Comparison and Combination of Two Novel Commercial Detection Methods

Pınar Duygulu
Department of Computer Engineering
Bilkent University
Ankara, Turkey, 06800

Ming-yu Chen and Alexander Hauptmann
Informedia Project
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Abstract

Detection and removal of commercials plays an important role when searching for important broadcast news video material. In this study, two novel approaches are proposed based on two distinctive characteristics of commercials, namely, repetitive use of commercials over time and distinctive color and audio features. Furthermore, proposed strategies for combining the results of the two methods yield even better performance. Experiments show over 90% recall and precision on a test set of 5 hours of ABC and CNN broadcast news data.

1. Introduction

In news videos, commercials are often inter-mixed with news stories. For efficient retrieval and browsing of the news stories, detection and removal of commercials are essential ([3, 4, 5, 6, 7]).

It is common to use black frames to detect commercials [3, 4]. However, such simple approaches will fail for videos of TV channels that do not use black frames to flag commercial breaks. Also, black frames used in other parts of the broadcast will cause false alarms. Furthermore, progress in digital technology obviates the need to insert black frames before commercials during production. An alternative makes use of shorter average shot lengths as in [6]. However, this approach depends strongly on the 'high activity' rate which may not always distinguish commercials from regular broadcasts.

In this study we propose two methods for commercial detection that use distinctive characteristics of commercials. In the first method, we exploit the fact that commercials tend to be repeated multiple times during various broadcasts. This observation leads us to detect commercials as sequences that have duplicates. The second method utilizes the fact that commercials also have distinctive color and audio characteristics.

Because the two methods capture different distinctive characteristics of commercials, they are orthogonal and complementary to each other. We propose strategies to combine the two different commercial detection algorithms which yield even more accurate results.

In Section 2, two methods proposed for commercial detection will be described separately. Then, in Section 3 the different strategies to merge the results of different detector results will be explained. Section 4 will present the detailed experiments. Section 5 will conclude with a summary of proposed work and discussion of the results.

2. Commercial Detection Methods

2.1. Commercials as Duplicate Sequences

Due to variability in shot segmentation, the same repeated commercial may have different numbers of detected shots, and the keyframes selected from each shot may also differ slightly. Therefore, the same commercial might appear as two different sequences as in Figure 1. The number of detected shots can be different due to missed shots in one of the sequences as shown in the top pair of Figure 1. Even if the lengths of the sequences are the same, the detetcted shots may be different as shown in the bottom pair of commercials. Furthermore, the extracted keyframes are often very similar but not identical. We define *duplicate sequences* as sequences that share identical or very similar consecutive keyframes with some missing keyframes allowed.

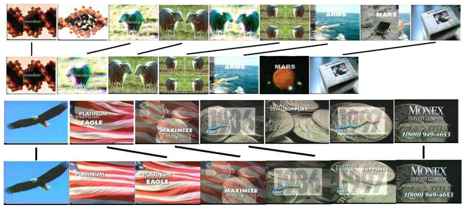


Figure 1: The variability of two example commercial pairs is shown. Matching frames are linked with lines.

We propose a heuristic pattern matching method for detecting *duplicate sequences*. The proposed method first detects *candidate repeating keyframes* (i.e. keyframes that have identical or very similar matching pairs) and then constructs the longest sequence that has consecutively similar keyframes with some missing elements allowed.

To detect *candidate repeating keyframes*, for each image in the data set, we find the most similar *N* images using feature similarity. *N* was limited to 50 to avoid some very common scenes of a TV channel (*e.g.* logo frames) that are shown in almost all news programs, analogous to stopwords in text. The similarity metric is based on these features: the average and standard deviation of HSV values obtained from a 5×5 grid; the mean values of twelve oriented energy filters (aligned uniformly with 30 degree separation) extracted from a 3×3 grid; Canny's edge detector results extracted from a 3×3 grid; and the size and position of frontal faces using Schneiderman's face detector algorithm [2].

If an image repeats itself *k* times, then we expect a discontinuity in the similarity values after *k* images. In order to catch this property, we take the derivatives of the similarity values. Then, we find the median of these values. The images are labeled as *candidate repeating keyframes* if the ratio between the largest value and meadian value is larger than a threshold (for the experiments the threshold is chosen as 100). The proposed method chooses the images in Figures 2(a)-(c) as *candidate repeating keyframes* and eliminates the rest. Frames in (a) and (c) have single similar images, and the keyframe in (b) has 8 similar images. (d) is eliminated since it is too common of a scene for weather news and repeats in almost all news programs. The image in (e) is from a regular news story. Therefore, it doesn't have duplicates and the discontinuity is not obvious.

Due to the errors in shot segmentation, similar sequences cannot be directly found by matching consecutive candidate keyframes. This is because interspersed with two matching candidate keyframes, there may be other keyframes that do not have any matching images. If we skip these non-candidate keyframes, and continue matching remaining candidates, then we have a chance to find a sequence which includes the missing keyframes. To detect matching sequences, the matching candidate keyframes are taken as the first elements of a possible matching sequence pair. The sequence is expanded only if there are other matching keyframes in close proximity. If such consecutively matching candidate frames are found, they are inserted as new elements into the matching sequences. Keyframes that are located in the interval between two inserted elements are also inserted to the sequences. This process repeats itself until no further matching pairs are found. This is performed for each candidate keyframe in the data set.

