Commercial Block Detection in Broadcast News Videos

Conference Paper · December 2014

DOI: 10.1145/2683483.2683546

CITATIONS
6

READS
153

4 authors, including:

Apoorv Vyas
École Polytechnique Fédérale de Lausanne
4 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Raghvendra Kannao
Indian Institute of Technology Guwahati
14 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Prithwijit Guha
Indian Institute of Technology Guwahati
61 PUBLICATIONS **196** CITATIONS
SEE PROFILE

Commercial Block Detection in Broadcast News Videos

Apoorv Vyas
Dept. of EEE
IIT Guwahati
Assam 781039, India
apoorv2904@gmail.com

Raghvendra Kannao
Dept. of EEE
IIT Guwahati
Assam 781039, India
raghvendra@iitg.ernet.in

Vineet Bhargava
Dept. of EEE
IIT Guwahati
Assam 781039, India
vineet.bhrgv@gmail.com

Prithwijit Guha
Dept. of EEE
IIT Guwahati
Assam 781039, India
pguha@iitg.ernet.in

ABSTRACT

Automatic identification and extraction of commercial blocks in telecast news videos find a lot of applications in the domain of broadcast monitoring. Existing works in this domain have used channel specific assumptions, machine learning techniques and frequentist approaches for detecting commercial video segments. We note that in the Indian context, several channel specific assumptions do not hold and often news and commercials have comparable frequencies of occurrence. This motivates us to use the machine learning techniques for classifying commercials in news videos. Our main contribution lies in the proposal of two features which are shown to outperform the existing audio-visual features – first, the MFCC bag of words (BoW) as audio track feature and second, overlaid text distribution as video shot feature. The shot feature space is further extended by appending contextual features which are categorized by SVM based classifiers. Additionally, we have used a post-processing stage to suppress the false positives. We have experimented with 54 hours of video acquired from three different Indian English based news channels and have obtained a F-measure of around 97%.

Keywords

Commercial Block Detection, MFCC Bag of Words, Overlaid Text Distribution, Contextual Features, SVM

1. INTRODUCTION

Television commercials are the most effective means of mass-marketing. The last decade has witnessed a steady growth in the number of television news channels leading to an explosion of multi-media data that can be acquired and analyzed for monitoring purposes. Product companies might be either interested in verifying the broadcast of their own commercial (as per contract) or on automatic notifications of their competitor's commercials and the shows they are sponsoring. Similarly, monitoring agencies might be interested in keeping a watch for controlling the time periods of commercial broadcasts. Groups focusing on analyzing or recording non-commercial video segments (e.g. news, discussions, interviews etc.), on the other hand, will be interested in detecting and removing commercial blocks from broadcast streams. All such varying necessities call for an efficient method for automatic detection of commercial blocks in broadcast news videos.

Existing literature in commercial block classification can be broadly divided into two categories – first, knowledge based methods and second, frequentist (repetition based) approaches. The knowledge based methods make assumptions on shot rate, logo presence, short silent regions, presence of blank frames etc. Sadlier et. al. [13] made an assumption on the presence of blank frames and depression in audio levels and have used the MPEG coded video features directly to separate commercial blocks from the TV pro grammes. Presence of specific product logos along with other features have been used to detect commercial boundaries [11, 18]. On the other hand, the absence of channel logo is also used to identify commercials [1]. Several researchers have used machine learning techniques by pre-training classifiers on audio-visual features extracted from commercials. These techniques include simple threshold based classification [4] (MPEG features), finite state machines [15, 17], SVM [7, 19, 6] and more recently an interactive ensemble learning method called Tri-Adaboost [9]. While knowledge based approaches work on case specific assumptions or multi-modal features, the frequentist approaches assume that the commercial segments repeat with higher frequency compared to the news, discussions and other non-commercials. These approaches work on the principle of finger-printing and hashing of audio-visual features [5, 3, 20, 16] and identify the commercials as the video segments having higher frequency of occurrence.

In this work, we focus on commercial block detection in Indian English news channel telecasts. In the Indian context, we have observed that the daily frequencies of broadcasting news segments are always very high and is often comparable to the frequency of specific commercials. Thus, a repetition based scheme is likely to fail in our case. Also, most Indian news channels do not strictly follow any format. Thus,

ICVGIP '14, December 14-18, 2014, Bangalore, India

http://dx.doi.org/10.1145/2683483.2683546 .

