Commercial Detection in Heterogeneous Video Streams Using Fused Multi-Modal and Temporal Features

[1,2] Masami Mizutani, [2] Shahram Ebadollahi, [2] Shih-Fu Chang

mizutani.masami@jp.fujitsu.com
mizutani@ee.columbia.edu

shahram@ee.columbia.edu

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 frame, which usually indicate the start of the 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 on WD (WindowDiff [11]) metric over just using the blank indicators. For the case when blank indicators do not exist, our approach results in almost 85% correct detection.

Table of Contents

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 8). In [5], 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 one obtained from detecting 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 only tested on videos comprised of news and commercials only and therefore does not necessarily scale to heterogeneous data set.

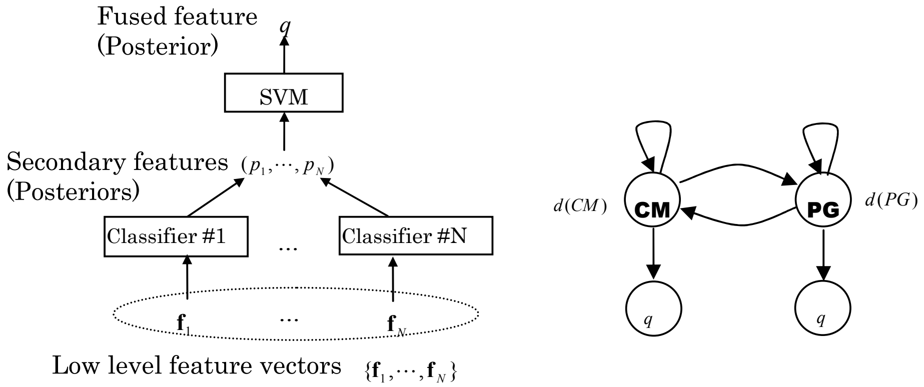
We aim to develop an algorithm for detecting commercial segments in video streams obtained from heterogeneous sources, which not only improves upon simple yet effective blank detection but also provide reliable detection performance when blank indicators are not available.

To achieve our goal, we systematically fuse audio/visual/temporal local features of commercials in the context of their global temporal characteristics (Figure 1). 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 (Figure 1.a). The results of the fusion of the decisions are then used as the probabilistic outcomes of a generative process modeling the global temporal characteristics of the commercials (Figure 1.b).

In this approach, 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[1]. We believe that it is important to use such general interest channels for conducting experiments on commercial detection in order to show that one is effectively learning both the global and local patterns of the commercials. Using binary-class data sets (i.e. news vs commercial) is an easy case, because commercials tend to display very distinct patterns from those obtained for the program segments.

[1] ABC, CBS, FOX, NBC, UPN, WB11. See Appendix B for the detail.



(a) Fusion method by discriminators

(b) Duration dependent generative model

Figure 1. System Framework: The fused feature (posterior) q is handled as the observation.

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 on WD (WindowDiff [11]) metric over the case that only uses blanks for commercial detection. We also show that the global temporal characteristics significantly improve the results compared to the results using just classifiers.

Most importantly, the fusion of the local and global characteristics of the commercials provides acceptable results when there are no blank frames to flag the start of commercials. Figure 8 shows that the blank frames are not consistent across the data set and on average our method which uses other local features provides superior performance.

In the remainder of the paper we first provide our observation of the characteristics of commercials in section2 and the definition of the problem we are trying to solve in section 3. We then discuss the modeling of individual detectors for commercials and the fusing of the individual decisions in section 4. Finally, we report our experiments and results in section 5 and conclude in section 6.

2. CHARACTERISTICS OF COMMERCIALS

Our observation says that, if we look at the entire video stream, commercials reveal a global temporal characteristic, which is the temporal spacing between the occurrences of commercial 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.

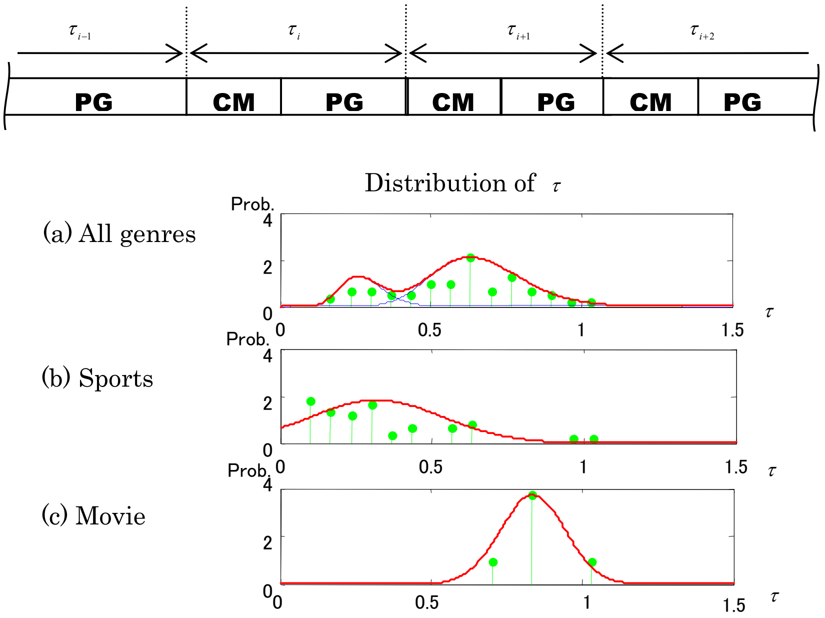


Figure 2. Inter-arrival times of commercial segments for 2 distinct genres and for all genres put together.

