank frames to flag the start of commercials. Figure 3 shows that the blank frames are not consistent across the data set and our method which uses other local features exceeds the worst and the best performances of the case of only using the blanks.

2. CHARACTERISTICS OF COMMERCIALS

Commercials do not occur randomly in the program stream. The timing of the insertion of a commercial segment is both dictated by the type of the program in which it is inserted and the content planning strategies of the broadcasters. These constraints together impose a distinct distribution of the inter-arrival times between the two consecutive commercial segments throughout the video stream.

Figures 1.b and 1.c show the difference between the inter-arrival times of the commercial segments during two different types of programming. During movies, commercial segments are more spread apart, whereas in sports, the frequency of their insertion tends to be higher.

According to the guidelines [1], commercials are made to be "appealing" to viewers in order to capture their attention. The appealing nature of commercials can be realized in both the semantics and the syntax.

From the syntactic point of view, the combination of audio, video and temporal characteristics could be exploited to catch viewers' attention. Discovering the exact features that contribute to the appealing nature of commercials needs an extensive study of the psychology of the perception of commercials. Here we attempt to select a set of features that are relevant to the psychological perception and are useful in quantifying the nature of commercials. These features are discussed in the next section.

Therefore, if we look at the entire video stream, commercials reveal a global temporal characteristic, which is the temporal spacing between the occurrences of commercials segments. Also, at the local scale, commercials display audio/visual/temporal characteristics that are due to their nature and are probabilistically distinct from the regular programs. These observations make the foundation of our approach for commercial detection.

3. PROBLEM STATEMENT AND APPROACH

We model the program stream by a two-state first-order Markov chain alternating between commercial (CM) and program (PG) segments according to a certain transition probability and an explicitly modeled duration of stay in each state.

According to this generative formulation of the problem, the program stream is a sequence of observations made at times corresponding to the scene change boundaries, where each observation is generated either due to CM or PG state of the system.

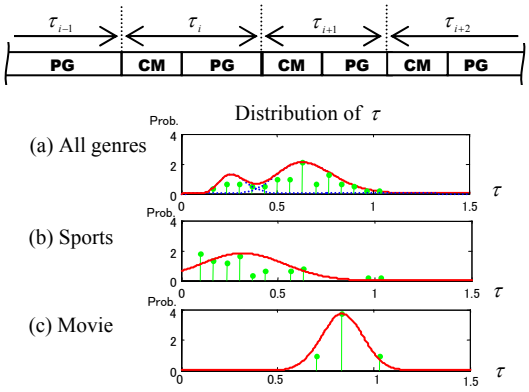

Figure 1. Inter-arrival times of commercial segments for 2 distinct genres and for all genres put together.

The problem of commercial segment detection is therefore transformed into that of inferring the optimal sequence of the hidden states of this duration dependent generative model. Note that we take into account the actual duration of stay in each state rather than the number of time steps spent in each state. We model the duration density functions as a uniform distribution for CM state and as a mixture of Erlang models for PG state.

The observation at each time step is obtained from the multiple discriminative classifiers. Each classifier is trained to distinguish between CM and PG segments using one kind of related raw features obtained from a local window. The posterior probability of each of the two possible hidden states is obtained by fusing the observation vector using a SVM classifier and obtaining the posterior probabilities of the different classes from the resulting margins [8].

This approach, therefore, captures the global characteristic in the generative model and the local characteristics in the discriminative classifiers. The final decision is obtained by inference in this framework and therefore fusing the decisions of both the generative and discriminative components to obtain the location of commercial segments.

4. FEATURE EXTRACTION AND FUSION

For each scene boundary obtained from a scene change detector [7], we extract a number of audio, visual and temporal features for a 15-second window after the boundary location (Figure 2). The 15-second window was chosen to be able to capture the thematic features of the minimum length of commercial segments. We also extract blank frame features from a 120-second window centered at the boundary location in order to include the boundaries of commercial clips before and after the candidate point.

The elements of the feature set are selected based on how well they might be able to represent or capture the appealing nature of the commercial video segments. These features are briefly explained here. We would like to refer readers to [10] for more details on them.

(1) Audio class histogram (ACH): we use an audio class classifier so that it can classify the sound during one second into one of four classes (silence, speech, music and music/speech). We compute the count of one second units having each audio class in the 15-second window. The rational for using this feature is that commercials might have a distinct pattern of audio classes from that of regular programs.

