A Novel Approach for Fast and Accurate Commercial Detection in H.264/AVC Bit Streams Based on Logo Identification

Klaus Schöffmann, Mathias Lux, and Laszlo Böszörmenyi

Institute of Information Technology (ITEC), University of Klagenfurt
Universitätsstr. 65-67, 9020 Klagenfurt, Austria
{ks,mlux,laszlo}@itec.uni-klu.ac.at

Abstract. Commercial blocks provide no extra value for video indexing, retrieval, archiving, or summarization of TV broadcasts. Therefore, automatic detection of commercial blocks is an important topic in the domain of multimedia information systems. We present a commercial detection approach which is based on logo detection performed in the compressed domain. The novelty of our approach is that by taking advantage of advanced features of the H.264/AVC coding, it is both significantly faster and more exact than existing approaches working directly on compressed data. Our approach enables removal of commercials in a fraction of real-time while achieving an average recall of 97.33% with an average precision of 99.31%. Moreover, due to its run-time performance, our approach can also be employed on low performance devices, for instance DVB recorders.

1 Introduction

Many free to air TV broadcast stations define advertisements as one of their main revenues. Therefore news, series, movies, etc. are interrupted by groups of advertisements. While commercial detection is certainly appealing for home users to skip unwanted content, also many professional applications in the domain of video indexing, retrieval, archiving, or summarization require removal of commercial breaks in order to focus on the actual content.

Although a lot of work has already been done in the area of commercial detection (see Section 2), most exact approaches work in the uncompressed domain (i.e. pixel data). Since practically every video is stored in compressed form, such approaches require to decode the video before commercial detection can be applied. However, as state-of-the-art encoding standards introduced increased complexity not only for encoding but also for decoding, working on pixel data has a serious drawback when run-time is an important criterion. Moreover, high resolution content as currently *HD* increases not only the requirement on decoding performance but also on required memory. In Digital Video Broadcasting (DVB) systems content is typically encoded with MPEG-2 or in case of HD with H.264/AVC[8]. Therefore, a commercial detection approach directly working on compressed data has a significant run-time advantage. However, existing approaches working in the compressed domain usually neither provide sufficiently exact results, nor do they support state-of-the art decoding standards (such as H.264/AVC). To the best of our knowledge, no approach has been presented until now, which can directly operate on H.264/AVC bit streams and achieves results comparable to techniques used in the uncompressed domain.

In this paper we present an approach, which is able to detect commercials in H.264/AVC bit streams in only a fraction of the time required for actual decoding. Our approach is based on the assumption that the broadcasting stations display a logo when sending real content while hiding the logo when sending commercials (or showing a different logo when sending self-advertisements[1]). Even though this assumption is not true for all broadcasting stations, it applies to many of them (it applies to the majority in Europe; in particular, to all popular German speaking channels). We emphasize that if our basic assumption is not fulfilled then - obviously - it cannot be applied. The novel idea of our approach is utilizing intra-prediction modes of macroblock partitions in order to detect whether a logo within a particular region is visible. This allows extremely fast processing since only a minimal part of decoding (namely the entropy decoding) has to be performed.

The paper is structured as follows. First we give a short overview on related work and state our research question. Then we describe our approach following by an evaluation on a recently recorded and annotated test data set. Finally we summarize the paper and present our conclusions.