2.1. Global Characteristic

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. For example, during a sports program (e.g. Basketball), commercials are often inserted when there is a break in the game, and during a movie program, they are inserted in a more or less regular points in time.

Figure 2.b and Figure 2.c show the difference between the inter-arrival times of the commercial segments during the 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.

These constraints together impose a distinct distribution of the inter-arrival times between the commercial segments throughout the video stream. Therefore, the pattern of the occurrences of the commercials throughout the program stream makes its global characteristic with respect to the entire program stream (Figure 2).

2.2. Local Characteristics

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 viewer's attention. For example, excessive utilization of overlay text during the commercials is a method for both capturing users' attention and for conveying the information. This text occurs in locations and in forms which are in a probabilistic way different from other types of programming.

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 use a set of features that we think might be useful in quantifying the nature of commercials. These features are "audio class", "color mood" [8], "characteristics of overly text" [9], and the "dynamics of the scene change" [10]. These features are discussed in the later section.

3. PROBLEM STATEMENT AND APPROACH

In this work, we assume that the locations of the scene changes (blank frames also regarded as scene change) are given. This is a reasonable assumption because the commercial/program boundaries are characterized by such abrupt transitions. Scene change detection usually performs well on these kinds of transitions [10].

We pose the problem of commercial detection as one of finding the scene changes of which the local features in observation windows and the global feature with respect to the previous commercial segments resemble the characteristics of the commercial segments. We take a combined generative and discriminative approach to this problem.

3.1. Generative Approach

We model the program stream by a two-state first-order Markov chain alternating between commercial (CM) and program[2] (PG) segments according to a certain transition probability and an explicitly modeled duration of stay in each state (Figure 3).

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 either due to CM or PG state of the system. The problem of commercial segment detection is therefore transformed into that of inferring the optimal sequence of the hidden states of this duration dependent HMM model. This can be solved by the duration Viterbi Algorithm.

The explicit duration model filters the state sequence to better correspond to the global characteristic of the commercials throughout the program stream. Modeling the duration of stay in different states of an HMM was previously used [6][7] in speech recognition. We model the duration density functions as a uniform distribution for CM state and as a mixture of Erlang models (EMM) for PG state.

[2] We sometimes call it as non-commercial segment in this paper.

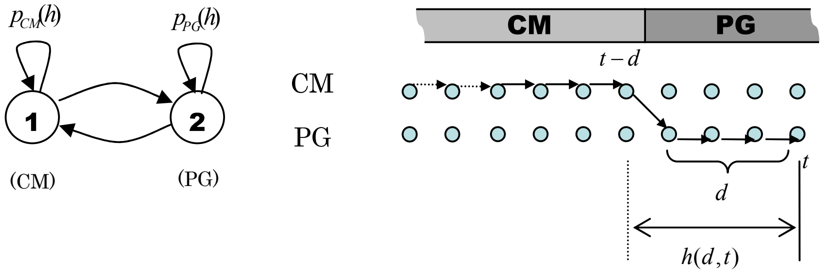


Figure 3. Duration Dependent HMM and Duration Viterbi: take into account of the actual duration

3.1.1. *Inference by the duration Veterbi Algorithm*

As mentioned before, the occurrence of commercial segments throughout the program stream obeys a certain pattern dictated by both the genre of the program they occur in and the content planning policies of the broadcast networks. We try to capture the inter-arrival time between two consecutive commercial segments (Figure 3). This is approximated to capturing the duration of commercial and program segments. The probability of the duration of commercial and program segments are embedded in the duration Viterbi algorithm [7] to take into account the global characteristic. The formulation of the duration Viterbi algorithm is as follows:

$$\delta_1(j) = \pi_j b_j(o_1)\,,\ \eta_1(j) = 0, \psi_1(j) = 0 \ (i, j = PG, CM)$$

$$\delta_t(PG) = \max\left[\pi_{PG} p_{PG}(h(t,t)) \prod_{\tau=1}^{t} b_{PG}(o_\tau),\ \max_{d=1}^{t-1}\left[\delta_{t-d}(CM) a_{CMPG} p_{PG}(h(t,d)) \prod_{\tau=0}^{d-1} b_{PG}(o_{t-\tau})\right]\right] \qquad \text{Eq. 1}$$

$$\delta_t(CM) = \max_{d=1}^{t-1}\left[\delta_{t-d}(PG) a_{PGCM}\, p_{CM}(h(t,d)) \prod_{\tau=0}^{d-1} b_{CM}(o_{t-\tau})\right] \qquad \text{Eq. 2}$$

$$\eta_t(j) = d^*, \psi_t(j) = i^*$$

As the conventional Viterbi algorithm [7], $\delta_t(j)$ is the maximum likelihood of starting from the initial state at time 1 and ending in state j at time t when observing the sequence $o_1 o_2 \cdots o_t$. π_j is the initial probability that the state starts with state j, and $a_{i,j}$ is the transition probability from state i to state j.