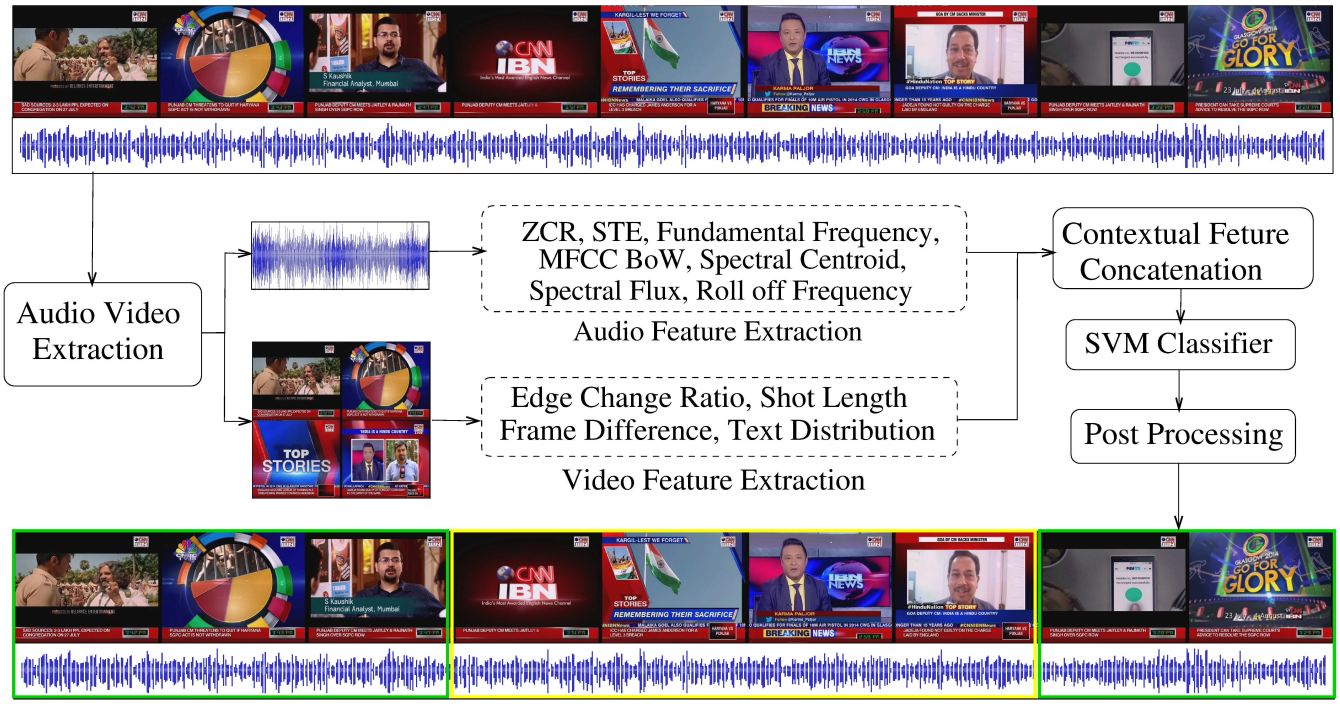


Figure 1: Illustrating the commercial detection scheme. As a pre-processing step the input video stream is segmented into shots based on color distribution uniformity. The audio tracks and image sequences are separated from each shot. The image sequences are used for extracting visual features while audio tracks are used to compute the audio features. Finally the combined audio-visual features (extended with context) are fed to a classifier, followed by a post-processing stage to identify the commercial blocks. The goal of this work is to identify the block of video segments containing commercials (marked in green) against other shows (news presentations, animations, discussions, weather reporting etc.) in broadcast news videos.

widely used assumptions like "presence of blank frames" and "absence of channel logo" can not be applied in the Indian context. This motivates us to employ the machine learning based approaches by training classifiers on audio-visual features of video segments.

In the proposed scheme, the video stream is first segmented into shots based on the uniformity of color distributions. The task of commercial detection then boils down to classifying these shots as either commercial or non-commercial segments. For this, we extract important audio and visual features from the shot and feed it to a classifier. The classification results are further improved by using a post-processing scheme which removes the false positives based on the temporal consistency of commercial and non-commercial blocks. The functional block diagram of the commercial block detection approach is illustrated in Figure 1(b).

