jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2012, VOL. 58, NO. 2, PP. 129–134
Manuscript received August 1, 2011; revised May, 2012. DOI: 10.2478/v10177-012-0018-2

GOP Structure Adaptable to the Location of Shot Cuts

Lenka Krulikovská and Jaroslav Polec

***Abstract*—In this paper we present a novel two stage algorithm for improving video coding efficiency. The proposed method combines video cut detection and adaptive GOP structure. At first, we have proposed a new technique of frames' comparison for the shot cut detection. The majority of existing methods compare pairs of successive frames. We compare actual frame with its motion estimated prediction. We also present adaptive threshold. The efficiency of novel technique for video cut detection was confirmed through experiment and compared to the commonly used ones in the terms of recall and precision. The next step is to situate I frames to the positions of detected cuts during the process of video encoding. Finally the proposed method is verified by simulations and the obtained results are compared with fixed GOP structures of sizes 4, 8, 12, 16, 32, 64, 128 and GOP structure with length of entire video. Proposed method achieved the gain in bit rate from 15,33% to 50,59%, while not degrading PSNR in comparison to simulated fixed GOP structures.**

***Keywords*—shot cut detection, Pearson correlation coefficient, motion estimation, adaptive threshold, video encoding, adaptive GOP structure.**

I. Introduction

PROGRESS in the multimedia compression technology and computer performance has led to the widespread availability of digital video. There is a corresponding growth in the need for methods to reliably detect shot boundaries within the video sequence and for higher compression ratio without degrading quality. The research was focused on this two fields separately for many years, but both fields have similar features. Therefore it is more convenient to focus of shot boundary detection and video encoding (adaptive GOP structure) together.

A. Shot Boundary Detection

The detection of shot boundaries provides a base for nearly all video abstraction and high-level video segmentation approaches. Therefore, solving the problem of shot-boundary detection is one of the major prerequisites for revealing higher level video content structure. Moreover, other research areas can profit considerably from successful automation of shot-boundary detection processes as well.

The shot is an elementary building block of video sequences and it is defined as sequence of consecutive frames, which were caught by one camera in one time in single action [1]. An example of abrupt cut is shown in Fig. 1.

Research described in the paper was financially supported by the Slovak Research Grant Agencies: KEGA under grant No. 119-005TVU-4/2010 and VEGA under grant No. 1/0602/11.

L. Krulikovská and J. Polec are with the Institute of Telecommunications, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Ilkovičova 3, Bratislava, SK-812 19, Slovak Republic (e-mails: krulikovska@ktl.elf.stuba.sk; polec@ktl.elf.stuba.sk).

There are a number of different types of transitions or boundaries between shots [1] to create higher video units.

A cut is an abrupt shot change that occurs in a single frame. A fade is a slow change in brightness usually resulting in or starting with a solid black frame. A dissolve occurs when the images of the first shot get dimmer and the images of the second shot get brighter, with frames within the transition showing one image superimposed on the other. A wipe occurs when pixels from the second shot replace those of the first shot in a regular pattern such as in a line from the left edge of the frames. Of course, many other types of gradual transition are possible.

Different approaches have been proposed to extract shots. The major techniques used for the shot boundary detection are pixel differences, statistical differences, histogram comparisons [2], edge differences, compression differences and motion vectors [3]–[5].

There are various possibilities for improving on the basic methods. The variety of basic methods opens up the possibility of combining several of them into a multiple expert framework, explored in [6]–[8]. Also, one can use an adaptive threshold setting, by using statistics of the dissimilarity measure within a sliding window [9]–[11].

B. Adaptive GOP Structure

Due to the need of flexibly delivering multimedia data to users with different available interests, access networks and resources, the efficiency of video encoding became very important task. The newest and the most efficient video coding standard is the H.264/AVC [12], [13]. New features of H.264 include motion estimation in variable block sizes, multiple reference frame motion compensation, spatial prediction for intra coding, small block size residual transform coding, adaptive and hierarchical block size transform, etc [12]. The encoders mostly used fixed group of pictures (GOP) size to encode video sequences. The GOP size can achieve different values, but once target size for GOP is selected, it is applied to whole coded sequence.

While fixed GOP structures are easy to implement, they prevent encoders from adapting to temporal variations in video sequences and thus prevent encoders from improving coding efficiency by selecting the frame type of each frame adaptively. The transitions between shots are the region, where static GOP structures achieved poor performance. Generally, if the video frames with smaller video content variance are coded as intra frames, we will waste a lot of bits in video coding. Conversely, if two shots changed and frames are coded using inter frames, it will also become inefficient.