2 Related Work

Research in the area of commercial detection can be classified in different ways. First, it can be based on visual or audio information or a combination of both. Some research projects (e.g. [6]) employ textual information acquired from text streams, optical character recognition (OCR) or speech recognition for classification. The authors of [3], for instance, employ common characteristics of advertisement blocks, like black frames and silence between commercials, to detect cuts and classify commercials based on OCR. Two different use cases can be identified for commercial detection: (i) recognize broadcasts of known commercials with high accuracy (e.g. with the help of fingerprinting or hashing) and (ii) detect previously unknown commercials. A further classification is to distinguish between approaches focusing on online (real-time) and offline processing. Some approaches need to analyze the video as a whole to determine thresholds and to compute features with temporal dependencies (e.g. windowed shot length average or shot boundary variance), while others focus on on-the-fly detection of commercials. There are also differences between commercial detection in compressed and uncompressed domain. While detection in compressed domain is faster, in general it is also a more challenging task and recognition rates are lower than in uncompressed domain. Some research groups focus on specific do-

[1] like program preview and broadcaster's merchandise

mains like for instance in [10], where commercial detection is applied to news broadcasts and anchorman detection is employed as a domain specific feature.

In [11] important features for detection of commercial broadcasts in Germany in the uncompressed domain are described. The authors focus on groups of monochrome or black frames in between commercials and an increased number of hard cuts in commercial blocks. Increased visual activity in commercials is reflected by the features *edge change ratio* and the *motion vector length.* They further increase the accuracy of their approach based on rules for maximum commercial length and other heuristics derived from German laws. Their evaluation based on German broadcasts showed that the approach resulted in a detection rate of 96.14% of the commercial block frames and 0.09% misclassifications.

In [5] classification of MPEG-2 video segments (based on a fixed number of I-frames) in compressed domain is presented. The authors employ logo recognition, black frame detection and color variance between I frames. They achieved a detection rate of 93% in terms of number of advertisements roughly in real-time (e.g. 1 min of processing for 1 minute of video content).

In [13] a system for real-time recognition of commercials within the first half second of broadcast in uncompressed domain based on color features is presented. The authors evaluate different hash functions to compute similarity of arbitrary video frames to already known commercials and achieve a recall of 96% with precision of 100%.

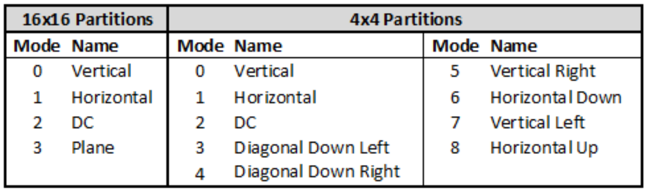
In [1] discriminative features for commercial detection that can be extracted within an MPEG encoding process in real-time are investigated as means for commercial detection by applying genetic algorithms. The authors focus on the selection of a set of features optimal for the task and chose *key frame distance*, *luminance* and *letterboxing.* Their best result is recall of 82% with precision of 97%. They furthermore present an approach for determining necessary thresholds for commercial detection by employing genetic algorithms. This approach outperforms thresholds set manually by experts in precision and recall as well as time needed to find the threshold. From the same research group another approach in compressed domain based on monochrome frames in between commercials is presented in [2]. While the approach allows real-time processing, classification of a commercial is done at the end of the commercials, which means that a commercial can only be identified after its last frame. The authors also use rather restrictive heuristics like a maximum commercial block length based on their test data set.

Evaluation results indicate in general precision levels beyond 90% and recall levels of 80% and more. Results of different research groups, however, cannot be compared directly as they use different test data sets from different broadcasting stations, different times and different regions. Evaluations also differ regarding the boundaries of the shots. Some groups use the second nearest to the recognized start and end frame of a commercial block for evaluation, others the exact frame number for determining what portion of the video has been classified correctly. Others use the number of correctly classified commercials for determining accuracy. Furthermore approaches are optimized in a different way. While for some application high recall (not missing a single commercial) is important other require a high precision (no real content frames get cut out).