Although $b_j(o)$ is originally the likelihood of feature vector o given state j, it can be replaced with the posterior probability of state j given feature vector o using Bayes rule as shown in Equation 3 with equal priors because the number of occurrence of the reguralizer from time 1 to t are the same regardless of any state in Equation 1 and 2.

$$P(CM \mid o) = \frac{1}{1 + \dfrac{P(o \mid PG)P(PG)}{P(o \mid CM)P(CM)}} \qquad \text{Eq. 3}$$

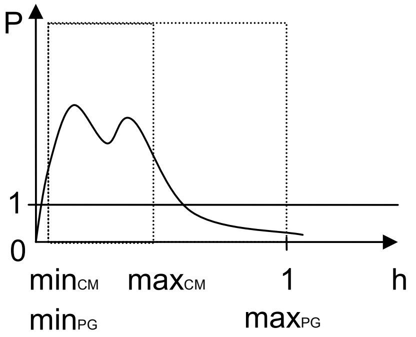


Figure 4. Duration probability models: the uniform distribution for CM and the EMM for PG are bounded by their minimum and maximum durations

t is a time step corresponding to the actual position in time of the t-th scene change boundary. $p_j()$ is the estimated density function of the duration of state j. d is the entering count staying at the same state, however, the actual time of duration d is evaluated in the probability model $p_j()$. For the sake of convenience, we define the function $h(t,d)$ which returns the actual duration of d shots before the t-th shot boundary normalized by the estimated maximum duration of all states as shown in Figure 4. The probability $p_j(h)$ is modeled as uniform distribution for CM, and as EMM for PG. They are bounded by the minimum and maximum durations of each state (Figure 4).

The mixture model is used to better capture the different nature of the inter-arrival times of the commercials inserted in different genres of programming. It is appeared in the duration model of program segments. On the other hand, as the duration of each individual commercial obeys a certain pattern dictated by the duration of the time slots sold by the carriers to the commercial owners, it is better to consider the duration of commercial segments follow the uniform distribution.

These two duration models for commercial and program segments are learned independently on genre. Especially, the parameters of EMM are learned by Expectation Maximization (EM) algorithm [12], which is detailed in Appendix A. Initial parameters are given by random. The number of mixtures is varied from 1 to 4 and the best is chosen by evaluating BIC (Bayesian Information Criteria). The estimated distribution is actually well fitted; the goodness of fitting is evaluated by Kolmogorov Smirnov test and we confirm null hypothesis can not be rejected at the significance level 5%.

3.2. Discriminative Approach

At each scene boundary location (each scene change location is a candidate for being either start or end of a commercial segment), we extract the local audio/visual/temporal features for a

15-second window placed after the boundary (Figure 5). The choice of 15-second window was made to be able to capture the thematic features of an average length of commercial segments. Features such as the mood of the commercial segment or the pattern of the location of the overlay text are better captured for long duration windows than the conventional short ones.

In addition, we extract the feature with regard to the blank frame for a 120-second window placed surrounding the boundary. The blank frame is considered as a good indicator of starting commercial. The choice of 120-second window was made to be able to include the boundaries of several commercial clips before and after the candidate point.

As a method to treat the derived features as the observations in the generative model, we use multiple discriminative classifiers which are trained to distinguish between CM and PG segments using one kind of related features obtained from a local window (Figure 1.a). The resulting margins from each discriminative classifier are transformed into the posterior probability of each of the two possible hidden states [11].

We deploy another SVM classifier for fusing the observation vectors of the posterior probabilities. The resulting margins from the fusion SVM classifier are transformed into the posterior probability in the same manner. The fused feature derived can be treated in the conventional and also duration Viterbi framework instead of the observation likelihood $b_j(o)$ in Equation 1 and 2.

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 correct location of commercial segments.

.

4. FEATURE EXTRACTION AND FUSION

Commercials are made to be interesting and eye-catching. There are certain tools at the disposal of the directors in order to achieve this goal [1]. In this work, we select a number of audio/visual/temporal local features that we believe to contribute to the “appeal” of the commercials (Figure 5). These are by no means a comprehensive set of features obtained in a principled way. For that, one has to conduct rigorous psychophysical studies. Below, we give a description of the features that we use and the way they are modeled in this work. These features were used by other researchers in similar or other contexts, and we do not have any claim on their novelty. Their combination, however, proves to be effective for capturing the characteristics of commercial segments and distinguishing them from other program segments in the program stream.

In the following, we first describe the local features used to capture the appealing aspects of the commercials and then will describe the method for fusing their decisions.

4.1. Features

4.1.1. Audio Class Histogram (ACH)

To capture characteristic of commercials, audio is important. We use an audio class classifier so that it can classify the sound during one second into either one of four classes: silence, speech, music and music/speech. The performance is shown in Table 1. We compute the count of one second units having each audio class in the 15-second window. Since there is much variation between different commercials, it is unrealistic to try to model the audio class transition during a commercial segment. Instead, we obtain the histogram of audio classes placed after the scene boundary. This histogram is used as the thematic feature for the entire duration commercial.

4.1.2. Commercial Pallet Histogram (CPH)

Color has been shown to have a direct relationship to the "mood" of the scene in movies [8]; i.e. the director uses certain colors to set the atmosphere of, for example, a horror scene or a happy scene. If we assume that there is a relationship between the "appeal" of the commercials and the mood and the atmosphere that they depict, then we can use a similar features to the "movie pallet histogram" suggested in [8] to capture the mood of the commercials.

We derive the "pallet histogram" in the 15-second observation window placed after the scene boundary. The three dominant colors of a keyframe in a shot are merged with weights, and it is quantized to one of the predefined number of pallet colors, which equally divide L*u*v* color space. We use 12 pallet colors. The histogram of the pallet colors within the observation window is constructed and used as a feature in further process. The total number of the colors in the histogram depends on the number of the detected shots in the observation window.

