Automatic TV Advertisement Detection from MPEG Bitstream

David A. Sadlier*, Dr. Sean Marlow, Dr Noel O'Connor and Dr Noel Murphy
Centre for Digital Video Processing
Dublin City University
Rep. of Ireland

Abstract

The Centre for Digital Video Processing at Dublin City University conducts concentrated research and development in the area of digital video management. The current stage of development is demonstrated on our Web-based digital video system called *Físchlár* [1], which provides for efficient recording, analysing, browsing and viewing of digitally captured television programmes.

Advertisement breaks during or between television programmes are typically recognised by a series of 'black' video frames simultaneously accompanying a depression in audio volume which separate each advertisement from one another by recurrently occurring before and after each individual advertisement. It is the regular prevalence of these flags that enables automatic differentiation between what is programme and what is a commercial break. This paper reports on the progress made in the development of this idea into an advertisement detector system that automatically detects the commercial breaks from the bitstream of digitally captured television broadcasts.

Keywords: MPEG-1; Black/Silent Frames; DC coefficients; Subband scalefactors

*Corresponding Author
Tel.: +353-1-7005871
Fax.: +353-1-7045508
E-mail address: sadlierd@eeng.dcu.ie

1. Introduction

For the development of an efficient video browsing/viewing tool for digitised television, it is desirable to present the user with the option of skipping irrelevant content. A typical television programme may be accompanied by beginning/end credits with one or more ad-breaks somewhere in the middle. To the user, these features of a programme/video are generally regarded as an insignificant part of the recorded material. Hence, by detecting the ad-breaks, the efficiency of programme browsing/viewing may be increased.

Advertisement breaks may be isolated from actual programme material by the flags that most terrestrial and some satellite television companies wave during their broadcast: a series of 'black' video frames simultaneously accompanied by a decrease in the audio signal occurring before and after each individual advertisement [2].

The *Físchlár* system captures television broadcasts and encodes the programmes according to the MPEG-1 digital video standard with the audio signal coded in line with the Layer-II profile.

An inherently dark or 'black' frame of a video may be recognised by its luminance histogram, which would be typically characterised by having most of its 'power' at the bottom end of the pixel amplitude spectrum, corresponding to black/dark pixels.

Thus, by comparing an average pixel value, representing an entire frame, against some given threshold, a decision on whether that frame may be considered 'black' or not, may be made.

Furthermore, a depression in audio volume for a particular video frame may be recognised as follows:

A summation of the absolute value of all the audio samples corresponding to one video frame may be defined as the ‘audio level’ for that frame, i.e. for a relatively silent frame, a low audio level would be expected. Thus, by comparing this audio level to some threshold, an audio depression (of magnitude defined by threshold) may be detected.

The abovementioned criteria propose a simplistic approach to the task of locating within a television programme, groups of black-frames/audio-depressions, which should provide for efficient detection of advertisement breaks.

However, the method does require direct access to both video pixels and audio samples. Therefore it necessitates a full decode of the captured programme from its compressed format [MPEG-1 (Layer-II)], which is highly undesirable from a computational point of view [2].

It was proposed that the same assessment and classification of the individual frames of a captured television signal might be more efficiently made as follows:

- For video; an examination of the DC Discrete Cosine Transform (DC-DCT) coefficients of a frame, which represent the weight of its zero-frequency content, with a view to establishing whether or not the frame is inherently dark enough to be labeled ‘black’. [The Discrete Cosine Transform (DCT) forms an integral part of the compression process used in the MPEG-1 standard].

- For Audio; an inspection of the weight of the scalefactors of the signal’s (low) frequency subbands with a view to establishing whether or not a video frame’s accompanying audio signal power is minimal enough for it to be labeled ‘silent’. [In the MPEG-1 Layer-II standard for compression of audio signals, the signal frequency spectrum is divided uniformly into 32 subbands which are then coded independently from one another.]

It was envisaged that black-frame/silence detection via the abovementioned principles would provide for advertisement detection from captured and encoded television programmes to the same degree of accuracy as could be obtained by the methods requiring implementation of a full audio/visual decode, but with the computational burden significantly reduced.

