TEMPORAL VIDEO SEGMENTATION FOR REAL-TIME KEY FRAME EXTRACTION

J. Calic[1], S. Sav[2], E. Izquierdo[1], S. Marlow[2], N. Murphy[2] and N.E. O'Connor[2]

[1]Department of Electronic Engineering, Queen Mary, University of London
[2]Centre for Digital Video Processing, Dublin City University

ABSTRACT

The extensive amount of media coverage today, generates difficulties in identifying and selecting desired information. Browsing and retrieval systems become more and more necessary in order to support users with powerful and easy-to-use tools for searching, browsing and summarization of information content. The starting point for these tasks in video browsing and retrieval systems is the low level analysis of video content, especially the segmentation of video content into shots. This paper presents a fast and efficient way to detect shot changes using only the temporal distribution of macroblock types in MPEG compressed video. The notion of a dominant reference frame is introduced here. A dominant frame denotes the reference frame (I or P) used as prediction reference for most of the macroblocks from a subsequent B frame.

1. INTRODUCTION

The trend of modern society towards increasing information access raises requirements for advances in the development of multimedia technologies. Advances in multimedia compression standards combined with increased performance of computers networks and storage media supports wide distribution of multimedia information. However, building large digital collections of data, in order to allow facile access, complicates searching, browsing and retrieval of data. Indexing techniques have been developed in order to support the operations mentioned above.

Regarded as an essential task of video indexing, shot segmentation [1] represents the process of temporal identification of the boundaries of camera shots. A shot can be defined as "a single sequence of frames in motion picture obtained by one camera without interruption" [2]. Connecting shot-level analysis with additional information allows indexing of the video content at the scene level [3]. Appending to that semantic boundary information, a collection of different scenes semantically related would be indexed as a story line [4].

Originally based on the uncompressed-domain video features, with the introduction of the MPEG standard the shot segmentation algorithms become more and more oriented towards compressed-domain, in order to allow real-time processing. Obviously, there is a trade-off between the accuracy of prediction offered by uncompressed domain features which are more information consistent, and real-time processing obtained by avoiding full decompression of encoded video bitstream which are limited to the MPEG video compressed domain features. The approaches primarily based on uncompressed or partially decompressed features are likely to produce better prediction of shot boundary occurrences. However, the increase of accuracy is inconsequential in comparison with the processing speed gained using only compressed domain features.

Extending the above consideration for compressed domain approaches, it emerges that as fewer and more basic features are used for prediction, the faster and less computationally demanding the algorithm. Evidently, once the algorithm reaches real-time the speed increase would not constitute a major factor. However, a simple approach would decrease the computational requirements.

In this paper we propose an efficient and robust method for shot change detection for MPEG compressed video based only on the temporal distribution of macroblock types. The experimental results are promising showing the effectiveness of the described algorithm for hard cuts detection. It also opens the possibility of extending the actual approach for detection of gradual transitions such as dissolves, wipes and fades. Moreover, reviewing other research covering shot detection, it is obvious that the method presented in this paper seems to assume the fastest and most computationally modest approach, with similar results.

The paper is organized as follows: Section 2 covers a review of related research. The proposed metric for evaluation of the frame differences is described in Section 3. Details of our approach for shot segmentation are presented in Section 4. Experimental results are presented in Section 5 while, finally, in Section 6 conclusion and a summary of the paper are presented.

2. RELATED WORK

First methods for shot segmentation used video features from the uncompressed domain, such as pixelwise difference [5], histograms [6], edge tracking [7], etc.
The introduction of the MPEG standard has redirected the effort for video segmentation in the compressed-domain. Methods based on partially decoded domain [8] or compressed domain features had been developed, mainly using the DCT coefficients [9, 10] or motion vectors [11].

Two approaches are particularly similar to our method, one the using spatio-temporal distribution of macroblock types for dissolve detection [2] and the second one tracking the scene change for a particular macroblock during the shot transition [12]. Both approaches are using information related to the spatial position of macroblocks within a frame. This paper introduces a method based only on temporal distribution of macroblock types.

