Analysis and Processing of Compressed Newsfeeds

Guido FALKEMEIER, Gerhard R. JOUBERT and Odej KAO
Technical University of Clausthal, Department of Computer Science,
Julius-Albert-Str. 4, 38678 Clausthal-Zellerfeld, Germany
Tel: +49 5323 953-140; Fax: +49 5323 953-149;
Email: guido.falkemeier@informatik.tu-clausthal.de,
gerhard.joubert@informatik.tu-clausthal.de, odej.kao@informatik.tu-clausthal.de

Abstract. In this paper methods for analysing and processing digitised compressed video newsfeeds are discussed. Due to computing, storage and communication limitations the material has to be processed in compressed form. This creates particular problems which are not encountered when decompressed video material is handled. The methods described allow news editors to directly extract information from the compressed material, creating the possibility that only those video clips which may be of interest need be decompressed. News editors can decide which sections of a newsfeed should be extracted and assembled for a particular news cast. The final cutting can then be done from the original broadcast quality material. This paper reports on results obtained with a joint research project with a commercial television station[1].

1. Introduction

Television stations receiving newsfeeds from various news agencies, such as Reuters, APTV or WTN on a continuous twenty four hour basis are confronted with a very large video processing task. Each newsfeed consists of a large number of assembled news clips.

Processing and selecting news items from a feed is usually done by copying the videos received via satellite onto conventional tapes. These are then manually scanned by editors for particular news items. Selected items are subsequently copied and assembled for news casts. In view of the increasing volume of material to be scanned and the limited number of professional workstations and copies of newsfeed videos available, it becomes increasingly difficult for editors to execute their task satisfactorily with present equipment and procedures.

Digital processing of videos could facilitate this process. Editors could, for example, work from their offices using desk computers to extract and view material stored on a central server. This reduces the need for specialised and very expensive professional video processing stations. In order to handle the large data volumes in a client/server environment the analogue videos must be digitised and then compressed into, for example, MPEG-1 format. This format does not supply broadcast quality material, which can at present only be attained by professional equipment. This necessitates the recording of the newsfeeds in both broadcast quality and (lower quality) digitised form.

Although compression of the digitised video's greatly reduces the storage and communication requirements, it creates serious difficulties regarding the processing of news feeds. It proves too time and space consuming to decompress the videos in order to find a particular news item in the data stream. It is thus desirable to have methods available by

[1] This project was executed with the financial support and co-operation of Radio Television Luxembourg (RTL).

which the beginning of a particular news item can be detected in the compressed video data. Subsequently only that particular clip need be decompressed and viewed by the editor.

In this paper methods which achieve this are discussed. The present pilot system extracts data from the compressed video stream and supplies editors with a list giving the start and end points of each news clip. This list is used to select, decompress and view particular news clips. In addition the time code of each frame can be extracted, thus facilitating the selection and assembly of the broadcast quality material from the broadcast quality copies.

Although the availability of digitised videos on a network seems to offer great advantages to editors, this is not necessarily the case in practice. The work flow practices of editors and their relationship with technicians executing the final assembly of news casts are directly affected by the digital video processing system. One aspect is, for example, that analogue material can be viewed at high speed. This is not possible with digitised videos made available on an affordable network. On standard networks the "high speed" playing of a video implies that only every n-th frame is displayed. This creates a display which is unacceptable to news editors. It is thus essential that additional instruments must be made available to editors in order to allow them to execute their task as efficiently and effectively as is the case with the traditional systems. In order to achieve this it is essential to supply editors with at least the following information:

- Start frames of the video clips contained in a newsfeed
- A summary (textual) of the news items contained in a continuous sequence of material.

As was mentioned already, this information must be extracted from the compressed video material, as a decompression is not feasible due to time, space and communication limitations.

In order to understand the information extraction process it is necessary to understand the structure of a newsfeed. A feed consists of a continuous stream of news items. The individual items are separated by a special frame sequence. This sequence can by divided into three subsequences. The first consists of only black frames followed by a few frames giving some information of the news item which is the next in the newsfeed (title, duration). These frames are again followed by a few black frames until the first news item starts. Figure 1 shows a typical sequence.