2. Background

The Digital Video Indexing Project at Dublin City University is a continuing research endeavor aimed at developing innovative technologies fundamental to the realisation of efficient video content management.

To demonstrate our research on digital video, we have developed a Web-based digital video system which we call *Físchlár* [from the Irish *fís* (vision) and *chlár* (programme)]. At present a user can pre-set the recording of TV broadcast programmes and can choose from a set of different browser interfaces which allow navigation through the recorded programmes. As our research develops we will plug in increased options such as personalisation and programme recommendation, automatic recording, SMS/WAP/PDA alerting, searching, summarising and so on.

To initiate the recording of a programme, a user browses the TV schedule and selects those programmes to be recorded - our system will then automatically record (digitally) that programme at broadcast time, much the same as a home Video Cassette Recorder. After we record a programme, we then automatically segment it using our shot boundary detection technique based on colour histogram comparison, so that the content becomes easily browsable through our various user interfaces. The analysed programme is then added to our archive of recorded programmes which a user can scroll through and then select one for browse/playback. As a user browses through a programme he/she can then stream the video to their desktop.

[SMS: Short Messaging Service; WAP: Wireless Application Protocol; PDA: Personal Digital Assistant]

3. 'Black/Silent' Frames

3.1 The MPEG Standard

The Moving Pictures Experts Group (MPEG), who meet under the International Standards Organisation (ISO), generate the international standards for digital video and audio compression. MPEG-1 [3] is a standard in five parts which individually address the issues of audio/visual multiplexing, video coding, audio coding, bitstream testing, and software implementation. A detailed description of the MPEG-1 standard for audio and video compression may be found via references [3] & [4] and is thus not dealt with here.

3.2 Black Frame Detection

Físchlár captures analog television signals and encodes them according to the MPEG-1 standard. An MPEG-1 video frame is divided into slices, which are subdivided into macroblocks which each contain 6 blocks of (8x8) pixels transformed by a 2D-DCT, 4 of which provide luminance information, leaving 2 for chrominance information. The video data is either explicitly given for the frame (I-frame) or provided implicitly via forward prediction (P- frame) or forward/backward prediction (B-frame).

The four luminance blocks (Y-blocks) of each macroblock provide the essential information on how dark or 'black' the macroblock effectively is, hence indicating how dark or 'black' the overall frame may be.

Each Y-block consists of a DC coefficient, which represents its mean luminance intensity, and a number of AC coefficients, which represent its non-zero frequency content. The DC value corresponds to an average intensity value for each block, to which the visual perception is highly sensitive, whereas, various frequency fluctuations may go unnoticed. It was thus assumed that a decision on the inherent darkness of a block could be made, with acceptable accuracy, via examination of the DC coefficients exclusively, i.e. AC variations may be ignored.

The proposal was that an average luminance intensity value for each frame could be determined from the DC-DCT coefficients provided within individual Y-block. This value was expected to be relatively low for inherently dark frames and higher for brighter frames. Thus by thresholding this value, a decision on whether the frame is 'black' or not may be made.

3.3 Silent Frame Detection

Físchlár captures television audio signals and encodes them according to the MPEG-1 Layer-II compression algorithm which encodes as follows:

The frequency spectrum of the audio signal (sampled at 32, 44.1, or 48kHz) is first divided uniformly into 32 subbands which approximate the ear's critical bands. These subbands are then individually assigned a bit-allocation according to the audibility of quantisation noise within that band (a pyschoacoustic model of the ear analyses the audio signal and provides this information to the quantiser).

Layer-II frames consist of 1152 samples; 3 groups of 12 samples from each of 32 subbands. A group of 12 samples gets a bit allocation and, if this is non-zero, a scalefactor. Scalefactors are weights that scale groups of 12 samples such that they fully use the range of the quantiser (the encoder uses a different scalefactor for each of the three groups of 12 samples within each subband only if necessary). The scalefactor for such a group is determined by the next largest value (given in a look-up table) to the maximum of the absolute values of the 12 samples, thus it provides an indication of the maximum power exhibited by any one of the 12 samples within the group.