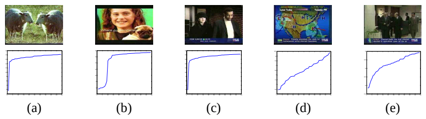


Figure 2: **Top:** Keyframe images, **bottom:** distances to the most similar 50 images.

2.2. Merging Color and Audio

Commercials have many distinctive characteristics in video and audio: news programs often have marks distinguishing them from commercials, like stock tickers. Most commercials contain background music while news contains mostly speech. Therefore, we can assume color and audio features are discriminative for commercials vs. news stories. In this study, a 5 by 5 125-bin HC square color histogram for images and the short time Fourier transform of 512 samples at 22050 kHz sampling rate for audio were used. The color histogram implictly includes a 'black frame' detection.

Both color and audio features are very diverse and abundant. Careful selection of distinctive features is important for decreasing noise which impairs the discriminative ability of a classifier and also for efficiency of computation. we use Fisher Linear Discriminant (FLD) for feature selection. The basic idea of FLD is to find the weighting of each dimension which maximizes the distance between different classes and minimizes the distance within the same class.

There are generally two choices for combining audio and image features: *feature synthesis* which merges different kinds of feature vectors into one integrated vector; or classifying different feature sets first and then combining the classification results into the final decision. Feature synthesis tries to represent the content of multiple media features as one integrated feature vector. It is a simple idea and an intuitive way to do the combination, but most experiments show that it does not perform well. The second approach classifies every feature set first and then combines the classified results. This approach tries to simplify the content of multiple media by assigning a higher level meaning to each set, by applying a binary classifier judgement to every feature set. We can then make discriminative decisions based on these judgments. The main drawback of this strategy is that detailed information contained in the feature sets is lost in the process of shrinking the dimensionality to one classifier result or judgment.

The basic idea of our combination approach is to obtain the benefits of both combination ideas. We apply FLD to every feature set and synthesize new feature vectors from every set. This step can be interpreted in two ways. First, it is feature selection. Second, it is like classification of the data since FLDs target function has an inherent ability to discriminate between classes of data. New feature vectors are not only selected from the raw data, but also generated by a discriminant function. Based on these new feature vectors, we construct a synthesized feature vector to represent the multimedia content and then apply classification to this representation. The details of the algorithm can be found at [9].

3. Combination Strategies

Our next challenge is to combine the two different commercial detection algorithms. In this section, we present two methods for combining the outputs of the two proposed commercial detectors. The first method is a heuristic which relies on the fact that the detected commercials are part of a sequence. The second method uses a high level SVM with the detection results of the two methods as input.

Sequence based: The simplest way to combine results of two different methods is to take the intersection of shots detected by both methods are true commercials. Since, the first method detects commercials as sequences, if an individual element of a sequence is known to be part of a commercial then the rest should be a commercial as well. Using this fact, we conservatively expand the intersection set of detected commercials, with all the elements of the sequences that have at least one element from the intersection set. Different commercial sequences usually occur grouped one after another. Therefore, if there is a small gap between two commercial sequences then there is a high likelihood that other commercials are in that gap. Thus in a final step of this strategy, frames that are labeled as highly likely commercials between already detected sequences are also labeled as commercials.

SVM based: It is possible to build a high-level classifier, in which the input features are the results of the two detection methods. Since the second method outputs confidence values, they can be directly used as input values. However, the first method only produces binary detection results. In this case, the length of the detected sequences can be used as a confidence value, since longer sequences are more likely to be a real commercial. As a second combination strategy, another classifier (Support Vector Machine) was built using these input values.

4. Experimental Results

The experiments are carried out on the data provided by the content-based video retrieval track (TREC-VID) of the 2003 Text Retrieval Conference [1]. The full data set consists of 120 hours of broadcast news videos from ABC World News Tonight and CNN Headline News from January through June 1998. Five news shows were selected for training and five for testing.

The common shot segmentations, defined by TREC-VID, are used as the basic units. One keyframe is extracted from each shot [8]. For CNN there were 411 keyframes labeled as commercials among 1362 training keyframes, while for ABC, there were 577 commercial keyframes among 1637 training keyframes. Logo images used for self-advertising of the news programs were not labeled as commercials.

Table 1 shows results for the first method on the CNN and ABC test sets. The detection results are compared for (i) taking detected candidate keyframes as commercials without finding sequences (indicated by *keyframe*) (ii) and, taking the elements of detected duplicate sequences as commercials (indicated by *sequence*). Results show that the power of the algorithm comes from detecting duplicate sequences, but not individual frames that have duplicates. In the next steps, we will only consider the elements of duplicate sequences as commercials detected by this method and the comparisons will be based on the results for *sequence*.