4.1.3. Text Location Indicator (TL)

In most cases, commercials contain overlay texts in order to point to important information. The characteristics of the overlay text in a commercial segment are different from those occurring in other types of programs. To capture this characteristic, we use the overlay text detector [9]. The performance is shown in Table 1. The result of the detector are obtained in every five frames and then mapped to a 16x16 binary grid (a grid element has a value of 1 if text is detected in the grid area). The cumulative grid through all frames in the 15-second window is then obtained, which captures the locations and frequencies of any texts occurred in the observation window.

4.1.4. Scene Change Rate (SCR)

In [2], the rate of scene changes in commercials was used for their detection in a heuristic manner. It was assumed that the rate of scene changes was always higher for commercials compared to other types of programs. This is generally not the case. One can often see commercials with very few scene changes. Here we model the scene change rate probabilistically for both commercials and program segments.

The scene change detector in [10] is used to obtain the scene boundaries and then the rate of scene changes is obtained for the observation window. The performance of the detector is shown in Table 1. Based on our empirical simulations, we found the scene change rate can be

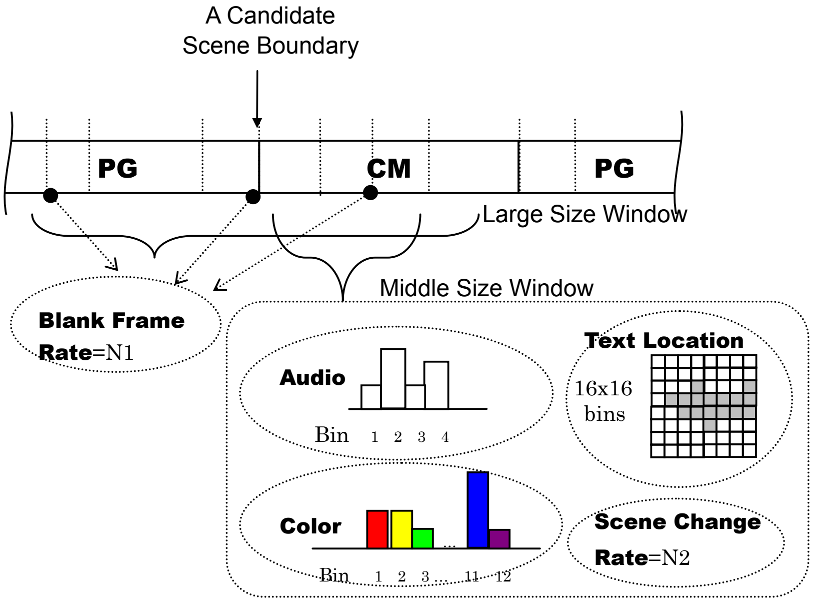


Figure 5. Local features of commercials: in the two observation windows placed a candidate point

adequately modeled by Poisson distribution.

For the observation window, we obtain the two likelihood values of it being commercial and program class based on its scene change rate. The likelihood values are used in the fusion stage to make a decision on the classification. The performance of this classifier is shown in Table 2.

4.1.5. Blank Frame Rate (BFR)

Blank frames are very good indicators of commercial segments when available [2]. We take advantage of the blanks as one of the features for detecting commercials. The blank frame is detected by evaluating the average and standard deviation of all the pixel intensity in the frame based on the following equations:

$$BlankFrame\ Indicator = \begin{cases} 1 & if\ \ mean(I) < 8\ \&\ std(I) < 20\ |\ mean(I) < 5\ \&\ std(I) < 30 \\ 0 & otherwise \end{cases} \qquad \text{Eq. 4}$$

If the average and the standard deviation below the thresholds, and if the average take the minimum values within a 30-frames window surrounding the evaluating frame, the frame is declared as the blank frame.

In this work, the probabilities of occurrence of the blank frames within a relatively large window in commercial and program segments are modeled using Poisson distribution, respectively. We use a 120-second window surrounding a candidate scene change boundary. 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.

For a scene boundary location using the models, we find the two likelihood values of the blank frame rate being a part of commercial and program class. The likelihood values are used in the fusion stage to make a decision on the classification. The performance of this classifier is shown in Table 2.

Table 1. Performance of low level detectors

	Audio class Classifier	Overlay Text detector	Scene change detector
Precision (%)	90 : 90 : 65 : 61	94	82

* Audio classes: silence, speech, music, music/speech

** Evaluated by sampling on our experiment data

4.2. Secondary Features

Above we described the extraction of different features around the candidate points. Because those features are different in nature, we derive a set of secondary features from them that well be used as the observation vectors in the commercial detection process.

For each of ACH, CPH and TL, we train a SVM classifier in order to distinguish between the patterns of those features for the observation window taken from commercial and program segments. RBF (Radial Basis Function) is selected for the kernel. As shown in Figure 6, all 15-second windows which reside in segments of each class are used as the training data for the class.

The confidence values for the two classes are obtained as the results of the application of the SVM classifiers. The performance of the SVM classifier for each kind of features is shown in Table 2. In order to treat the outputs more conveniently and consistently with other features, we converted them to the posterior probability using a sigmoid function as addressed in [11].

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 given the feature vector using Bayes rule with equal priors (Equation 3).