(2) Commercial pallet histogram (CPH): in [5] the mood of the scenes in movies was captured and represented by the "movie pallet histogram". This representation was shown to be effective for classifying movie scenes into different mood classes such as a horror scene or happy scene. Here we use a similar concept and extract a color feature called "commercial pallet histogram" in the 15-second window. The rational behind this representation is that commercials may use certain moods to be more appealing to viewers.

(3) Text location indicator (TL): commercials usually use overlay text to both capture viewers' attention and convey information. Here we detect overlay text using the system described in [6] and map them to a 16x16 binary grid in every 5 frames (a grid element has a value of 1 if text is detected in the grid area). Then the grid values are accumulated over all frames in the 15-second window in order to capture the locations and frequencies of texts occurrence throughout the video.

(4) Scene change rate (SCR): observations show that commercials possess distinct scene cuts. The distributions of the scene change rates for CM and PG segments are modeled to capture this characteristic of commercials. Based on our empirical simulations, we found the scene change rate can be adequately modeled by Poisson distribution.

(5) Blank frame rate (BFR): as mentioned before, blank frames are very strong indicators of commercials when available. We use a blank frame detector that compares the average pixel intensity of a frame with a threshold, and compute the number of the detected blank frames within the entire 120-second window. The choice of the large window size is better for evaluating the number of commercial clips around the candidate point, which are connected by blank frames. The distributions of the number of the blank frames for CM and PG segments are modeled by Poisson distribution based on our empirical simulations.

From the above raw features, we take a two-step process to fuse them into a single posterior probability of the current candidate being in CM or PG state. The state posterior probability is then used with the 2-state Markov model (section 3) to infer the current state. First, for each of ACH, CPH and TL, we train a SVM classifier with RBF (Radial Basis Function) in order to distinguish between the patterns of those features in the observation window taken from commercial and program segments. In order to treat the outputs consistently with other features, we convert them to the posterior probability of commercial using the sigmoid function [8]. Similarly, as for each of SCR and BFR, the likelihood values derived from the classifier based on the probability model are converted to the posterior probability of commercial using Bayes rule with equal priors. The performance of the classifiers is shown in [10].

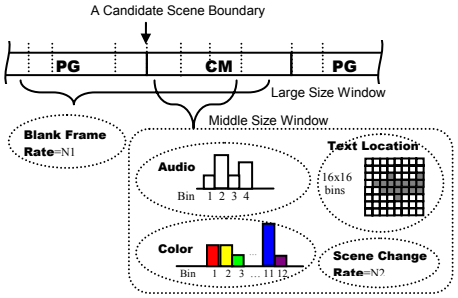

Figure 2. Local features of commercials: in the two observation windows placed around a candidate point

We then train another SVM classifier with RBF kernel to fuse the above posteriors based on individual features into a singe posterior probability of commercial. Such an approach is equivalent to late fusion of multiple features, in contrast with an alternative that directly fuses multiple raw features. We did not adopt the latter because of the great diversity of the features.

5. EXPERIMENTS AND RESULTS

We use a heterogeneous data set of 36 hours of MPEG-1 streams including about a total of 9 hours of commercial segments. The data set consists of 49 individual programs taken from 6 general interest channels (ABC, CBS, FOX, NBC, UPN, WB11) and 6 genres (news, drama, animation, movie, entertainment, sports). Ground truth for the boundaries of the commercial segments was obtained manually. The data set is divided into 3 parts, out of which two parts are used for training and one for testing.

In order to show the efficiency of our approach, we conduct experiments using three kinds of feature sets: a) only blank feature, b) all features except blank, c) all features including blank. The experiments are conducted for three different scenarios: 1) local point decision, 2) inference using the conventional Viterbi but without modeling the duration, 3) inference using the modified Viterbi to reflect both the temporal transition characteristics and explicit durations of the states.

Two types of metrics are used to assess the performance and to compare the results. The first metric is a version of the *F1 metric* as suggested in [3], which we use for counting correctly classified boundaries. This metric has a problem handling short duration false alarm segments occasionally. As a complementary metric, we use *WindowDiff (WD)* [9] that is widely used in the field of text segmentation, which is more appropriate to see how many

Table 1. Results of commercial detection: Using all features in the context of the global characteristic gives the best results (lower-right element). Note for WD, lower values are better.