Figure 1: Black and Content Frames between two news items

2. Finding Black Frames

The first step is to find the start frames of each group in a newsfeed. This means that the sequences of black frames must be determined directly from the compressed videos. In order to achieve this it is necessary to know how MPEG compression is achieved and the format of the files generated. The MPEG-1 compression method, used in the case of the videos considered in this paper, has been standardised and the reader is referred to the literature [2 - 4]. To summarise: a MPEG compressed video consists of three different frame types. These are:

- Intra-frames (I-frames)
- Predicted frames (P-frames)
- Bidirectional predicted frames (B-frames).

The I-frames are "base" frames which are directly compressed. I-frames can thus also be decompressed without reference to other frames. The P- and B-frames are referenced to I-frames before compression. They can thus not be directly decompressed. I-frames are selected at regular intervals. The MPEG standard allows for a certain freedom of choice. In the compressed videos considered here an I-frame is selected every 12 - 15 frames.

In order to identify a sequence of black frames in a compressed video the analysis can be limited in the first instance to a search for the I-frames in a black sequence. This is due to the fact that at least one I-frame is always contained within a black frame sequence. The first problem to be solved is thus whether black I-frames can be determined from a compressed stream of frames.

2.1 The Theoretical Case

The I-frames are compressed using the discrete cosine transform (DCT) with a following Huffman coding, which is similar to that used in JPEG compression. It can be shown that this transform results, in the case of a black frame, in the value 128 for the DC-component, whereas the AC-components all have the value 0. These values are obtained for each of the three constituents Y (intensity), C_b (colour difference blue) and C_r (colour difference red) of the colour model used by MPEG [5]. In accordance with the MPEG standard the coding of the macro blocks must be inspected. Each macro block comprises six 8×8 pixel blocks of which four represent the Y component and one each the C_b and C_r components respectively. Thus the occurrence of black frames can be determined in the compressed video.

This theoretical solution does not, however, produce reliable results in practice. This is due to the fact that even very little noise in the original significantly changes the coding. In order to analyse the sensitivity of the compression technique to noise in the original image, a black image with a single white pixel was analysed. The result differs significantly from that for a completely black frame. One possibility is to skip the macroblock containing the noise. To implement this a threshold value for the number of macroblocks which can contain noise, but still allows the image to be classified as "black", must be set.

The determination of a suitable threshold, however, results in the following problem: Consider the image shown in Fig. 2. This image is clearly not "black". The coding of the white macroblocks is, however, identical to that of the black blocks. This means that there are only two macroblocks which differ from the "black" (and "white") macro blocks. These are the two blocks containing the transition from black to white and from white to black. Thus, even a threshold value of two is not usable in practice. This method is thus also unsuitable for finding black sequences in news feeds.

Figure 2: False detected black frame

From these considerations it becomes clear that the search for sequences of black frames is easily solved in theory. In practical situations, however, the identification of such sequences in environments where noisy images are inevitable, this approach or simple adaptations of it do not work. More sophisticated methods are thus needed

.

2.2 Detection by Partial Decoding

By a more detailed analysis of the DC-components of the 8×8 pixel blocks contained in the macro blocks, leads to a further possibility for finding sequences of black images. This requires the partial decoding of the DC-coefficients by use of the Huffman coding scheme. It should be noted that the compute intensive inverse DCT is not required.

The reason for the fact that only the DC components need be treated results from the strong correlation between the original image and the codec of an image in which only the respective DC components are considered. This aspect is described in [6] in more detail. In the method developed here only the DC components of the I-frames are thus considered. The instrument used to identify a black image are the respective histograms of the images considered. Fig. 3 shows histograms for a completely black as well as a randomly chosen video frame. The differences are clear. The black image has only one luminance value in contrast to normal images, which show a histogram covering a range of values. This characteristic can be used to differentiate not only between monochromatic images and others, but also between monochromatic images, such as “black”, “red”, “green”, etc.