The proposal was that an audio power level for each video frame could be determined by superposition of the scalefactors corresponding to the groups of audio samples to which the frame is associated. This power level was expected to be relatively low for 'silent' frames and higher for louder frames. Thus by thresholding this value, a decision on whether the frame is 'silent' or not may be made.

Further still, it was expected that examination of just the bottom end of the frequency spectrum would provide sufficient information on which this decision could still be accurately made, since it is typical of an audio signal to have most of its energy corresponding to relatively low frequencies.

By trial and error examination of various audio signals, it was decided that at least 10 of the low frequency subbands' scalefactors must be included in the silence investigations such that the results rendered were both sensible and consistent.

[Since the maximum frequency encoded by the MPEG-1 Layer-II standard is 20kHz, the first 10 subbands corresponds to an examination of the energy present from
0-6kHz].

4. Ad-break Detection

4.1 Recognition of Advertisement breaks

As explained, the occurrence of black-frame/audio-depression series may indicate the existence of an ad-break. However, it is possible, and maybe quite probable, that these indicators also occur during the valuable material of the programme itself. For example, they are not uncommon when news programmes cut back and forth from anchorperson to news reports, or during scene changes during a soap opera. To combat this problem, and its consequence of detection/removal of valuable programme content, some strict conditions had to be enforced.

- It was noted that the 'black/silent' frame series occurring between individual advertisements tended to be of at least 6 frames in length. Thus to aid against detection of freak black-frame/audio-depression occurrences not associated to ad-breaks which may sporadically occur during programmes, it was decided only to recognise them if they exhibit a series of **<u>at least 6 consecutive ‘black/quiet’ frames</u>**. i.e. a series of 10 ‘black/quiet’ frames between adjacent advertisements would be detected, while a series of 5 occurring between anchor person and news report would not.

- Upon examination of 20+ advertisement breaks from various television stations, the longest advertisement recorded was that of 76secs, with approximate average advertisement duration of 25secs. Thus it was decided that upon detection of a series of (at least 6) ‘black/silent’ frames, if another distinct series was not detected within a **<u>window of 90 seconds</u>**, then the initial series must therefore not correspond to an ad-break and should be ignored. This prevented against recognition of rogue ‘black/quiet’ frame series which may randomly occur during relevant programme material.

  A consequence of this condition was that the system would fail upon occurrence of advertisements that were longer than 90secs in duration (76secs was longest advertisement recorded so 90secs provided for some tolerance).

  Frame rate for *Físchlár* is 25 frames/sec. Thus 90secs corresponds to 2250 frames.

- Upon examination of 20+ advertisement breaks from various television stations, it was determined that the average number of individual advertisements within an ad-break was approximately seven.

  It was decided, to further prevent against the possibility of relevant programme material being mistakenly recognised as advertisement, that the **<u>recognition process</u>**

would not succeed if the number of advertisements within one ad-break was less than three.

i.e. upon detection of a series of (at least 6) ‘black/quiet’ frames, at least three more series must be detected, within the respective 90sec windows of each other, for the overall detection to be recognised as an ad-break (detection of 4 ‘black/quiet’ frame series corresponds to 3 individual advertisements). This prevented against possible mistaken recognition due to the unlikely event of up to three rogue series (consisting of at least 6 consecutive ‘black/quiet’ frames) occurring within 90secs of each other, during relevant programme material.

The above three conditions would be individually weak since they focus on features which are undoubtedly inconsistent attributes of a television broadcast. However it is the combined effect of all three clauses which is expected to provide the success in accurately preventing mis-recognition of programme content for advertisement material. Figure-1 explains how detection of sporadic ‘black/quiet’ frame series are interpreted and recognised (shaded frames represent frames which are both ‘black’ and ‘silent’).