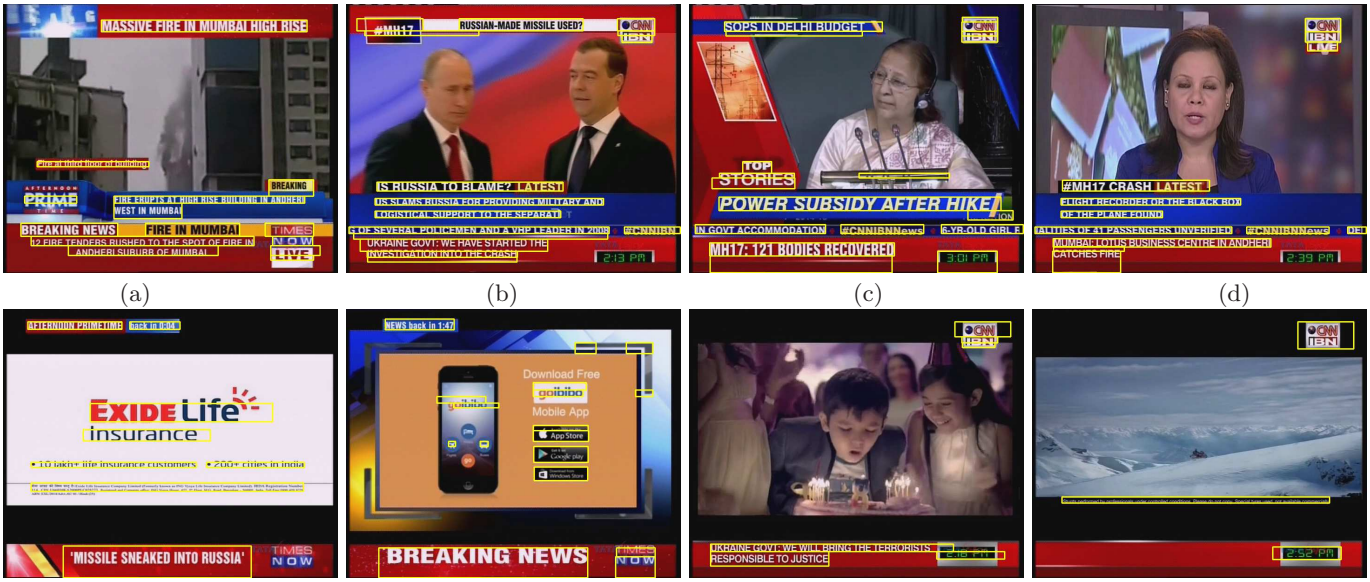
Researchers have proposed a number of features for the classification of commercial video segments – the most notable visual features being second order statistics of edge change ratio (ECR) and frame difference [7, 18], shot length [4] and presence/absence of text bands [8]; and the most commonly used audio features are zero crossing rate (ZCR) [17, 19], spectral centroid, flux and roll-off frequency [9] and short time energy (STE) [19]. Our main contributions lie in the proposal of one audio feature based on MFCC Bag of Words and one visual feature based on the distribution of text in horizontal bands in image frames.

The mel-frequency cepstral coefficients (MFCC) contain very important audio information and have been used extensively in speech processing community for speech recognition and speaker identification. In our work, the shots were of unequal length and hence the direct use of MFCC features was not possible. We have rather clustered all MFCC features obtained from fixed size windows and used them in the Bag-of-Words (BoW, henceforth) framework. We observed that the MFCC features in BoW framework work extremely well and alone gives a shot level accuracy in the range of 90% which was better than any other audio feature. Our next contribution is in the use of overlaid text distribution in identifying commercial shots. Earlier work have established that the presence/absence of overlaid ticker text bands in certain positions can differentiate commercial shots from the non-commercial ones [8]. However, these approaches used some manually marked positions for the ticker bands. We generalize this method by computing the amount of detected text in an array of horizontal bands. It was observed that the proposed feature provided the best accuracy among all the visual features.

This paper is organized in the following manner. Brief descriptions of the audio-video features along with their contextual extensions are presented in Section 2. A post-processing scheme for the removal of classification (SVMs trained over contextual features) errors is described in Section 3. The experimental results on 54 hours of video from three different Indian English based television news channels are presented in Section 4. Finally, we conclude in Section 5 and sketch the future scope of the proposed work.