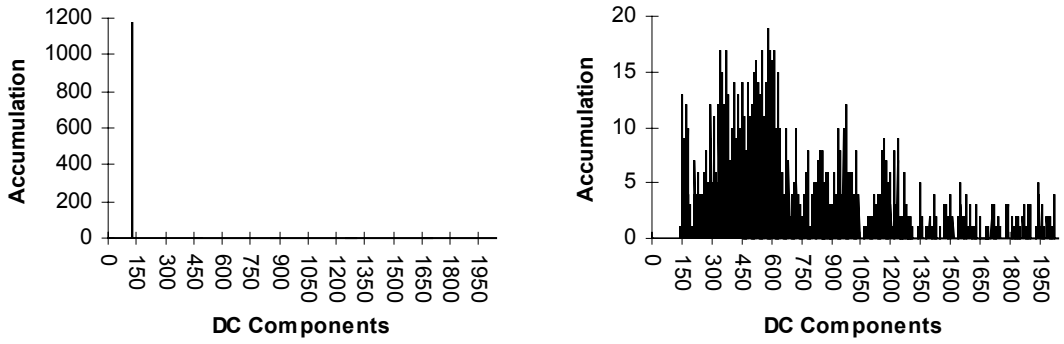

Figure 3: Histogram of a black frame (left) and a normal frame (right)

frames. By implementing a partial decoding of the DC components of I-frames, calculating the histograms and using a suitably chosen threshold value, black frames can be identified. Numerous tests with this simple method showed that sequences of black frames can be detected with nearly hundred percent accuracy. It should be noted that noise in an image has

little or no effect on the DC components, as the higher noise frequencies mainly affect the AC values. This results in the relative insensitivity to noise of this method.

A further point of interest is the accuracy with which a whole newsfeed is analysed. In the identification of black frame sequences only the I-frames are considered. A particular black frame found is thus not necessarily the first frame of a black sequence. In the first phase of the project the goal was only to detect the existence and approximate position of a black frame sequence. Once such a sequence has been detected an analysis of the B- and P-frames prior to and after the I-frame can supply time code information about the start and end frames of the sequence. This can be achieved by analysing the types of the macroblocks of the compressed B- and P-frames, thus also obviating the need for decompression.

3. Detection of Content Frames

It was mentioned already that content lists are supplied in news feeds at the beginning of a sequence of news items. This information is important to editors as it gives an indication of the theme and the length of a particular news clip. The content lists are embedded in black frame sequences. The goal is thus, once a black frame sequence has been identified, to localise at least one frame containing the content list. Subsequently the textual information can be extracted.

Two methods to achieve this can be used:

- The frame, midpoint between the first frame of the first black sequence and the last frame of the second black sequence, is calculated. The closest I-frame is then extracted. The probability of this I-frame containing the required textual information is high as the information frames are usually situated in the middle of the header.
- The differences between the DC components of the sequence of frames in the header, starting from the first frame of the first black sequence, can be computed. This can be used to localise the frames containing textual information.

The second possibility results in:

Definition 1: DCI_k indicates the ordered set of DC components of the I-frame k. The DC components are ordered in the sequence in which the 8×8 blocks appear. l is the number of all I-frames in the sequence being considered. With $x_{m,n} \in DCI_k$ and $y_{m,n} \in DCI_{k+1}$ the distance D between two succeeding sets is given by:

$$D_{i,i+1} = \sum_{m,n} \left| x_{m,n} - y_{m,n} \right|$$

Here m, n indicate the number of rows and columns, $i=1..l$.

The distance metric from Def. 1 defines a sequence of difference images. Within this sequence two difference images will show up a greater distance than the rest. In the first instance the difference between the last I-frame without text content and the first I-frame with text content is relatively large, as is also the difference between the last I-frame with text content and the following I-frame without text content. This is shown in Fig. 4. The data was determined from sample news feeds from Reuters (left) and WTN (right).