4.2 Black-frame & Silent-frame Thresholds

A number of threshold techniques were investigated. However, the following adaptive method provided the most consistency in the results obtained, and was thus chosen as the appropriate scheme.

For video, an overall mean DC-DCT value was calculated by averaging over all individual frames (= DC-DCT$_{avg}$). The ‘black-frame’ threshold was then expressed as some factor times this number. By trial and error examination of various ad-break clips, the video threshold which gave the most sensible and consistent results was:

$$\mathbf{V_{th} = 0.48 * DC\text{-}DCT_{avg}} \qquad \mathbf{(1)}$$

For audio, an overall mean audio level value was calculated by averaging over all individual frames (= audio_level_{avg}). The 'silent-frame' threshold was then expressed as some percentage of this number. By trial and error examination of various ad-break clips, the audio threshold which gave the most sensible and consistent results was:

$$\mathbf{A_{th} = 0.073 * audio\_level_{avg}} \qquad \mathbf{(2)}$$

4.3 Procedure

4.3.1 Video Examination

- The DC-DCT coefficients of each Y-block of each frame were stripped from the video bitstream of the MPEG file.
- An average luminance DC-DCT coefficient was then calculated for each video frame of the sequence.
- The overall mean value of the average frame coefficients for the clip was determined. The 'black-frame' threshold was then defined as the value corresponding to 48% of this number. [see **(1)**]
- Each frame's average DC-DCT coefficient value was compared to the threshold and if equal/less than, then the frame was labeled 'black'.

4.3.2 Audio Examination

- The scalefactors corresponding to the first 10 subbands of the encoded audio signal were stripped from the bitstream.
- An audio level for each video frame was determined by averaging the scalefactors corresponding to its associated audio signal.
- The overall mean value of the video frame audio levels for the entire clip was determined. The ‘silent-frame’ threshold was then defined as the value corresponding to 7.3% of this number. [see **(2)**]
- Each video frame’s representative audio level was compared to the threshold and if equal/less than, then the frame was labeled ‘silent’.

4.3.3 Search for Simultaneous ‘Black/Silent’ Frames

- Unless 4 distinct series of at least 6 consecutive simultaneously ‘black/silent’ frames were detected within 90secs (2250 frames) of each other, any such series were ignored.
- Upon detection of 4 distinct series of at least 6 consecutive simultaneously ‘black/silent’ frames, within 90secs (2250 frames) of each other, the minimum requirements for such a sequence to be recognised as an ad-break have been completed.
- Further such series detected within the same allowable window [90secs (2250 frames)] from the end of the previous would also be recognised as being part of the same ad-break.
- Until there is no further detection of such series within the allowable window, all detected series are recognised as corresponding to the same ad-break.

- The detected ad-break is then said to have begun with the first ‘black/silent’ frame of the first detected/recognised series and ended with the last frame of the last detected/recognised series.
- The process begins again upon detection of the next (unrelated) series.

5. Results & Examination

5.1 Test Material

Físchlár provided 10 short television programme clips from 4 different channels [labeled (a), (b), (c) & (d)] in MPEG-1 Layer-II format. The recordings were meticulously chosen such that they exhibited significant content diversity with at least one complete ad-break somewhere in the middle.

5.2 Results

The abovementioned procedures were executed on all 10 clips.

Evaluation of the system was performed by comparison of results achieved against a manual record of the true location of the ad-breaks, to the nearest second, within each clip.

Results are tabulated in Table-1.

[For shorthand purposes, a detected series of at least 6 consecutive ‘black/silent’ video frames as previously described, is henceforth labeled as a **flag**.]

5.3 Result Examination

The absence of **falsely identified** content is evident from the results in Table-1.

In clip '**(b)** Sports Show', the detected end of the ad-break is at 770 seconds, but the true end is later, at 790 seconds. This is recorded as 20 **missed** seconds of ad-break material.

The total number of seconds in the true ad-break (= 790 – 603) = 187 seconds.

The total number of seconds detected as ad-break (= 770 – 603) = 167 seconds.

A superposition of these quantities over all 10 clips was performed to give four overall values, which are illustrated in Figure-2.