3 Commercial Detection in H.264/AVC

In Europe, a trend towards digital TV, especially DVB-S and DVB-T can be observed. Also consumer products capable of receiving and displaying high definition video content are getting more and more common. As already mentioned, high definition content broadcasted in DVB is encoded in the H.264/AVC format. Compared to MPEG-2, H.264/AVC offers in general better visual quality at lower bitrates at the cost of higher computational effort for both the encoder and the decoder [12]. Concepts employed for encoding H.264/AVC are more sophisticated than the ones used in MPEG-2. For commercial detection in compressed domain we interpret particular compression concepts as features for logo detection. The idea of taking advantage of features already extracted by the encoder is not new; it has been discussed in the context of MPEG-2 DC coefficients and motion vectors (see e.g. [1], [2] or [5]). In the following we describe the selection of the appropriate H.264/AVC features and a distance function.

In contrast to earlier standards, H.264/AVC[8] allows partitioning of macroblocks to partitions of the size 16x16, 16x8, 8x16, 8x8, 8x4, 4x8, and 4x4 pixels. The possible partitions of intra coded macroblocks are limited to 16x16 and 4x4 pixels. Intra prediction is used for every partition of an intra coded macroblock. Four possible intra prediction modes for 16x16 partitions and nine possible intra prediction modes for 4x4 partitions are defined by the standard, as shown in Table 1 and Fig. 1.



16x16 Partitions		4x4 Partitions			
Mode	Name	Mode	Name	Mode	Name
0	Vertical	0	Vertical	5	Vertical Right
1	Horizontal	1	Horizontal	6	Horizontal Down
2	DC	2	DC	7	Vertical Left
3	Plane	3	Diagonal Down Left	8	Horizontal Up
		4	Diagonal Down Right		

Table 1. Intra prediction modes in H.264/AVC

As mentioned earlier, our commercial detection approach is based on the observation that many broadcasters hide their logo while broadcasting commercials. For the suggested algorithm we assume that commercial content has no logo, while regular program content shows a logo of the broadcaster on a fixed position. In our experience the location of a logo and its (structural) design changes rarely. So we further assume that position and size of the logo are known for each broadcaster. For a particular use case this constraint can be tackled by providing this information to be retrieved from a server, which is updated every time the position or structure of the logo changes for a particular broadcaster.

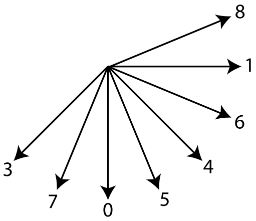


Fig. 1. Intra prediction modes of 4x4 partitions

Another approach is to let the user paint a boundary box around the logo. Obviously, if the position of the logo is not fixed (e.g. move around or rotate), the performance (i.e. recall) of our algorithm will degrade.

For detection of frames containing no broadcaster logo we employ intra prediction modes used for intra coded macroblocks in the area of the logo. We extract the intra prediction modes of the macroblocks - the so called *Intra Prediction Layout* (IPL) - within the logo region. According to the prediction modes introduced in Fig. 1, Fig. 2 shows a visualization of how the nine 4x4 intra prediction modes are applied to a particular logo. Note that it is not necessarily required to use all 4x4 partitions the logo is contained in. In some cases it might be sufficient to use only some relevant partitions within the logo area.

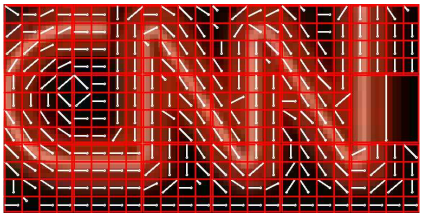


Fig. 2. Prediction modes (white) of intra-coded partitions (red) in the area of the CNN logo