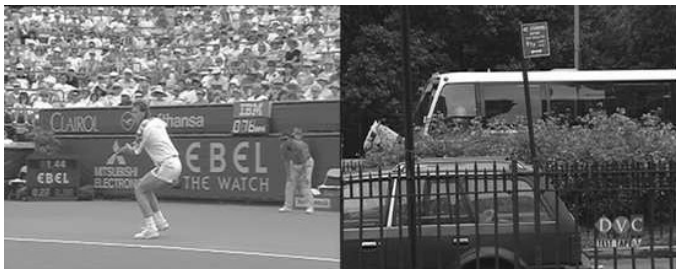

Fig. 1. An example of abrupt cut.

This can be solved by using adaptive GOP structure with positioning I frames to the places of shot changes. Adaptive GOP structure (AGS) [14], [15] is a new technique that can be used for enhancing the coding performance of the scalable extension of H.264/AVC. The AGS scheme adaptively changes the sizes of GOP structure according to the temporal characteristics of a video sequence to improve the coding efficiency.

In this paper we propose a novel method of GOP structure adaptable to the positions of shot transitions. This approach is based on shot cut detection and subsequently the size of GOP structure is adapted to the video content by placing I frames to detected abrupt cut. The proposed method was evaluated through experiments and obtained results were compared with selected sizes of fixed GOP structure.

The paper is structured as follows: in the second section a proposed method of adaptive GOP structure is described. Results obtained by the simulations for adaptive and fixed GOP structures are displayed in the third section. All results are summarized and discussed in the conclusion.

II. Proposed Method

The proposed approach is based on two principles – the detection of shot transitions and applying I frames to the positions of detected cuts.

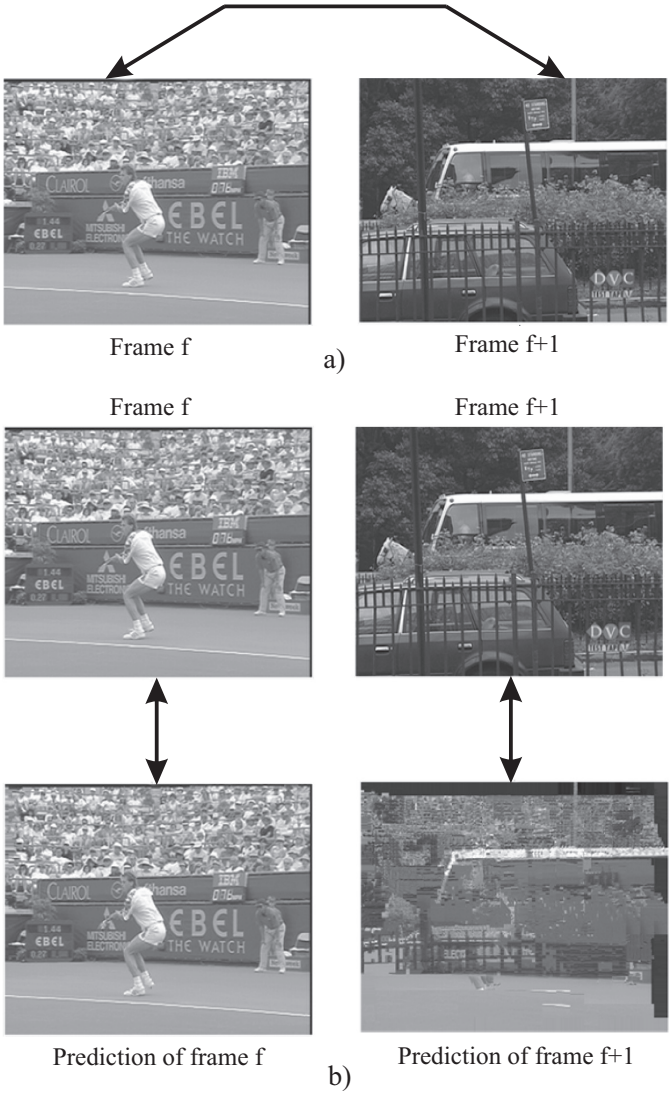

Fig. 2. The principle of comparison in shot cut detection process for the (a) majority of published methods and (b) proposed method.

A. Detection of Video Cuts

In general, abrupt transitions are much more common than gradual transitions, accounting for over 99% of all transitions found in video [2]. Therefore, we focus only on the detection of an abrupt cut.

The novelty of presented method is in the evaluation of the positions of abrupt cut. The most of existing methods calculate similarity of two consecutive frames by chosen metric and classify the frames as cut or non-cut based on the comparison of obtained similarity metric value and defined threshold. These approaches can achieve high detection accuracy, but suffer for high sensitivity to object or camera motion within shots. This leads to increased number of false detections and decreasing the detection accuracy.