3. FRAME DIFFERENCE METRICS

MPEG-2 encoders compress video by dividing each frame into blocks of size 16x16 called *MacroBlocks* (MB) [13]. Each MB contains information about the type of its temporal prediction and corresponding vectors used for motion compensation. The character of the MB prediction is defined in a MPEG variable called *MBType,* and it can be: *Intra* coded, *Forward* referenced, *Backward* referenced and *Interpolated.*

Since the MPEG sequence has a high temporal redundancy within a shot, a continuously strong inter-frame reference will be present in the stream as long as no significant changes occur in the scene [14]. The "amount" of inter-frame reference in each frame and its temporal changes can be used to define a metric, which measures the probability of a shot change in a given frame. We propose to extract only MBType information from the MPEG stream and, by analyzing it, measure this "amount" of inter-frame reference.

Without loss of generality we assume that in any analyzed MPEG stream, the *Group Of Pictures* (GOP) will have the standard structure [IBBPBBPBBPBBPBB]. Observe that this frame structure can be split into groups of three having the form of a triplet: IBB or PBB. In the sequel, both types of the reference frames (I or P) are denoted as R_i, the front bi-directional frame of the triplet as B_i, while the second bi-directional frame is denoted as b_i. Thus, the MPEG sequence can be analyzed as a group of frame-triplets in the form: { $R_1 B_2 b_3$ $R_4 B_5 b_6$ … $R_i B_{i+1}$ b_{i+2} …}. This convention can be easily generalized to any other GOP structure.

As mentioned above, a high visual similarity within a sequence should result in high percentage of predicted MBs in both bi-directional B frames and predicted P frames and lack of intra coded MBs. More precisely, if two frames are strongly referenced then the most of the MBs in predicted frame would have the corresponding prediction type: forward, backward or interpolated, depending on the type of reference [15]. Thus, we can define a metric for the visual frame difference by analyzing the statistics of MBTypes in each frame.

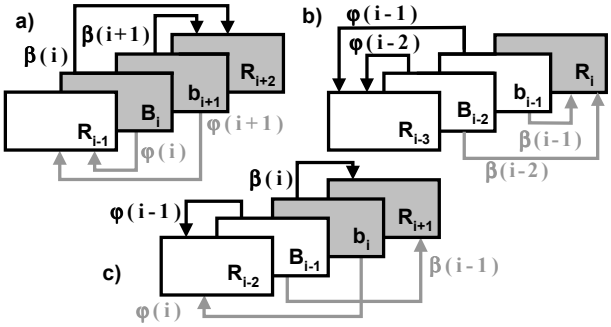

Figure 1. Shot change positions in a frame triplet

The possible locations of a content change (i.e. cut) in a frame triplet are depicted in Figure 1. If the front referenced frame B_i is the first frame with different visual content (a), the next reference frame R_{i+2} predicts backwards a significant percentage of MBs in both B_i and b_{i+1}. If the content change occurs at the rear reference frame R_i (b), then the bi-directional frames B_{i-2} and b_{i-1} will be mainly predicted forwards by the previous reference frame R_{i-3}. Finally, if the content change occurs at b_i (c), then B_{i-1} will be strongly predicted forward by the previous reference frame $R_{i-2,}$ while b_i will be predicted backwards by the next reference frame R_{i+1}.

Let $\Phi_T(i)$ be the set containing all forward referenced MBs and $B_T(i)$ the set containing all backward referenced MBs in a given frame with index i and type T. In the same manner, we define sets of intra coded MBs as $I_T(i)$ and interpolated MBs as $\Pi_T(i)$. Then we denote the cardinalities of the corresponding sets as: $\varphi_T(i)$, $\beta_T(i)$, $\iota_T(i)$ and $\pi_T(i)$. The metric $\Delta(i)$ used to determine a visual difference measure within a frame triplet is defined as:

$$\Delta(i) = k_{\varphi B}\varphi_B + k_{\varphi b}\varphi_b + k_{\beta B}\beta_B + k_{\beta b}\beta_b + k_{\iota B}\iota_B + k_{\iota b}\iota_b + k_{\pi B}\pi_B + k_{\pi b}\pi_b$$

By analyzing the prediction character and behavior in one frame triplet, we can estimate the changes in visual content within. Depending on the frame type, there are three different linear combinations of variables $\varphi_T(i)$, $\beta_T(i)$, $\iota_T(i)$ and $\pi_T(i)$ for both bi-directional frames in a frame triplet. Each linear combination has two main coefficients that are directly proportional to the visual content change within predicted and reference frame in a frame triplet (k=+1), and two that are inversely proportional (k=-1) to it. Additional factors k_π and k_ι are describing overall change in a triplet, one in direct (k_ι) and one in inverse (k_π)

proportion. The coefficient values are determined by the rule of thumb, and are presented in Table 1.