2. CHARACTERIZING COMMERCIALS

The broadcast news videos are recorded as sequences of



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: Illustrating text distribution in news broadcast videos. The figures show the outputs of scene text region detection [14] in (a)-(d) non-commercial (news, discussions, field shots etc.) and (e)-(h) commercial videos. The non-commercial presentations in Indian television news channels contain a lot of ticker text bands; mostly concentrated in upper and/or lower $25-30\%$ of the scene area. However, during commercial breaks, most of these ticker bands disappear leaving only one or two bands at lower or upper portion of the screen; and, the product related text regions are distributed throughout the scene, mostly around the central region.

one hour long videos, each of which are shot segmented using inter-frame color distribution consistency. The advertisement and non-commercial video shots can be characterized based on a few general observations involving shot length, image motion and overlaid text statistics, music content and spectral properties of associated audio streams. These observations motivate us to compute 4 video (Sub-section 2.1) and 7 audio (Sub-section 2.2) features. More so, we observe that the commercials are generally presented in a block and hence in most cases the (near) left and right neighbors of a commercial shot are also possibly commercials. This motivates us to use the contextual features formed by concatenating the features of neighboring shots (Sub-section 2.3). The video features used for characterizing commercials are described next.

2.1 Video Features

We observe that the positions and contents of overlaid text bands play a significant role in discriminating commercial video segments from others. Apart from that, researchers have also used the shot length and second order statistics of edge change ratios (ECR) and frame differences (FD). However, the edge change ratios and frame differences are only to be computed in image regions excluding the overlaid text bands as these regions have their own motion patterns. Thus, text region detection is a necessary first step for us for the computation of video features. We have used the method proposed in [14] for the purpose of text detection. A few results of text region detection in advertisements and non-commercials are shown in Figure 2. It can be observed that the nature of text distributions are significantly different in commercials compared to other video segments. Brief descriptions of these video features are presented as follows.

Text Masked Edge Change Ratio (ECR) marks the dynamic nature of commercial background and estimates the net motion at object boundaries. Commercials have to convey maximal information in short duration and hence the motion content in commercials is generally higher compared to non-commercials [7]. We mask the detected text regions for ignoring the animation effects (rolling, moving in/out, blinking etc.) of text in the overlaid bands. The second order statistics (mean and standard deviation) of ECR computed over an entire shot (excluding the text regions) form a $2D$ feature for estimating motion content.

Text Masked Frame Difference (FD) estimates the amount of change between consecutive frames in a video shot. This is used along with ECR as a second feature for motion. ECR fails in cases, where object boundaries may not move but the internal region might change color [7]; e.g. a static red object can turn green. Such cases are very common in animated shots. The frame difference is used to handle such scenarios where the ECR may fail. The second order statistics of the frame differences computed over an entire shot (excluding the text regions) form the second set of $2D$ features for expressing the net amount of change.

Shot Length (SL) is defined as the length of a shot in frames and forms a $1D$ feature for the video segment. A general observation reveals that the commercial shots are of shorter duration compared to the others [7]. For example, in a 6 hour long video segment, we have obtained 4092 commercial shots and 1161 non-commercial shots even with the regulation that commercial blocks can not exceed 24 minutes in any one hour program.

Overlaid Text Distribution (OTD) plays a major role in characterizing commercials [8, 9]. It is observed that a lot of text bands are generally displayed during news presentations, interviews, debates and talk shows (Figure 2(a)–(d)) which are concentrated in the upper and lower parts of the

frame. On the other hand, in case of commercials, most of these bands disappear leaving only a single band, either at the top or the bottom of the image while product related information in different text bands are distributed through out the image (Figure 2(e)–(f)). We have divided the frame into 5 horizontal strips of equal height and estimate the second order statistics of the amount of text in each of these strips over the entire shot. The mean and standard deviation in each of the 5 strips form the $10D$ feature vector for overlaid text distribution in a video segment.

These 4 video features involving text and motion statistics only partially describe the visual mode of the commercials. The features employed for characterizing the audio modality of commercials are described next.

2.2 Audio Features

The audio streams associated with advertisement videos are observed to have high music content, faster change and higher volume compared to the non-commercials [7]. This motivates us to use a few existing audio features that capture these characteristics of audio streams [12]. We extract these features from audio streams associated with the video segments. The audio stream is first divided into non-overlapping frames of 20 msec duration and the features are computed from each frame. The statistics of these features computed from the entire stream are used as the audio descriptors for commercial classification. The different audio features used in this work are described next.