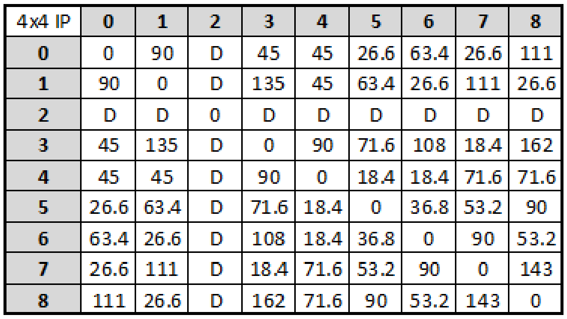
We further focus on intra coded frames (I frames) only and leave aside predicted (P) and bidirectional predicted (B) frames. Motivated by a high definition DVB use case, where HD content is broadcasted according to the DVB standard [4], we can assume that 2 seconds is the maximum distance between two consecutive I frames. In other words for a 25 fps video every 50-th frame is an I frame. Restricting the approach to I frames assures that macroblocks in the area of the logo use intra coding and we can obtain an IPL for each frame investigated. As the H.264/AVC standard [8] allows an intra coded macroblock to be encoded either as one 16x16 partition or as 16 4x4 partitions, we transform intra prediction modes of 16x16 partitions to 4x4 intra prediction in order to ease the comparison of the respective IPLs (and simply use 4x4 partitions as the basis).

The transformation is quite simple as three out of four intra prediction modes for 16x16 partitions (namely Vertical, Horizontal, and DC) are also used for 4x4 partitions. There is one intra prediction mode (Plane) which is not used for 4x4 partitions. In our transformation rules we use the 4x4 DC mode for that one as it seems to be the most similar mode. Therefore, our transformation rules are:

- 16x16 Vertical → 4x4 Vertical
- 16x16 Horizontal → 4x4 Horizontal
- 16x16 DC → 4x4 DC
- 16x16 Plane → 4x4 DC

On every I frame we extract the IPL from the area of the broadcaster's logo, and compare it with the a-priori known prototype IPL [2] for the selected broadcaster. For comparison we use a distance function based on angle differences between each possible pair of 4x4 intra prediction modes given by a matrix $M = (m_{i,j})$ with $i, j \in \{0, 1, \ldots 8\}$ shown in Fig. 3. In M, every column and every row represents one 4x4 intra prediction mode, as defined in Table 1 / Fig. 1, and D represents the difference from or to 4x4 DC intra prediction mode. For the IPL of the n-th I frame l_n, we calculate the normalized distance to the prototype IPL l_S denoted as $d(l_n, l_S)$. Both IPLs l_n and l_S have by definition the same number of 4x4 partitions k. The prediction mode of the i-th partition, $i \leq k$, of an IPL is l_n^i for the n-th frame and l_S^i for the prototype IPL respectively.

$$d(l_n, l_S) = \frac{\sum_{i=0}^{k} m_{l_n^i, l_S^i}}{k} \tag{1}$$



4x4 IP	0	1	2	3	4	5	6	7	8
0	0	90	D	45	45	26.6	63.4	26.6	111
1	90	0	D	135	45	63.4	26.6	111	26.6
2	D	D	0	D	D	D	D	D	D
3	45	135	D	0	90	71.6	108	18.4	162
4	45	45	D	90	0	18.4	18.4	71.6	71.6
5	26.6	63.4	D	71.6	18.4	0	36.8	53.2	90
6	63.4	26.6	D	108	18.4	36.8	0	90	53.2
7	26.6	111	D	18.4	71.6	53.2	90	0	143
8	111	26.6	D	162	71.6	90	53.2	143	0

Fig. 3. Distance matrix M for comparing IPL

To tackle the problem of single I frames containing visual content in the same color as the logo in the logo region we compute the average $d_n^{avg}(l_n, l_S)$ over r I frames. This process flattens single peaks but retains sequences of high values that occur in commercial blocks.

[2] This a-priori known prototype IPL could also be stored on a server and be updated when the structure of a logo changes.

$$d_n^{avg}(l_n, l_S) = \frac{\sum_{i=-\lfloor \frac{r}{2} \rfloor}^{\lfloor \frac{r}{2} \rfloor} d(l_{n+i}, l_S)}{r} \tag{2}$$

Based on the averaged distance for each I frame n we classify all I frames with $d_n^{avg}(l_n, l_S) > t$, where t is a predefined threshold, as I frames containing commercial content. We further employ a heuristics to extend the quality of classification. Therefore, we only assume sequences of 15 and more I frames as commercials. This is implicitly based on the heuristics that a commercial block has a minimum length of 30 seconds.