	T(i)=R	T(i)=B	T(i)=b
$k_{\varphi B}$	+1	-1	+1
$k_{\varphi b}$	+1	-1	-1
$k_{\beta B}$	-1	+1	-1
$k_{\beta b}$	-1	+1	+1
$k_{\iota B}$, $k_{\iota b}$	+0.5		
$k_{\pi B}$, $k_{\pi b}$	-0.5		

Table 1. Coefficients in the linear combination for Δ(i)

4. SHOT CHANGE DETECTION

The raw difference metric defined in the previous section has a strong noise that makes further processing of the data almost impossible. However, we know that the source of this noise is in the discontinuous nature of the difference metrics. Since the metrics value is determined separately for each frame and the content change is based on frame triplets, low-pass filtering with kernel proportional to triplet length would eliminate the noise. The filter with Gaussian pulse response is applied:

$$h(i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{i^2}{2\sigma^2}}$$

Where $i \in [-4\sigma, 4\sigma]$, and $\sigma=1.5$. The value for σ is chosen to maximize the smoothing within one frame triplet.

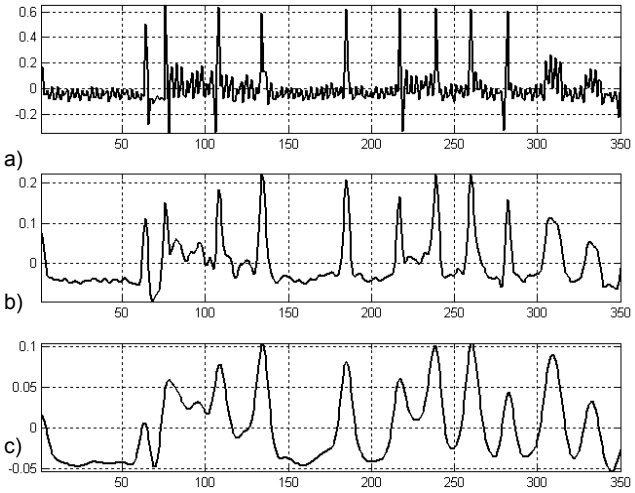

Figure 2. a) Raw metric, b) Metric after the noise suppression, c) Smoothed metric

Metric with suppressed noise is calculated as a convolution of Gaussian filter pulse response and the raw noisy metrics:

$$\Delta = \Delta_N \otimes h$$

Example of noise suppression from a difference metrics is given in Figure 2.

After noise suppression, the same filtering procedure is applied to eliminate small spurious peaks and to smooth the difference metrics function. As in noise suppression, the filtering kernel is Gaussian, but with parameter $\sigma=3$. The positions of the central points in the shot change are determined by locating local maxima of the smooth metrics curve.

5. RESULTS

The collection of C++ classes called Mpeg Development Classes, implemented by Dongge et al. [16], was used as the main tool for manipulating the MPEG streams, while Berkeley mpeg2codec was used as the reference MPEG codec. Test sequences were produced by Multimedia & Vision Research Lab, Queen Mary, University of London, while some were provided by Centre for Digital Video Processing, Dublin City University, Dublin, Ireland.

5.1. Shot detection statistics

Since the algorithm comparison in temporal video analysis is mainly based on evaluation of the shot change detection, we will present the statistics of the experimental results starting with shot detection statistics.

The applied statistical performance evaluation of temporal segmentation of the video sequences is "based on the number of missed detections (MD's) and false alarms (FA's), expressed as recall and precision" [17] :

$$Recall = \frac{Detects}{Detects + MD's}, \quad Precision = \frac{Detects}{Detects + FA's}$$

We took manually detected positions of the shot boundaries as the ground truth, defining in that way the number of missed detections and false alarms. There were three main categories of video material analyzed:

- NEWS; long monotonous sequences with mainly abrupt changes,
- SOAP OPERA; average shot length with some gradual changes and editing effects
- COMMERCIALS; short shots with a lot of gradual changes and editing effects.

The shot changes detection procedure showed excellent results for different types of changes with almost 100% accuracy, as seen in Table 2.

	Detect	MD's	FA's	Recall	Precision
News	87	2	6	98%	94%
Soap	92	2	9	98%	91%
Commercials	127	9	16	94%	88%

Table 2. Shot changes detection results

6. CONCLUSIONS

A novel shot change detection technique based on the motion information extracted from the MPEG video stream is proposed. First, an algorithm for the frame difference metrics extraction that uses simple statistics of the MacroBlock types was introduced. Second, the noise suppression method and the metrics-smoothing algorithm were described. Finally, the experimental results were introduced in Section 5. The results showed high accuracy in abrupt shot detection. However, the method is sensitive to high motion during the gradual changes, and thus those transition types are detected with lower, but reasonable precision.