	(a) Only blank		(b) Other features		(c) All features	
	F1 (%)	WD (%)	F1 (%)	WD (%)	F1 (%)	WD (%)
1) Point Decision	87	23	80	71	90	38
2) Viterbi	87	22	84	31	91	16
3) Duration Viterbi	89	21	85	25	92	15

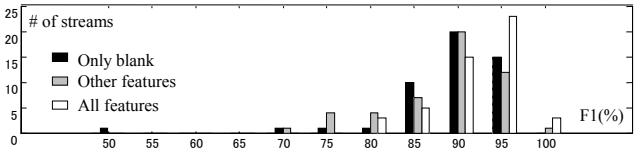

Figure 3. Blank features although powerful for commercial detection, are not consistent. Figure shows that the number of erroneous results (error > 30% or F1 < 70%) is higher for the case where only blank features are used.

discrepancies occur between ground truth and detection result. The lower the score obtained from this latter metric, the better is its performance, while for the former, the reverse is true.

The detection results on all 49 test streams across 3 rounds are shown in Table 1. The first column on the left shows that the local point decisions obtained using only the blank features have high accuracy and filtering the results using the global temporal characteristic of the commercials does not improve the results significantly.

According to the middle column, if we do not use the blank feature, the result of the point decisions made to classify the observation windows into CM or PG class is not good, but when the local decisions are put into the context of the global characteristic of the commercials temporal pattern, the results are drastically improved. It also shows the duration dependent Markov model further improves the conventional Markov model.

The rightmost column shows that when all features including the blanks are employed in the context of the global characteristic of commercials, then one can obtain the best results (WD=15%, F1=92%).

This results show that our approach which assesses the local characteristics of commercial segments in the context of their global inter-arrival pattern throughout the program stream is the most effective in locating commercial segments. Even when blank frames are not available, we can detect the commercial block boundaries with acceptable accuracy (WD=25%, F1=85%).

Figure 3 shows the distribution of detection accuracy over individual video sequences. This graph shows that the blank feature is not consistently present in the program stream to flag the start of the commercials and therefore it is not a reliable feature for commercial detection. Some programs, due to the lack of blank features in the video, suffer very low detection accuracy. However, by fusing all features (including blanks) we can achieve better consistency in the results for commercial detection on a heterogeneous data set. This way we can leverage the blank frames, which are strong indicators of commercials, when they are available, and enhance upon them (8% on WD) by using other features as described in the paper.

6. CONCLUSION AND FUTURE WORK

In this paper, we described a systematic approach for detecting commercial segments using their local and global characteristics in the program stream. We also reported the results of a set of comprehensive experiments on a heterogeneous data set (36 hours from 6 channels) and demonstrated the efficiency of our approach in detecting commercial segments. Our experiments show that fusing the local and global characteristics of commercials using the combined generative and discriminative model provides the best results. Even when the blank indicators are not used, the framework results in the acceptable commercial boundary detection.

One improvement that could be done in this work is to automatically select the most discriminative features for distinguishing between CM and PG segments from a large pool of extracted features. The feature set that we used was hand selected with attempt to match psychological perception, and the reported results are subject to this selection.

7. REFERENCES

[1] E. J. Heighton et al, "Advertising in the broadcast media", p.97-120, Wadsworth Publishing Company, 1976

[2] S. Marlow, et al, "Audio and Video Processing for Automatic TV Advertisement Detection", Proc. of ISSC, 2001

[3] N. Dimitrova, et al, "Evolvable Visual Commercial Detector", in the proceeding of CVPR'03, Vol. II. p.79, 2003

[4] P. Duygulu, et al, "Comparison and Combination of Two Novel Commercial Detection Methods", Proc. of ICME, 2004

[5] C. Wei, et al, "Color-Mood Analysis of Films Based on Syntactic and Psychological Models", Proc. of ICME 2004.

[6] D. Zhang, S. Chang, et al, "Accurate Overlay Text Extraction for Digital Video Analysis", Proc. of ITRE 2003.

[7] Di Zhong, "Segmentation, Index and Summarization of Digital Video Content", PhD Thesis Graduate School of Arts and Sciences, Columbia University, 2001

[8] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods", Advances in Large Margin Classifiers, MIT Press, 1999.

[9] Pevzner, L, and M. Herst, "A Critique and Improvement of an Evaluation Metric for Text Segmentation", Computational Linguistics, 28 (1), p.19-36, 2002

[10] M. Mizutani, et al, "Commercial Detection in Heterogeneous Video Streams Using Fused Multi-Modal and Temporal Features", ADVENT Technical Report No. 204-2004-4, http://www.ee.columbia.edu/dvmm/newPublication.htm