Finally, each feature is now transformed into the posterior probability of commercial and these 5 posterior probabilities collectively make the new secondary feature vectors x, which is summarized in Table 2

4.3. Fusion

We train another SVM classifier with RBF in order to fuse the derived secondary features. The secondary features which reside in segments of each class are used as the training data for the

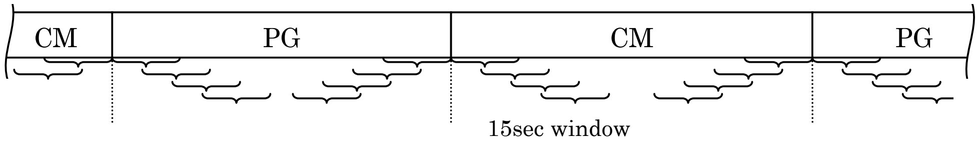


Figure 6. Training Data Location for discriminative classifiers.

class. The confidence value from the SVM classifier is transformed to the posterior probability in the same manner, which is used as the 'fused' feature in Viterbi frameworks.

Table 2. New Secondary Feature Vectors.

Variable	Original	Meaning	Classifier / Model	Range	F1*	Recall/Precision
x_1	ACH	Posterior of CM (=1-PG)	SVM, sigmoid	0～1	75.0%	78.3% / 72.0%
x_2	CPH	Posterior of CM (=1-PG)	SVM, sigmoid	0～1	66.2%	75.1% / 59.2%
x_3	TL	Posterior of CM (=1-PG)	SVM, sigmoid	0～1	65.7%**	68.2% / 63.4%
x_4	SCR	Posterior of CM (=1-PG)	Poisson, Bayes rule	0～1	75.6%	77.4% / 72.0%
x_5	BFR	Posterior of CM (=1-PG)	Poisson, Bayes rule	0～1	87.4%	94.3% / 81.3%

* Evaluated on all 49 test streams across 3 rounds.

** Text is evaluated only where it exists.

5. EXPERIMENTS AND RESULTS

5.1. Training and Test Procedure

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) point decision, 2) inference using the conventional Viterbi, 3) inference using the modified Viterbi to reflect explicit duration of the states.

We use 36 hours MPEG-1 streams which consist of 49 test program streams in the United States across 6 general interest channels as listed in Appendix B. The streams consist of 6 genres: news, drama, animation, movie, entertainment, and sports. Commercial occupies about 25%, 9 hours. The specification of MPEG-1 video is 1.6Mbps bitrate, 30fps frame rate and 352x240 pixel size, and that of MPEG-1 Audio layer II is 192kbps bitrate, 32kHz sampling rate and stereo channel. These qualities are almost same with experimental data used in TREC-VID2003 [14].

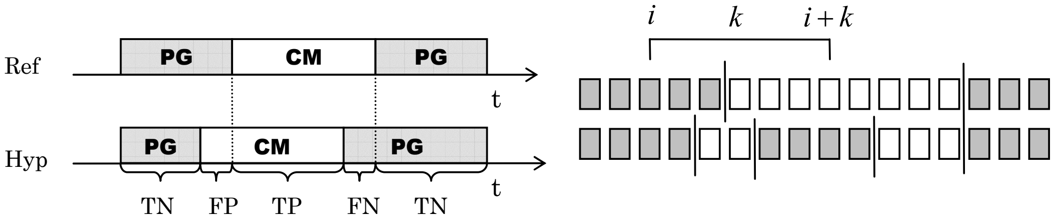


Figure 7. F1 (left) and WD (right) Metrics.

Ground truth of the commercial segment boundaries are chosen and labeled by manual among the scene change boundaries. Very short films are treated as program segments such that a logo of a film production company sometimes appears for a few second between commercial and program segments.

The experiment data is exclusively divided into 3 rounds and the two of them are used as training data across multiple channels and multiple genres, and the rest is as test data (Table 3). For example, at the 1st round, 32 program streams are used as training data and 17 program streams as test data. Totally, 49 programs for about 36 hours including 9 hours commercial are tested across 3 rounds. In training phase, the 3-fold cross validation is carried out with the leave-one-out method in order to avoid overfitting parameters on models.

We use the performance F1 metric in introduced in [4]. Recall and precision rates are defined as follows:

$$\begin{aligned}&\text{Recall} = \text{TP}/(\text{TP}+\text{FN}),\ \text{Precision} = \text{TP}/(\text{TP}+\text{FN})\\&\text{F1} = 2\times\text{Recall}\times\text{Precision}/(\text{Recall}+\text{Precision})\end{aligned} \qquad \text{Eq. 5}$$

TP (True Positive) and TN (True Negative) are correctly detected as commercial and program segments respectively, and FP (False Positive) and FN (False Negative) are incorrectly detected vise versa. The number of the candidate scene changes is used for the each period in Figure 7 in order to show how much improvement is archived from the SVM point decision.

As a problem, the recall and precision metric does not reflect well the small false alarms within the detection result. As a complementary metric, we also use WindowDiff (WD) [13] that is usually used in the research of text segmentation to see how many discrepancies occur between ground truth and detection result.

$$WindowDiff(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|\,b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})\,| > 0) \qquad \text{Eq. 6}$$

ref is the ground truth of the test sequence and hyp is the detection result. $b(i, j)$ represents the number of commercial and program boundaries between scene change position i and j. N represents the number of scene change positions in the sequence. In our experiment, k is set to the half of the average size of commercial and program segments in the sequence. The equation gives a score between 0 and 1, and the detection result which has all the boundaries correctly receives a score of 0.