We are investigating the possibilities of improving the real time gradual shot changes detection using multidimensional clustering of the MPEG compressed features. Additional MPEG features should make the proposed algorithm less sensitive to strong motion during shot changes without losing the real time capabilities.

ACKNOWLEDGEMENTS

The first and third authors acknowledge support from the EPSRC Hierarchical Video Indexing Project under the Grant number R01699/01.

7. REFERENCES

[1] H. J. Zhang, "Content-based Video Browsing and Retrieval", *Handbook of Multimedia Computing*, CRC Press, Boca Raton, Florida, USA, 1999.

[2] S. B. Jun, K. Yoon, H. Y. Lee, "Dissolve Transition Detection Algorithm Using Spatio-Temporal Distribution of MPEG Macro-Block Types", *ACM Multimedia Electronic Proceedings*, 2000. http://www.acm.org/sigs/sigmm/MM2000/ep/jun/ACMMM2000Dissolve.pdf

[3] J. S. Boreczky, L. A. Rowe, "A Comparison of Video Shot Boundary Detection Techniques", *Storage and Retrieval for Image and Video Databases IV*, Proceedings SPIE 2670, pp. 170-179, 1996.

[4] N. O'Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy, A. Smeaton, "News Story Segmentation in The Fischlar Video Indexing System", *Proceedings of ICIP 2001 – International Conference on Image Processing*, Thessaloniki, Greece, October, 2001.

[5] H. J. Zhang, A. Kankanhalli, W. Smoliar, "Automatic Partition of Full-Motion Video", *Multimedia Systems*, vol. 1, no.1, pp. 10-28, 1993.

[6] A. Nagasaka, Y. Tanaka, "Automatic Video Indexing and Full-Video Search for Object Appearances", *Visual Database Systems*, vol. II, pp. 113-127, 1992.

[7] R. Zabih, J. Miller, K. Mai, "A Feature-based Algorithm for Detecting and Classifying Scene Breaks", *Proceedings of ACM Multimedia '95*, pp. 189-200, 1995.

[8] J. Meng, Y. Juan, S. F. Chang, "Scene Change Detection in a MPEG Compressed Video Sequence", *Proceedings SPIE 2419*, pp. 14-25, 1995.

[9] B. L. Yeo, B. Liu, "Rapid Scene Analysis on Compressed Video", *IEEE Transaction on Circuits & Systems for Video Technology*, vol. 5, no. 6, pp. 533-544, December 1995.

[10] S. W. Lee, Y. M. Kim, S. W. Choi, "Fast Scene Change Detection Using Direct Feature Extraction from MPEG Compressed Videos", *IEEE Transaction on Multimedia*, vol. 2, no. 4, pp 240-254, December 2000.

[11] V. Kobla, D. S. Doermann, K. I. Lin, C. Faloutsos, "Compressed Domain Video Indexing Techniques Using DCT and Motion Vector Information in MPEG Video", *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases V*, vol.3022, pp. 200-211, February 1997.

[12] S. C. Pei, Y. Z. Chou, "Efficient MPEG Compressed Video Analysis Using Macroblock Type Information", *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 321-333, December 1999.

[13] D. LeGall ., J. L Mitchell., W. B. Pennbaker, C. E. Fogg, *"MPEG video compression standard"*, Chapman & Hall, New York, USA, 1996.

[14] J. Calic and E. Izquierdo, "Towards Real-Time Shot Detection in the MPEG Compressed Domain", *Proceedings of WIAMIS'2001 - Workshop on Image Analysis for Multimedia Interactive Services*, Tampere, Finland, May 2001.

[15] J. Calic and E. Izquierdo, "Temporal Segmentation of MPEG video streams", submitted to *Special issue on Image Analysis for Multimedia Interactive Services, EURASIP Journal on Applied Signal Processing*, 2001

[16] L. Dongge, I. K. Sethi, "MDC: a software tool for developing MPEG applications", *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, vol. 1, pp.445-450, 1999.

[17] U. Gargi, S. Strayer, "Performance Characterisation of Video-Shot-Change Detection Methods", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 1, February 2000.