4 Evaluation

We have investigated 19 different channels [3] popular in German speaking countries. All of them suppress their logo while broadcasting commercials. Our evaluation is based on roughly 10 hours of DVB-S recordings from nine of those channels with different genres (see Table 2). The proportion of commercial material in the recorded content of 10 hours (i.e. 594 minutes) is 20.98% (i.e. 124.66 minutes) in average.

We manually classified every second of content into *Commercial* or *Real Content* and used that classification as ground truth. Self advertisement (e.g. commercials intros, teasers, and program preview) was also classified as *Commercial*. We run our commercial detection approach on the test data and evaluated *Recall* and *Precision*. As test setting we used $D = max(m_{i,j})$, $r = 5$ and $t = 0.28$. All videos have been encoded with the x264 [9] encoder with a bit rate of 4096 Kb/s. As the DVB specification [4] suggests using 2 seconds as a maximum time interval between two random access points in the bit stream, we have encoded our 25fps videos with a *Group-of-Picture* (GOP) size of 50 frames. Thus, the evaluation of our test data reflects the performance of our approach if used in an on-the-fly manner for content received via DVB.

Our approach achieves an average recall of 97.33% with an average precision of 99.31% in total. As shown in Table 2, for almost all recordings we achieved results near to 100% for both precision and recall. However, for one channel in our evaluation recall decreased to about 84%. The reason is that this channel uses the same logo while showing self advertisement as while showing real content. In that case self advertisement is not detected as ”commercial” (which was our requirement) and, thus, recall decreases.

Regarding run-time, our approach performs very well due to several reasons. First, as only I Frames are considered, the number of frames to process is strongly reduced. Moreover, as our approach solely requires information already available after entropy decoding (no DCT/Integer transform, no motion compensation, no pixel interpolation, no deblocking), the run-time is further reduced to roughly

[3] 3sat, ARD, ATV, BR, CNN, Das Vierte, DSF, Eurosport, Kabel-1, N24, n-tv, ORF 1, ORF 2, Pro 7, RTL, RTL II, Sat 1, Super RTL, VOX, WDR, ZDF

Genre	**Channel**	**Minutes**	**Commercial proportion**	**Recall**	**Precision**
Documentary	N24	53	21%	97.67%	100.00%
Documentary	ZDF	30	9%	99.39%	99.51%
Feature film	SuperRTL	110	18%	99.76%	99.56%
Feature film	Pro7	39	20%	99.23%	96.23%
News show	CNN Int.	58	19%	99.00%	100.00%
Sports	DSF	106	24%	98.82%	99.65%
Thriller	SAT 1	25	28%	100.00%	99.39%
Reality show	ATV	98	25%	98.10%	99.47%
Live show	RTL	75	20%	84.03%	100.00%

Table 2. Evaluation results

one quarter (refer to [7] for a workload characterization of H.264/AVC). In addition, if the logo is used in the upper part of the image, we can simply skip entropy-decoding for all macroblocks after the logo. If the logo is in the bottom part of the image, preceding macroblocks need to be entropy-decoded due to the variable length of macroblocks (i.e. encoded with CAVLC or CABAC). Depending on the position of the broadcaster's logo, commercial detection required between 4.46 % and 6.97% of the full decoding time for our test data. In other words, if a decoder would be able to decode a 60 minutes sequence in 50 minutes (i.e. in real-time), a commercial detection tool using our approach would require less than 3.5 minutes to detect all commercials.