The **Zero Crossing Rate (ZCR)** measures how rapidly an audio signal changes and provides partial information about the music content of an audio stream [19, 17]. The **Short Time Energy (STE)** is defined as the sum of squares of signal values in a frame. It is noted that the commercials generally have higher audio amplitude for attracting user's attention and hence, have higher STE [19]. The spectral features are good descriptors for identifying audio streams with higher frequencies and faster change. For example, higher **Spectral Centroid (SC)** indicate the presence of higher frequencies (mostly in music), higher **Spectral Flux (SF)** signify rapid change of power spectrum and **Spectral Roll-Off Frequency (SRF)** discriminates music from (non)pure speech [9]. The **Fundamental Frequency (FF)** also plays an important role in discriminating advertisements (dominated by music) from non-commercials (dominated by speech) – i.e. audio streams from the former category are expected to have higher fundamental frequency compared to those from the later [12]. The second order statistics of each of these features computed over all the frames of a shot form the features for the associated audio stream.

Mel Frequency Cepstral Coefficients (MFCC) are one of the most popular short term features and are most commonly used in a variety of speech processing applications [12]. The MFCC features in a Bag of Words framework have been used earlier in the context of video concept detection [10]. We propose to use the same framework for the purpose of commercial detection. We have computed the first 15 Mel Frequency Cepstral Coefficients from all the frames extracted from the entire dataset. The resulting set of $15D$ vectors are subjected to k-means based vector quantization to form 100 cluster centers (words). The MFCC feature vectors computed from the set of frames of the audio stream associated with a shot are compared against these 100 words to form a histogram representing the contributions of each word. Thus, from the audio stream of each shot, we obtain a word histogram which is used further as feature for commercial classification.

2.3 Contextual Features

It is a well known fact that the commercials and news segments appear in continuous blocks of some minimum duration and thus the shots on the left and right neighborhood of a video segment can be used to strengthen our inference about the category of the current segment. Hence, to include the effect of the neighbors of a shot we append their features to extend the feature vector with the contextual features [7]. Thus, if a neighborhood of n_c shots is taken (both leftwards and rightwards), we effectively have $2n_c + 1$ contextual features for each dimensions of the audio-visual features. In all our experiments, we have used $n_c = 5$ following the proposal in [7].

SVM based classifiers are trained over the contextual feature space. However, even after extensive training, the classification errors do persist. We next describe the adopted (two-stage) post-processing scheme for improving the overall classification performance by suppressing false positives and false negatives.

3. POST PROCESSING

Television news broadcasts can be seen as sequences of Commercial and Non-commercial blocks consisting of several shots. Thus the block wise broadcast pattern of TV news broadcast can be used to reduce the false detections. Post processing is realized through two stages – first, at the *shot level* followed by *block level* post-processing.

Shot Level Post Processing – The shots on boundaries of advertisement and non-commercial blocks are usually misclassified due to the contextual features. In shot level post processing we try to locate the exact block boundaries by correcting the labels of the shots near the borders. We have adopted the shot level post processing scheme from [9]. We first identify the set $\mathbf{S_p}$ of shots for shot level post processing from set of shots $\mathbf{S} = \{s_1, s_2 \ldots s_i \ldots\}$. $\mathbf{S_p}$ is given as follows

$$\mathbf{S_p} = \{s_i : sgn(\sum_{t=1}^{N_l} \hat{y}_{i-t} l_{i-t}) \neq sgn(\sum_{t=1}^{N_r} \hat{y}_{i+t} l_{i+t})\} \quad (1)$$

where, N_l and N_r are the respective number of shots considered in the left and right neighborhood of s_i; $sgn(.)$ is the sign function; l_i is the likelihood of predicted label $\hat{y}_i$ for shot s_i. $l_i \in [0, 1]$ and $\hat{y}_i \in \{-1, 1\}$. Now every subset $\mathbf{S_c} = \{s_{i-u}, \ldots s_i \ldots s_{i+v}\}$ of consecutive shots in $\mathbf{S_p}$ may have advertisement and non-commercial block boundary. The block boundary $[b_t, b_{t+1}]$ in subset $\mathbf{S_c}$ is located by,

$$b_t = argmax_{t \in [i-u, i+v+1]} (|\hat{y}_t l_t - \hat{y}_{t+1} l_{t+1}|) \quad (2)$$