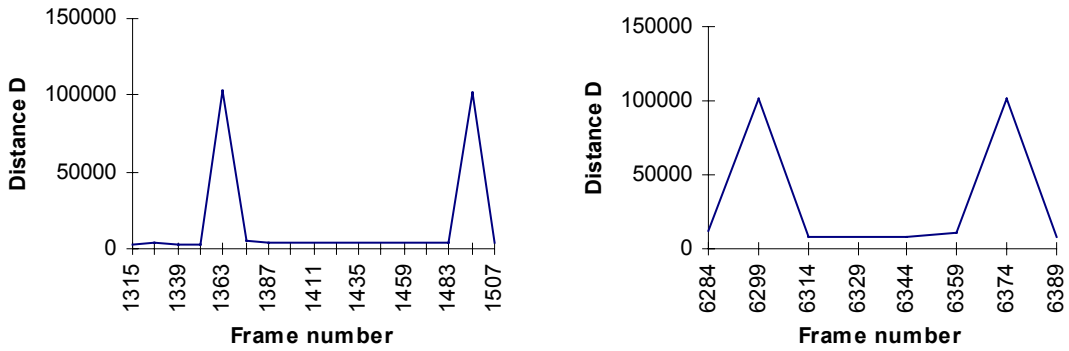

Figure 4: Sequence of difference images (Reuters and WTN)

The two frames with higher differences can be identified directly. It can also be seen that the frame sequence containing textual information is longer in the case of Reuters than WTN. Furthermore, the black sequence before the text frames is, in the case of Reuters, longer than the black sequence following the text sequence. The black sequences are relatively short in the WTN feed.

In order to find an I-frame which does contain textual information, the following procedure is used: The two peaks, as in Fig. 3, are determined. An enclosed I-frame is then searched for.

In order for a change from black sequence to text sequence to be located between frames i and $i+1$ or vice versa, the following condition must be met:

$$D_{i,i+1} > 4 \cdot D_{av}, \text{ with } D_{av} = \frac{\sum_{i=1}^{l-1} D_{i,i+1}}{l}$$

The result is two I-frames a and e in the interval 1...l. For the I-frame containing textual information it must hold that a > b > e. This condition is met by a number of I-frames, i.e. b is not unique. Anyone of these candidates can be selected for the extraction of the textual information.

The news agency APTV uses a different format for announcing the contents of a feed. In this case a list of all items in the feed is first given, followed by textual information about the subsequent items at the start of each. This results in the sequence of difference images, see Fig. 3, containing three peaks instead of two. The frames needed for the extraction of textual information about the particular clip are located between the second and third peak. The condition given above remains the same, except for the factor four, which must be reduced. This is the result of the higher value of D_{av} obtained from the three frames with large differences. In practical tests good results were obtained with a factor of two.

Practical tests with a variety of news feeds showed that the methods described here can reliably extract frames containing textual information.

4. Information Extraction of the Content Frames

The frames containing the text information are characterised by a dark background and bright, sometimes coloured, lettering. The text is set up automatically, so that the letters and the lines are depicted in a regularly spaced grid. In each content frame some key-words occur, such as DURATION, SHOT and SOUND, which are necessary for the interpretation of the information. The key words and the font style depend upon the news agency.

Additional to the statements on the length of the video clips, the language and the date of recording, the location (country or city) and a title of the story are given. Though this data is

not sufficient to completely describe the contents of a video clip, they are a useful indication of the contents, also for use in other information retrieval applications. These could, for example, be the on-line available databases of publishers, Internet distributions of the daily press or the news wires of the news agencies. By the connection of these media all information relevant to the video clip is available in compact form at the editor's desk.

This information can be combined with additional information on the same topic, also items published at an earlier or later date, and used to retrieve information on the particular topic from archives. A search for a certain video clip can be done by utilising a text pattern matching algorithm.

After detecting the position of the content frames by using the methods explained in section 3, the respective frames can be extracted from the MPEG-compressed data stream, converted to the RGB colour model and saved as a file of a standard picture format, such as JPEG, GIF, TIFF, etc. The text information of the frame is then analysed by a suitable OCR system.

The extracted frames are usually not well-suited for optical character recognition as:

- the text may be presented in different colours,
- the background may have various grey shades, which could be recognised as noise,
- many of the OCR-systems expect dark text on a white background,
- the letters can contain some noise and
- the content frame is equivalent to a resolution of 200 dpi, which is not sufficient for character recognition.

The result is that standard OCR systems have difficulties producing reliable results. Thus, for example, reading a "O" instead of an "Q" and an "i" instead of a "1" is quite common.