5.3.1 Precision & Recall

For a better insight into the individual accuracy of each clip, two important figures of merit for the ad-detection system were calculated:

- The **Recall** measure looks at the percentage of detected material corresponding to true ad-breaks.

Recall =

$$\frac{100 * [\text{Length of ad-break (seconds)} - \text{No. seconds missed}]}{\text{Length of ad-break (seconds)}}$$

- The **Precision** measure is a percentage showing how accurate the system is at exclusively detecting ad-break material.

Precision =

$$\frac{100 * [\text{Length of ad-break (seconds)} - \text{No. seconds missed}]}{\text{Length of ad-break (seconds)} - \text{No. seconds missed} + \text{No. seconds falsely identified}}$$

Example:

For clip '**(b)** Sports Show' the Recall/Precision figures were calculated as follows (from the information in Table-1):

Length of ad-break (= 790 – 603) = 187 seconds

No. seconds falsely identified = 0 seconds

No. seconds missed = 20 seconds (ad-break-end detected 20 seconds early)

Precision = 100 * [(187 – 20) / (187 – 20 + 0)] = **100**

Recall = 100 * [(187 – 20) / 187] **= 89.3**

Results following similar calculations performed on all clips are presented in Table-2.

6. Conclusions

6.1 Result Evaluation

In all, 10 clips comprising 315 minutes of digital video, incorporating 11 ad-breaks, were analysed. The following are the main points to be noted:

- The system detected the occurrence of all 11 ad-breaks.
- A total of 14 'black/silent' flags which were not associated with ad-breaks were detected (7 of these occurring in one clip alone). However, the number of falsely detected ad-breaks was zero.
- All detections attained a Precision percentage of 100%.
- 8-out-of-11 of the detections attained a Recall percentage greater then 98%.

Clips '**(b)** Sports Show, **(d)** Comedy Quiz, & **(a)** News Broadcast' performed relatively poorly compared to others. It was noted that their common downfall was the ad-break-end being detected prematurely. Manual investigation showed that the reason for this was that for all three clips, the final flag in the ad-breaks (occurring between final ad and return of programme) consisted of less than 6 consecutive frames, thus violating one of the three conditions imposed on the detection process. Consequently, the system did not detect the occurrence of the final advertisement within the ad-break, but identified the ad-break-end with the end of the last <u>detected</u> flag (which actually corresponded to the end of the penultimate advertisement).

Numbers of individual advertisements comprising overall ad-breaks were counted for the three clips. It was then concluded that clips '**(b)** Sports Show' and '**(d)** Comedy Quiz' resulted in 1-out-of-7 and 1-out-of-6 ads missed within their respective ad-breaks, giving Recall percentages of **89.3** and **86.9**. However, clip '**(a)** News Broadcast' resulted in 1-out-of-4 ads missed within its ad-break, which represented non-detection of a much larger proportion of the overall ad-break. Hence, this clip only attained a meagre Recall percentage of **68.6**.

The consistently high Precision figures indicated high success in the prevention of mis-recognition of content not corresponding to ad-breaks.

6.2 Further Work

The system succeeded very well until, on 3-out-of-11 occasions, the conditions incorporated to combat false identification of ad-breaks, actually prevented the detection of the final advertisement within their respective ad-breaks. For this reason (and due to the fact that the system will fail for (i) occurrence of individual advertisements longer than 90secs in duration, and (ii) ad-breaks consisting of less than 3 advertisements), any future work could possibly involve further optimisation of the trade-off between consistently precise ad-break detection and the prevention of false identification, either by changing the parameters in the existing conditions or by replacing them with improved ones.

In introducing this subject, it was mentioned that most advertising television stations feature the characteristic of ‘black/silent’ frame gap in between individual advertisements. In fact, of the six advertising television channels captured by *Físchlár*, two do not exhibit this trait. They instead have each individual advertisement run directly into each other with no pauses in between. Consequently, the discussed method of ad-break detection would fail for these broadcasts and so an alternate method is required. A proposed technique is to examine the rate of shot-cuts over an entire programme. This rate is expected to be notably high during ad-breaks since advertisements characteristically exhibit a consistently high rate of activity. This, coupled with some timing aspects (e.g. broadcast time is usually sold in discrete units) could provide sufficient information enabling an alternative method for automatic advertisement/programme differentiation.