Once the block boundaries b_t are located in each subset we reestimate the label likelihood product for each shot in $\mathbf{S_c}$ as

$$\widetilde{(\hat{y}_j l_j)} = \begin{cases} \frac{1}{b_t - i + u + 1} \sum_{k=i-u}^{b_t} \hat{y}_k l_k; & j \in [i-u, b_t] \\ \frac{1}{i+v-b_t} \sum_{k=b_t+1}^{i+v} \hat{y}_k l_k; & j \in [b_t+1, i+v] \end{cases} \quad (3)$$

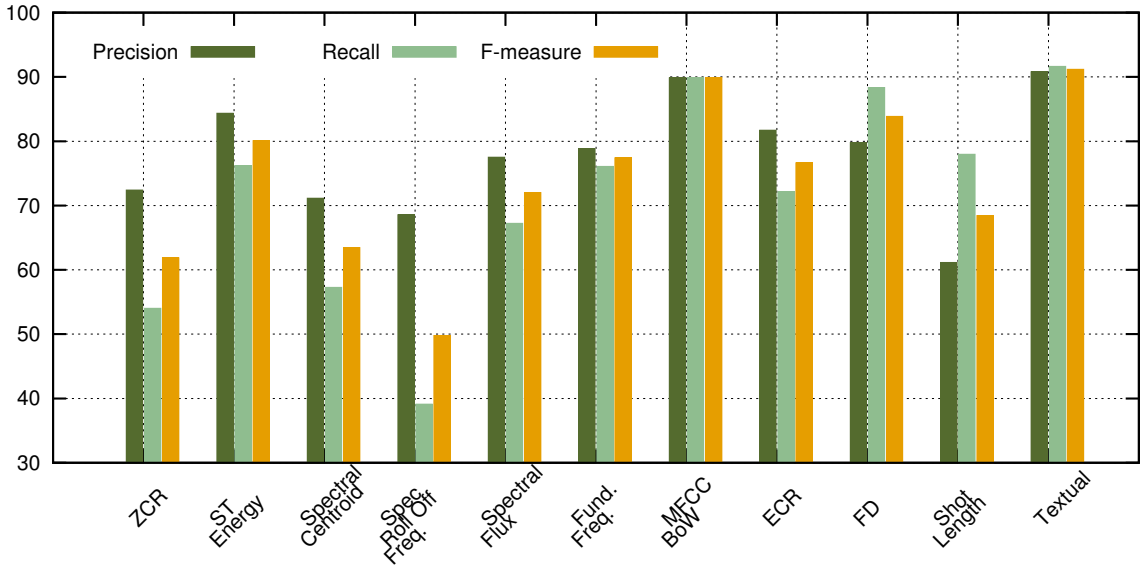


Figure 3: Performance analysis of the audio-visual features. SVM based classifiers are trained on each of the 11 features for the task of identifying commercial shots and the classification performances of each feature are shown in terms of precision, recall and f-measure. It can be noted that, the overlaid text region distribution and MFCC Bag of Words features outperform the other features in the respective video and audio domains.

Thus, the process of shot level post-processing helps us in locating the the block boundaries while refining the labels of the shots near the borders of advertisement and non-commercial blocks.

Block Level Post Processing – Here, we exploit the temporal consistency of the advertisement and non-commercial blocks duration constraint to further reduce the mis-classifications. We note that the duration of a non-commercial block can never be less than 60 seconds. So, we merge all the commercial blocks which are less than 60 seconds apart. Similarly, the duration of a commercial block is never less than 60 seconds. Thus, we also re-label the advertisement blocks (of size lesser than a minute) as non-commercials.

4. RESULTS

We have experimented with 54 hours of broadcast news videos, 18 hours from each of the three Indian English news channels *NDTV*, *TIMES NOW* and *CNN-IBN*. We have selected 6 hours of video from each channel to form the training data set with a total of 18 hours of video data for training. After shot segmentation, in each channel, we have obtained approximately 5000 shots from 6 hours of training data and around $10,000$ shots for 12 hours of testing data.

In our experiments we have observed that the number of commercial shots and non-commercial shots differ quite significantly. The number of commercial shots in the training set for each channel was found to be around 4 times the number of non-commercial shots. This is due to the fact that a commercial has a lot more motion compared to the non-commercial ones resulting in larger number of shorter duration shots for the former category. This difference in the number of shots of commercial and non-commercial categories in training sets gave rise to the problem of data imbalance. To tackle this problem, we first performed k-means clustering on the commercial and non-commercial features separately and collected training samples from different clusters in proportion to the cluster sizes. This scheme ensured that we don't pick all the training features from just a few clusters. From our experiments on validation sets we observed that this process of constructing training data set yields much better results compared to that of random sampling.