Table 3. Training and test condition: 49 programs are divided into 3 rounds.

	News	Drama	Animation	Entertain.	Movie	Sports	Total
All round	11 (437)	29 (972)	3 (89)	4 (145)	1 (119)	1 (326)	49 (2088)
Training data							
R1	4 (131)	21 (710)	2 (60)	3 (115)	1 (119)	1 (326)	32 (1461)
R2	7 (306)	18 (619)	3 (89)	3 (116)	1 (119)	1 (326)	33 (1575)
R3	11 (437)	19 (615)	1 (29)	2 (59)	0	0	33 (1140)
Test data							
R1	7 (306)	8 (262)	1 (29)	1 (30)	0	0	17 (627)
R2	4 (131)	11 (353)	0	1 (29)	0	0	16 (513)
R3	0 (90)	10 (357)	2 (60)	2 (86)	1 (119)	1 (326)	16 (948)

* Number of programs (minutes)

5.2. Results

5.2.1. Effectiveness of fusion

The detection results on all 49 test streams across 3 rounds are shown in Table 4 shows that even using only the blank feature gets the pretty good result (F1=87%, WD=23%). This means that blank frames are good indicators of commercial segments and aggregating blank frames evaluating BFR is very meaningful for detecting commercial. Since BFR is already a temporal feature based on a relatively large observation window, there is not very much significant improvement on both F1 and WD after two kinds of Viterbi smoothing.

The result (b-1) shows that, even not using such a blank feature, local features can capture the characteristics of the commercial segments (F1= 80%). However, the WD score (71%) means there is a lot of small fragmentation of false alarms and boundary discrepancies, because we are using a comprehensively heterogeneous data set and the synthetic variation of currently considering features is so diverse in commercials and programs. For example, in the ending credit of a certain drama program, telops flows on the right side in the display and a commercial is already starting on the other side. We labeled such an ambiguous segment as a program segment in that case; however, the local characteristic of it must be very similar to that of commercial segments. Due to such a reason, it is inevitable that the result (b) of the point decision tend to have such a bad WD score.

However, the result (b-2) shows such false alarms can be eliminated and the WD score was largely improved to 31% from 71% using the conventional Viterbi. Moreover, the result (b-3) shows that the duration Viterbi works better (WD=25%) and is better to capture that commercials display certain global and local features regardless blank.

We tried another experiment applying specific duration models to the duration Viterbi that are learned dependently on genres (only news and dramas due to the shortage of training data) instead of using the general duration model learned independently on genre. However, the improvement was within 1% on both F1 and WD from the case of using duration models learned independently on genres, and there was not very much significant difference.

The result (c-1) shows that using all features is most excellent on F1 with regard to the point decision, however, it is inferior on WD to the case using only BFR. This means fusing all features increases the number of scene changes correctly classified, but the newly correctly classified 'seeds' tend to be fragmented due to the nature of the other features as we see in the result (b-1). However, such seeds bring better detection result after the two kinds of Viterbi smoothing. The result (c-3) is the best results (WD=15%, F1=92%), although there was not very much significant difference between result (c-2) and(c-3), because the point decision is accurate enough for the conventional Viterbi to eliminate small fragments.

Table 5 shows the F1 in the case where the actual time is taken into account. Generally, there are drops of about 6% in comparison with Table 4, because the correctly and falsely classified scene change boundaries have the different number of frames.

5.2.2. Consistency of the performance

Figure 8 shows the histograms of F1 (upper) and WD (bottom) of detection results when using the three kinds of feature sets with the duration Viterbi. The widths of the histograms are both 5%.

The two graphs show that the case of using only the blank feature achieves the good performance on F1 and WD, however, it is not always reliable compared with the case using all the features. The lack of blank frames in commercial segments or the fail to detect them gives a large loss on the performance (F1<70%). Although using all the other features without BFR also achieves good performance in most program streams, the bottom graph shows some detection results tends to have many false fragmentations which do not appear in F1 as addressed the reason. However, when all the features are fused in our framework, it can be expected that they are working well complementarily with BFR for our purpose.

5.2.3. Analysis of false alarms

We examined the false alarms in case of using our approaches. We found the detected parts as commercial (FP) in program segments is similar to a certain type of commercials. One of them is in an entertainment genre program in NBC channel which provides the information on artists and leisure events with rhythmical music and eye-catching visual effects. However, this result can be considered as reasonable because our probability models are designed to capture the parts which have high likelihood of commercials in terms of the activity of audio, color, overlay text location and temporality. We will further explore to make the most of this ability for the application of commercial detector in our future works.

Table 4 Results of commercial detection in the case where the number of scene change is under consideration: Using all features in the context of global characteristics gives the best results (lower-right element)

	a) Only blank feature		b) Other features		c) All features	
Metric	F1	WD	F1	WD	F1	WD
1) Point Decision	87% (94%, 81%)	23%	80% (85%, 75%)	71%	90% (94%, 86%)*	38%
2) Viterbi	87% (95%, 81%)	22%	84% (89%, 79%)	31%	91% (96%, 87%)	16%
3) Duration Viterbi	89% (95%, 83%)	21%	85% (90%, 80%)	25%	92% (95%, 88%)	15%