Figure 3 shows the confidence values of the second method for true commercials in the training sets. It can be observed that for CNN almost all commercials have very high confidence values, but in ABC some commercials have low confidence values. Table 2 shows the recall and precision values when the frames having higher values than a threshold is detected as commercials. We set the thresholds to either 0.5 or to the average confidence value of the true commercials in training set, which was 0.90 and 0.83 for CNN and ABC respectively.

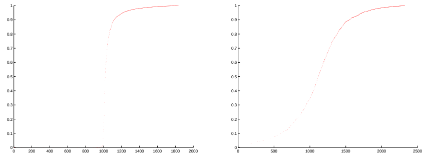
As can be seen from Tables 1 and 2, the second method performs better on CNN, while the first method does better on ABC. The goal of the proposed combination methods is to reach higher performance (higher recall or precision values) in both sets.

Table 3 shows the results of the first combination strategy. *Common* stands for the common frames obtained by using a simple intersection on detected commercials by two methods (with 0.5 as the threshold for the second method). Then, all frames in the sequences detected by the first method which include at least one element from this common set are taken as detected commercials (represented by *in-sequence*). The last step adds in the frames with high confidence values that lie between two sequences of commercials (represented by *final*). As the results show, all these steps produce higher precision compared to the individual results. For CNN, recall values are lower than the results of the second method when the threshold is set to 0.5, but higher than the results when threshold is set to 0.90 which corresponds to taking commercial keyframes with high confidence. Compared to the first method there is a 10% increase in recall for CNN. For ABC, the combined results are much higher than the results of either method.

As a second combination strategy, a high-level SVM is built which uses the detection results of two methods as inputs, namely the confidence values for second method and the length of the sequence that covers the detected commercials for the first method. As the results of Table 4 show this strategy has higher recall than the first strategy, but lower precision.

Table 1: Performance on test sets using Method 1. *keyframe* stands for the results when only candidate keyframes are taken as detected commercials. *sequence* stands for the final results when commercials are detected as repeating sequences.

		recall	precision	F1
CNN	keyframe	0.5620	0.6226	0.5907
	sequence	0.7445	0.8248	0.7826
ABC	keyframe	0.6274	0.7464	0.6817
	sequence	0.8406	0.9032	0.8708



CNN ABC

Figure 3: Confidence values for true commercials on training data for Method 2.

5. Summary and Discussion

In this study, two novel methods are proposed for detection and removal of commercials in broadcast news. The first method views commercials as sequences that repeat over time, and detects duplicate sequences. The second method builds an FLD classifier using distinctive color and audio features. Color and audio based methods have very high recall values, especially in CNN. Sequence based methods have lower recall values but higher precision which is a desirable property for keeping important news stories. In both methods, most of the false alarms correspond to logos or self advertisements which were not considered as commercials while truthing. We observe that combining different characteristics of commercials produces better results, although no one strategy is clearly superior. The results show that recall and precision up to 95% is possible with the proposed system.

Table 2: Performance on test sets using Method 2. Frames having higher values than the *threshold* values are labeled as commercials.

	threshold	recall	precision	F1
CNN	0.50	0.9294	0.9455	0.9374
	0.90	0.8273	0.9444	0.8820
ABC	0.50	0.7487	0.8060	0.7763
	0.83	0.6049	0.8682	0.7130

Table 3: Results of the first combination strategy.

		recall	precision	F1
CNN	common	0.6983	0.9829	0.8165
	in-sequence	0.7445	0.9474	0.8338
	final	0.8443	0.9507	0.8943
ABC	common	0.6395	0.9584	0.7671
	in-sequence	0.8232	0.9596	0.8862
	final	0.8873	0.9517	0.9184

Table 4: Results of second combination strategy.

	recall	precision	F1
CNN	0.9611	0.9186	0.9394
ABC	0.9099	0.8692	0.8891

Acknowledgments

This work was supported by the Advanced Research and Development Activity (ARDA) under contract numbers MDA908-00-C-0037 and MDA904-02-C-0451, and by the National Science Foundation (NSF) under Cooperative Agreement No. IIS-0121641.

References

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

[2] H. Schneiderman, T. Kanade, “Object Detection Using the Statistics of Parts”, International Journal of Computer Vision, 2002.

[3] R. Lienhart, C. Kuhmunch, W. Effelsberg, “On the detection and Recognition of Television Commercials”, In proceedings of IEEE International Conference on Multimedia Computing and Systems, 1997.

[4] 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

[5] Retrieval of commercials by video semantics C. Colombo, A. Del Bimbo, and P. Pala, IEEE International Conference on Computer Vision and Pattern Recognition, 1998.

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

[7] Evolvable visual commercial detector L. Agnihotri, N. Dimitrova, T. McGee, S. Jeannin, D. Schaffer, J. Nesvadba, IEEE International Conference on Computer Vision and Pattern Recognition, 2003.

[8] H. Wactlar, M. Christel, Y. Gong and A. Hauptmann, “Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library”, IEEE Computer, vol. 32, no. 2, pp. 66-73, February 1999.

[9] A. Hauptman et.al., “Informedia at TRECVID 2003 : Analyzing and Searching Broadcast News Video ”, TREC Video Retrieval Evaluation Publications, 2003.