References

[1] Lee H., Smeaton A., O'Toole C., Murphy N., Marlow S. & O'Connor N., *The Físchlár Digital Video Recording, Analysis, and Browsing System*, RIAO 2000 – Content-based Multimedia Information Access. Paris, France, 12-14 April 2000.

[2] Carroll, D., *Advertisement Detection*, Internal technical report, Dublin City University 2000. (contact corresponding author)

[3] *The Official MPEG Committee Website*. http://www.cselt.it/mpeg/

[4] Marshall D., *Video and Audio Compression Website*. http://www.cs.cf.ac.uk/Dave/Multimedia/node196.html

[5] Rao, K.R. & Hwang, J.J., *Techniques & Standards for Image, Video and Audio Coding*. Prentice Hall, 1996.

[6] Bourquin B., Frey M. & Wetzel R., *NOMAD Project*. http://www.fatalfx.com/nomad/

[7] Lienhart, R., Kuhmunch, C. & Effelsberg, W., *On the Detection & Recognition of Television Commercials*, Proc. IEEE Conf. on Multimedia Computing and Systems, pp. 509 - 516, Ottawa, Canada, 1996.

[8] Fry, D., Hampshire, E., Hargrove, T., *Automatic Detection of Commercials within Pre-Recorded Cartoon Shows*. Preliminary project design report. http://www.cse.scu.edu/projects/2000-01/project12/

[9] Li, Y. & Jay Kuo, C.-C., *Detecting the Commercial Breaks in Real TV Programs Based on Audiovisual Information*, Proc. SPIE Vol. 4210, pp. 225-236, Internet Multimedia Management Systems.

About the Authors

David A. Sadlier received his Bachelor of Electronic Engineering Degree from the National University of Ireland in 2000. His research interests are mainly in signal processing for digital audio/video analysis and is currently pursuing a Masters degree of E. Eng in this field at Dublin City University.

Dr. Sean Marlow, MIEI, CEng, received a BSc (Hons) degree in Electrical and Electronic Engineering from Queen's University, Belfast in 1976 and a PhD in signal processing from the same University in 1979. From 1979 to 1980 he was assistant lecturer in University College, Cork. From 1980 to the present he was employed at

Dublin City University (formerly NIHE Dublin), first as lecturer, then as senior lecturer in electronic engineering. Dr Marlow is joint Irish representative on the COST 211ter (Video Compression for Telecomms) Management Committee. He is co-Director of the Visual Media Processing Group and Centre for Digital Video Processing

Dr. Noel O'Connor has been a lecturer in the School of Electronic Engineering at Dublin City University Since July 1999. He is currently the Programme Chair for the School's new BEng in Digital Media Engineering degree programme. Noel is the Irish representative to the European COST 211 action and has also been an Irish representative to the ISO/IEC MPEG standards body on a number of occasions. His plans for future research include continuing his work on video compression, video object segmentation and investigating the role of image and video analysis techniques in future multimedia archiving and indexing applications.

Dr. Noel Murphy received his degree in theoretical physics from Trinity College Dublin in 1985 and subsequently began work as an investigative researcher in Computer Vision at the School of Electronic Engineering in DCU. In 1986 he began as a part-time lecturer at the School while he finished his PhD on an information theoretic approach to visual perception. He has held a permanent lecturer position in the School since 1994.

Fig. 1. Interpretation of ‘black/silent’ frame series

Fig. 2. Total number of seconds corresponding to ad-breaks, detected ad-breaks, missed ad- breaks and falsely identified ad-breaks.

Table 1. Results of experimentation on 10 video clips provided by 4 television stations comprising 11 advertisement breaks.

Table 2. Precision & Recall values based on results in Table 1.