*Recall and Precision rates

Table 5 Results in case where actual time is under consideration: the performance is dropped by almost 6% against Table 4

	a) Only blank feature	b) Other features	c) All features
Metric	F1	F1	F1
1) Point Decision	81% (▲6%)	73% (▲7%)	86% (▲9%)*
2) Viterbi	81% (▲6%)	78% (▲6%)	88% (▲3%)
3) Duration Viterbi	83% (▲6%)	78% (▲7%)	87% (▲5%)

* Difference from Table 4

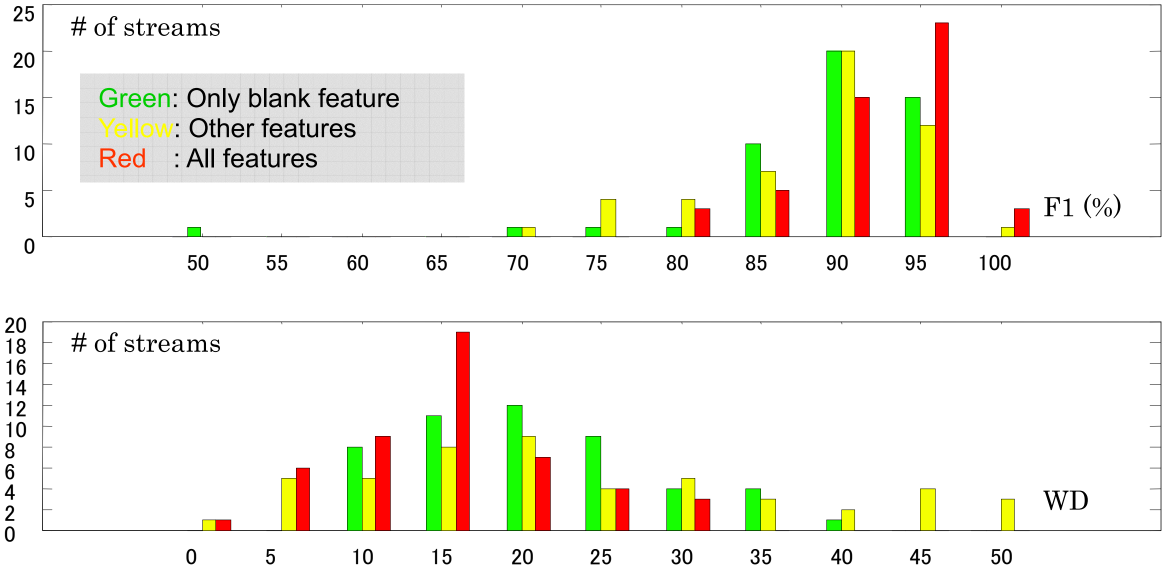


Figure 8. Histograms of detection results in case of using the duration Viterbi: The upper graph is on F1, and the bottom is on WD. Blank features although powerful for commercial detection, are not consistent

.

6. CONCLUSION AND FUTUER WORK

In this paper, we described a systematic approach for detecting commercials 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 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 (F1=92%, WD=15%). Even when the blank indicators are not used, the framework results in the acceptable commercial boundary detection (F1=85%, WD=25%).

This graph in Figure 8 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. 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're available and enhance upon them by using other features as described in the paper.

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 and the reported results are subject to this selection.

We also like to apply this framework to the TRECVID [14] data set, which is binary in nature (e.g. news vs. commercials) to be able to better compare our results against those reported by others on the same data set.

7. ACKNOWLEDGMENTS

The first author would like to thank all the members of DVMM laboratory at Columbia University for their suggestions and discussions.

8. REFERENCES

[1] Elizabeth J. Heighton, Don R. Cunningham, “Advertising in the broadcast media”, p.97-120, Wadsworth Publishing Company, 1976

[2] S. Marlow, D. A. Sadlier, K. McGeough, N. O’Connor, N. Murphy, “Audio and Video Processing for Automatic TV Advertisement Detection”, Proceedings of ISSC, 2001

[3] A. Hauptmann, M. Witbrock, “Story Segmentation and Detection of Commercials in Broadcast News Video”, Advances in Digital Libraries Conference (ADL’98), Santa Barbara, CA, April 22-24, 1998

[4] L. Agnihotri, N. Dimitrova, T. McGee, “Evolvable Visual Commercial Detector”, in the proceeding of CVPR’03, Vol. II. p.79, 2003

[5] P. Duygulu, M. Chen, A. Hauptmann, “Comparison and Combination of Two Novel Commercial Detection Methods”, Proceedings of ICME, 2004

[6] H. Gu, L. Lee, “Isolated-Utterance Speech Recognition Using Hidden Markov Model with Bounded State Durations”, IEEE Trans. on Signal Processing, Vol. 39, No. 8, 1991, 1743-1752.

[7] L. Rabiner and B.-H. Juang. “Fundamentals of Speech Recognition”, Prentice Hall Signal Processing Series, Prentice Hall, Englewood Cliffs, New Jersey 07632, 1993. A. V. Oppenheim, Series Editor.

[8] C. Wei, N. Dimitrova, S. Chang, “Color-Mood Analysis of Films Based on Syntactic and Psychological Models”, Proceedings of ICME 2004.

[9] D. Zhang, B. Tseng, C. Lin and S. Chang, "Accurate Overlay Text Extraction for Digital Video Analysis", Proceeding of IEEE International Conference on Information Technology: Research and Education (ITRE 2003).

[10] Di Zhong, “Segmentation, Index and Summarization of Digital Video Content”, http://www.ee.columbia.edu/~dzhong/

[11] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods", in A. J. Smola, P. Bartlett, B. Scholkopf and D. Schuurmans, Advances in Large Margin Classifiers, MIT Press, 1999.