We have attempted the task of commercial shot classification by using the audio-visual features individually using SVM classifiers with linear kernels. The performances of individual features in terms of precision, recall and f-measure are shown in figure 3. From the figure it can be seen that the text distribution and the MFCC Bag of Words features have out-performed the other features reported in the literature. This performance analysis was carried out by using the contextual features.

The audio-visual features of a shot are concatenated along with their contextual extension and are used further to train SVM based classifiers for commercial identification. We have used the LibSVM toolkit for training the SVMs over the feature space [2]. Each of the three different channels have distinct styles of presentation. Thus, a classifier trained on a certain channel will provide best testing performance on that channel itself. This fact is also experimentally verified in our work.

We have trained 4 different classifiers; of which three are trained on the individual channels and a generic classifier trained over a dataset comprising of features from all the three channels. The classification performance of these four classifiers along with improvements after two stages of post-processing are presented in Table 1. We observe that a classifier trained on a particular channel exhibits best classification performance on that channel itself. As expected, the generic classifier has shown comparatively poorer performance on the combined data set or the individual channels. Also, the classification performances were found to consistently improve with the two stages of post-processing.

Table 1: Performance Analysis of generic classifier and classifiers trained on different channels using precision (P), recall (R) and f-measure (F). We note that classifiers trained over specific channels provide better performance on that channel only and a little degraded performance over others. This may be due to minor variations in presentation styles across channels. We also note that the different stages of post-processing help in the overall improvement of classification performance.

TRAINED ON	*Results*	Before Post Processing			With Post Processing I			With Post Processing II		
	Channel	P	R	F	P	R	F	P	R	F
NDTV	NDTV	**95.13**	**95.96**	**95.54**	**95.39**	**95.76**	**95.57**	**96.46**	**96.08**	**96.27**
	TIMES NOW	94.89	92.22	93.58	95.02	92.16	93.57	95.72	93.05	94.37
	CNN-IBN	93.96	87.42	90.57	93.97	87.21	90.46	96.85	90.07	93.33
CNN-IBN	NDTV	74.94	94.18	83.47	75.2	94.13	83.64	75.58	96.26	84.67
	TIMES NOW	92.35	90.91	91.62	92.68	90.7	91.68	94.32	92.69	93.5
	CNN-IBN	**94.46**	**96.78**	**96.11**	**95.49**	**96.78**	**96.13**	**96.67**	**97.99**	**97.33**
TIMES NOW	NDTV	83.51	96.27	89.44	83.1	96.42	89.26	84.6	97.22	90.47
	TIMES NOW	**96.64**	**95.73**	**96.18**	**96.48**	**95.88**	**96.18**	**97.81**	**96.74**	**97.27**
	CNN-IBN	92.98	90.35	91.65	92.9	90.5	91.69	95.65	95.11	95.38
GENERIC	NDTV	77.23	88.58	82.52	77.62	88.45	82.68	78.78	95.72	86.43
	TIMES NOW	82.8	85.8	84.28	82.68	85.59	84.11	84.73	94.9	89.53
	CNN-IBN	89.34	94.55	91.87	89.41	94.45	91.86	90.75	99.08	94.74
	Combined	**83.33**	**89.73**	**86.4**	**83.43**	**89.58**	**86.39**	**84.89**	**96.22**	**90.38**

5. CONCLUSION

We have proposed an approach to combine audio-visual cues to identify and localize commercial blocks in broadcast news videos. The main contributions of our proposal are the use of overlaid text distribution in horizontal bands as visual and MFCC Bag of Words as audio features. We have experimented with existing video features like second order ECR statistics, shot length and other audio features. The overlaid text distribution and MFCC Bag of Words features were found to outperform the ones proposed in earlier works. The experiments were conducted on 54 hours of video data acquired from 3 Indian English television news channels and have obtained classification performances (in terms of F-measure) around 97%. We have observed that classifiers trained on a specific channel perform better on that channel itself compared to others; which might be due to the differences in presentation styles of overlaid bands and audio settings.