These problems can partly be solved by image preprocessing. The content frame can be represented as a m×n array p with $p(i,j)=(r(i,j), g(i,j), b(i,j))$, where $r(i,j), g(i,j), b(i,j) \in [0,255]$ and m, n are the dimensions of the frame. For optical character recognition a monochrome image is needed. Therefore the colour frame must be converted first to a halftone image g, with $g(i,j) \in [0,255]$. A simple method is to combine and normalise the three RGB channels by the following formula:

$$h(i,j) = [r(i,j) + g(i,j) + b(i,j)] / 3 \quad for \quad all \quad i = 1,..,m, \quad j = 1,..,n$$

In the second step the halftone image is binarised. For this purpose the method with a local adaptive threshold [7] can be used. A monochrome image b consists of only two grey levels: 0 (black) and 255 (white). By introduction of a threshold value t the grey levels of the halftone image can be separated. The threshold value t can be determinated through calculation of an image histogram. Problems with the noise can be avoided, when the neighbouring points are considered and the threshold is locally adjusted. Through the binarisation unnecessary grey levels of the background can be eliminated. The required black text on a white background can be obtained by inverting the image.

To close small gaps in the contours of the letters morphological operators such as OPENING and CLOSING [5] can be used. Morphological image processing modifies the spatial form or structure of objects (letters) within an image. The majority of morphological operators can be defined as “hit or miss” transformations. A small odd-sized mask (typically 3×3) is scanned over a binary image. If the binary-valued pattern of the mask matches the state of the points under the mask (hit), then an output point in spatial correspondence to the centre point of the mask is set to some desired binary state. For a pattern mismatch (miss), the output point is set to the opposite binary state.

Two particular morphological operations are OPENING and CLOSING. CLOSING of an image with a compact structuring element without holes smoothes contours of objects,

eliminates small holes in objects and fuses short gaps. OPENING of the image eliminates small objects (noise).

Tests showed that these transformations resulted in images which are more amenable to OCR systems. The result is a considerably more reliable interpretation of the textual information contained in context frames. The output of the OCR system is a file containing the recognised text lines. From these, location, date and a short description can be determined.

This information can be used as input for search machines of online data bases, or to find the respective reports of the daily press. In addition a full text query on the archived videos can be performed to find previous reports on the same topic.

The sum of information found this way is presented to the editors as an automatically created HTML page.

5. Conclusions

In this paper methods which reliably extract information from compressed news feeds are presented. Working directly on the compressed material enables the implementation of the methods on standard computer networks due to the low demands made on computing and communication performance. Some of the methods proposed are already in regular use. Additional functions are, however, required to make the digital system fully compatible with present work practices as well as offering enhanced capabilities compared to analogue processing techniques.

Future investigations will be directed at generalising the black frame detection methods to enable their application to MPEG-2, etc. compressed video material. Furthermore, it is to be investigated whether the additional use of statistical methods can improve the detection accuracy by reducing the sensitivity to changes in the characteristics of the original material.

In order to give better support to news editors in compiling news casts it is desirable to automatically generate links to relevant information available on internet, which may be obtained through datamining or other techniques. The extracted and compiled information must be stored in a way which will facilitate its future recovery. Considerable work is required to define and standardise appropriate data structures.

References

[1] S. Geisler und Usa Luengjiranothai, Automatische Erkennung von Themenwechseln in MPEG kodierten Videos, Studienarbeit an der TU Clausthal, February 1998.

[2] ISO/IEC 11172-1, Information technology --- Coding of moving pictures und associated audio for digital storage media at up to about 1,5 Mbit/s ---, Part 1: Systems, 1993.

[3] ISO/IEC 11172-2, Information technology --- Coding of moving pictures und associated audio for digital storage media at up to about 1,5 Mbit/s ---, Part 2: Video, 1993.

[4] ISO/IEC 11172-4, Information technology --- Coding of moving pictures und associated audio for digital storage media at up to about 1,5 Mbit/s ---, Part 4: Compliance Testing, 1993.

[5] W.K. Pratt, Digital Image Processing, John Wiley and Sons, Inc New York, 1991.

[6] B-L. Yeo and B..Liu, Rapid scene analysis on compressed video, IEEE Transactions on circuits and systems for video technologie, vol.5, Dezember 1995, pp. 533-544.

[7] P. Zamperoni, Methoden der digitalen Bildsignalverarbeitung, Vieweg, 1989.