[12] N. Vasconcelos, A. Lippman, “Statistical Models of Video Structure for Content Analysis and Characterization”, IEEE Transactions on Image Processing, vol. 9, n. 1; January 2000

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

[14] TRECVID 2003 Guidelines, http://www-nlpir.nist.gov/projects/tv2003/tv2003.html

[15] Jeff Bilms, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," ICSI TR-97-021 April 1998

[16] “TV Anytime, Specification Series: S-3 on Metadata (Normative), Appendix B”, Feb, 2001

APPENDIX A EM ALGORITHM FOR EMM

As [15], we can define an auxiliary function Q for the mixture of any general distribution as follows (the meanings of symbols are consistent with [15]):

$$Q(\Theta,\Theta^g) = \sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l \mid x_i,\Theta^g) + \sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(x_i \mid \theta_l))p(l \mid x_i,\Theta^g) \cdots (1)$$

where $p(l \mid x_i,\Theta^g) = \dfrac{\alpha_l^g p_l(x_i \mid \theta_l^g)}{\sum_{l=1}^{M}\alpha_l^g p_l(x_i \mid \theta_l^g)}$

We can solve α_l without any concrete distribution, which give the maximization of Q function in the same way in [15].

$$\alpha_l = \frac{1}{N}\sum_{i=1}^{N} p(l \mid x_i,\Theta^g)$$

Erlang distribution is now as follows:

$$p_l(x_i \mid \theta_l) = \varepsilon_{r_l,\lambda_l}(x_i) = \frac{\lambda_l^{r_l} x_i^{r_l-1} e^{-\lambda_l x_i}}{(r_l-1)!}$$

Taking the log of it, the following equation is derived:

$$\log(p_l(x_i \mid \theta_l)) = r_l \log\lambda_l + (r_l-1)\log x_i - \lambda_l x_i - \log((r_l-1)!)$$

Taking derivatives of (1) with λ_l and setting 0, we can easily get the following equation.

$$\sum_{i=1}^{N}\left(\frac{r_l}{\lambda_l} - x_i\right)p(l \mid x_i,\Theta^g) = 0$$

Finally, we get λ_l:

$$\lambda_l = \frac{r_l \sum_{i=1}^{N} p(l \mid x_i,\Theta^g)}{\sum_{i=1}^{N} x_i p(l \mid x_i,\Theta^g)} \cdots (2)$$

Taking derivatives of the equation (1) with r_l is difficult, because it includes factorial of r_l, which is discrete. As the purpose is to solve r_l which give the maximization of function Q, we can find it from the following equation by searching the discrete values for r_l in a specific range exhaustively. (i.e. r_l=1,..,20)

$$\begin{aligned} r_l &= \arg\max_{r_l} Q(\Theta,\Theta^g) = \arg\max_{r_l} \sum_{i=1}^{N}\log(p_l(x_i \mid \theta_l))p(l \mid x_i,\Theta^g) \\ &= \arg\max_{r_l} \sum_{i=1}^{N}\{r_l \log\lambda_l + (r_l-1)\log x_i - \lambda_l x_i - \log((r_l-1)!)\}p(l \mid x_i,\Theta^g) \end{aligned}$$

where λ_l is given by the equation (2)

APPENDIX B PROGRAM GUIDE

Totally, 49 program streams for 36 hours are used as listed in Table 6. The boundaries of programs do not always meet in the unit of 30 minutes or 1 hour. Commercial occupies about 9 hours, 25% of all the data. The genre definition is based on TV anytime forum [16].

Table 6 Program guide

CH(date)	6:00PM	6:30PM	7:00PM	7:30PM	8:00PM	8:30PM	9:00PM	9:30PM	10:00AM	10:30PM	11:00AM	11:30PM
WB11 (Fri. 3/12/04)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	INFO (News)		DRAMA (SitCom)	DRAMA (SitCom)
UPN9 (Sat. 3/13/04)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	Movie				INFO (News)		DRAMA (SitCom)	–
FOX5 (Sun. 3/14/04)	INFO (News)	Animation	Animation	DRAMA (SitCom)	Animation	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA	INFO (News / Sports News)		DRAMA (SitCom)	DRAMA (SitCom)
NBC (Tue 3/16/04)	INFO (News)	INFO (Nightly News)	ENT (Gossip)	ENT (Gossip)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA (SitCom)	DRAMA	DRAMA		INFO (News)	–
	12:00PM	12:30PM	1:00PM	1:30PM	2:00PM	2:30PM	3:00PM	3:30PM	4:00PM	4:30PM	5:00AM	5:30PM
ABC7 (Mon. 3/15/04)	INFO (News)	ENT (Quiz)	DRAMA		DRAMA		DRAMA		ENTERTAIN (Talk Show)		INFO (News)	
CBS2 (Thurs. 3/18/04)	INFO (News)	Sport Event (NCAA Basketball Tournament)										INFO (News)

Genre description

1 Blue:	Information	(Subtype: Daily News, Sports News, Politics/National Assembly, Others)
2 Red:	Drama	(Subtype: SitCom, Comedy, General light drama)
3 Yellow:	Animation	
4 Green:	Entertainment	(Subtype: Quiz, Talk show, Gossip)
5 Light Blue:	Movie	
6 Light Green:	Sport Events	(Subtype: Basketball)