Commercial block detection is a preliminary first step towards our larger goal of commercial analytics for the purposes of monitoring and competitive business intelligence. The present work has only explored the position of text regions in images for commercial detection but not the text content. We believe that text content and style will lead to more advanced features for commercial block detection, thereby identifying the products which are being advertised.

The classification of commercials involve the contributions of multiple features of different kinds. We have also shown the relative performances of each feature in classifying the commercials. In more recent experiments, we have observed that the classification success is a local phenomenon in the feature sub-spaces. In most cases, when some sub-set of features fail, another sub-set picks up and this feature combination is a very localized phenomenon. Along with identifying new features, the future work also lies in the experimentations with locally weighted ensemble classifiers and learned multiple kernel combinations for commercial classification.

6. REFERENCES

[1] N. Banic. Detection of commercials in video content based on logo presence without its prior knowledge. In *MIPRO, 2012 Proceedings of the 35th International Convention*, pages 1713–1718, 2012.

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[3] M. Covell, S. Baluja, and M. Fink. Advertisement detection and replacement using acoustic and visual repetition. In *8th IEEE Workshop on Multimedia Signal Processing*, pages 461–466, 2006.

[4] N. Dimitrova, S. Jeannin, J. Nesvadba, T. McGee, L. Agnihotri, and G. Mekenkamp. Real time commercial detection using mpeg features. In *International Conference on Information Processing and Management of Uncertainty in Knowlwdge-based Systems*, pages 481–486, 2002.

[5] P. Duygulu, M. yu Chen, and A. Hauptmann. Comparison and combination of two novel commercial detection methods. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 2, pages 1267–1270, June 2004.

[6] G. Fei and S. Ping. The detection of tv commercial based on multi-feature fusion. In *International Conference on Multimedia Technology*, pages 1–4, 2010.

[7] X.-S. Hua, L. Lu, and H.-J. Zhang. Robust learning-based tv commercial detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[8] A. Jindal, A. Tiwari, and H. Ghosh. Efficient and language independent news story segmentation for telecast news videos. In *IEEE International Symposium on Multimedia*, pages 458–463, 2011.

[9] N. Liu, Y. Zhao, Z. Zhu, and H. Lu. Exploiting visual-audio-textual characteristics for automatic tv commercial block detection and segmentation. *IEEE Transactions on Multimedia*, 13(5):961–973, 2011.

[10] M. Maehling, R. Ewerth, J. Zhou, and B. Freisleben. Multimodal video concept detection via bag of

auditory words and multiple kernel learning. In *Advances in Multimedia Modeling*, volume 7131, pages 40–50. Springer Berlin Heidelberg, 2012.

[11] L. Meng, Y. Cai, M. Wang, and Y. Li. Tv commercial detection based on shot change and text extraction. In *Second International Congress on Image and Signal Processing*, pages 1–5, 2009.

[12] P. Rao. Audio signal processing. volume 83 of *Studies in Computational Intelligence*, pages 169–189. Springer Berlin Heidelberg, 2008.

[13] Sadlier, A. David, M. O'Connor, S. O'Connor, E. Noel, and N.Murphy. Automatic tv advertisement detection from mpeg bitstream. In *International Workshop on Pattern Recognition in Information Systems (In Conjunction with ICEIS 2001)*, PRIS, pages 14–25, 2001.

[14] P. Shivakumara, T. Q. Phan, and C. Tan. A laplacian approach to multi-oriented text detection in video. *IEEE PAMI*, 33(2):412–419, February 2011.

[15] X. Wang and Z. Guo. A novel real-time commercial detection scheme. In *International Conference on Innovative Computing Information and Control*, pages 536–536, 2008.

[16] X. Wu and S. Satoh. Ultrahigh-speed tv commercial detection, extraction and matching. *IEEE Circuits and Systems for Video Technology*, 23(6):1054–1069, 2013.

[17] S.-H. Yang, C.-W. Fan, and Y.-C. Chen. An improved automatic commercial detection system. In *IEEE Visual Communications and Image Processing*, pages 1–4, 2011.

[18] B. Zhang, B. Feng, P. Ding, and B. Xu. Tv commercial detection using constrained viterbi algorithm based on time distribution. In *9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 2010–2014, 2012.

[19] L. Zhang, Z. Zhu, and Y. Zhao. Robust commercial detection system. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 587–590, 2007.

[20] D. Zhao, X. Wang, and Y. Qian. Fast commercial detection based on audio retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1185–1188, 2008.