5 Summary & Conclusion

We have presented a method for commercial detection in compressed domain utilizing encoding concepts of H.264/AVC. Experiments show that the method has very high precision and reasonable recall: While we minimize the number of false positives we miss in certain cases intros, teasers, and self advertisement of broadcasters. Beside the high precision and reasonable recall the contribution of this approach lies in the run-time performance and efficiency. Less than 7% of the decoding process has to be done to reach the accuracy documented in this paper. Therefore the proposed method can also be applied to HD content, where real-time or even faster analysis in the uncompressed domain is still a challenging problem. Also the implementation is rather easy to reproduce. In comparison to other approaches we use a minimum number of heuristics. We assume that the size of a commercial block is greater than 30 seconds and that the broadcasting logo is suppressed as long as commercials are broadcasted. Although not implemented in our current version, the accuracy of our approach could be improved if P or B frames containing intra coded macroblocks in the area of the logo are considered too. However for utilizing P and B frames we need higher decompressing effort.

References

1. L. Agnihotri, N. Dimitrova, T. Mcgee, S. Jeannin, D. Schaffer, and J. Nesvadba. Envolvable visual commercial detector. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, volume 02, page 79, Los Alamitos, CA, USA, 2003. IEEE Computer Society.
2. N. Dimitrova, S. Jeannin, J. Nesvadba, T. McGee, L. Agnihotri, and G. Mekenkamp. Real-time commercial detection using mpeg features. In *9th Int. Conf. On Information Processing and Management of Uncertainty in knowledge-based systems (IPMU 2002)*, Annecy, France, 2002. Invited paper.
3. L.-Y. Duan, J. Wang, Y. Zheng, J. S. Jin, H. Lu, and C. Xu. Segmentation, categorization, and identification of commercial clips from tv streams using multimodal analysis. In *Proceedings of the 14th annual ACM international conference on Multimedia (ACM Multimedia 06)*, pages 201–210, New York, NY, USA, 2006. ACM.
4. dvb.org (ETSI). ETSI TS 101 154 v1.7.1, Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream, February 2007.
5. R. Glasberg, C. Tas, and T. Sikora. Recognizing commercials in real-time using three visual descriptors and a decision-tree. In *IEEE International Conference on Multimedia and Expo 2006 (ICME 2006)*, Toronto, July 2006. IEEE.
6. A. G. Hauptmann and M. J. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *ADL '98: Proceedings of the Advances in Digital Libraries Conference*, page 168, Washington, DC, USA, 1998. IEEE Computer Society.
7. M. Holliman and Y. Chen. MPEG Decoding Workload Characterization. *Proc. of Workshop on Computer Architecture Evaluation using Commercial Workloads*, pages 23–34, 2003.
8. ISO/IEC JTC 1/SC 29/WG 11. ISO/IEC FDIS 14496-10: Information Technology - Coding of audio-visual objects - Part 10: Advanced Video Coding. March 2003.
9. Laurent Aimar and Loren Merritt and Eric Petit and Min Chen and Justin Clay and Mans Rullgard and Radek Czyz and Christian Heine and Alex Izvorski and Alex Wright. x264 - a free h264/avc encoder.
10. S. Li, H. Li, and Z. Wang. A novel approach to commercial detection in news video. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing 2007 (SNPD 2007)*, volume 2, pages 86–90. IEEE, August 2007.
11. R. Lienhart, C. Kuhmunch, and W. Effelsberg. On the detection and recognition of television commercials. In *International Conference on Multimedia Computing and Systems (ICMCS'97)*, volume 00, page 509, Los Alamitos, CA, USA, 1997. IEEE Computer Society.
12. J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi. Video coding with H. 264/AVC: tools, performance, and complexity. *Circuits and Systems Magazine, IEEE*, 4(1):7–28, 2004.
13. A. Shivadas and J. Gauch. Real-time commercial recognition using color moments and hashing. In *Fourth Canadian Conference on Computer and Robot Vision (CRV '07)*, volume 00, pages 465–472, Los Alamitos, CA, USA, 2007. IEEE Computer Society.