Our proposed method compares the actual frame with its motion compensated prediction. It is based on assumption the prediction of frames within one shot would be very similar (or nearly identical) to the evaluated frame, while the prediction of frame in the place of shot should be very bad (different), because we predict the content of next shot from the content of previous shot. Therefore if we compare actual frame with its motion compensated prediction by selected metric, it should show high similarity for frames within one shot and high dissimilarity for cut frames.

Fig. 2 illustrates the differences among the proposed methods and existing ones. The arrows indicate which frames are compared. We can see that prediction within one shot is very similar to actual frame (tennis player) and the predicition is very poor in the place of shot. We can notice the legs and contour of tennis player in the prediction of bus. Thus, proposed method should be able to correctly distinguish cut and non-cut frames and should be more robust to object or camera motion within shot.

The other advantage of proposed method is that it can be performed directly during video encoding process without any delays or additional computational complexity, because the block of motion estimation is standard part of currently used video encoders.

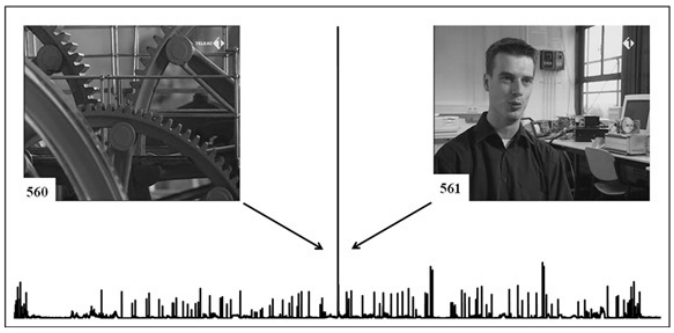

Fig. 3. An example of shot cut detection process.

The cuts are expected in local extremes for both proposed and existing methods as the similarity evaluation of cut frames by selected metrics would cause significant increase (for metrics with the zero value for identical images) or significant decrease (for metrics with the highest value for identical images – mainly correlation based metrics) of the metric value in comparison with previous and next frames. The example of shot cut detection process is shown in Fig. 3.

For evaluating the similarity of the frame and its prediction we have chosen Pearson correlation coefficient as a representative of correlation metrics. In statistics, the Pearson's correlation coefficient (sometimes also referred to as the Pearson product-moment correlation coefficient) has been widely employed to measure the correlation (or strength of linear dependence) between two variables X and Y [13]. The value for a Pearson correlation coefficient can fall between 0 (no correlation) and 1 (perfect correlation). Generally, correlations above 0.80 are considered as really high. Therefore the lowest values will be determined as cuts. The Pearson correlation coefficient for 2D signals like video sequences is expressed as follows [16]:

$$PCC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(f\left(i,j\right)-f^{m}\right)\left(f_{p}\left(i,j\right)-f_{p}^{m}\right)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(f\left(i,j\right)-f^{m}\right)^{2}\left(f_{p}\left(i,j\right)-f_{p}^{m}\right)^{2}}}, \quad (1)$$

where M and N stand for dimension of frames f and its prediction f_p. $f\left(i,j\right)$ and $f_p\left(i,j\right)$ express the pixel intensity for $(i,j)^{th}$ element of frames. f^m is the mean pixel intensity of the frame f and f_p^m is the mean pixel intensity of its prediction f_p.

The use of threshold is needed for automatic shot boundary detection. We can use fixed or adaptive threshold. For fixed threshold it is needed to select an appropriate value, otherwise the shot boundary detection would achieve poor results. The efficiency of shot cut detection algorithm can be evaluated by recall, precision and F1 score.

The recall measure, also known as the positive true function or sensitivity, corresponds to the ratio of correct experimental detections over the number of all true detections and it can be calculated as follows:

$$r = \frac{C}{C+M}, \quad (2)$$

where C represents correctly detected cuts and M missed detections.

The precision measure is defined as the ratio of correct experimental detections over the number of all experimental detections and it is computed as:

$$p = \frac{C}{C+F}, \quad (3)$$

where C stands for correctly detected cuts and F for false detections.

F1 score measure is a combined measure that results in high value if, and only if, both precision and recall result in high values. F1 score measure is computed like:

$$r = \frac{2.p.r}{p+r}. \quad (4)$$

B. Adaptive GOP Structure

If we encode two video sequences by the same way, it is not guaranteed we achieve the same compression ratio. It is caused by video content variation in different video sequences. Therefore it is desirable to adapt the video encoding process to the content of particular video sequence. Current video encoders enable this by adapting the size of GOP structure or type of frame to the video content. This approach is called adaptive GOP structure.

Our aim is to predict the position of I frames based on shot transitions with aim to improve video encoding efficiency without degrading the quality. We place I frames to the positions of abrupt cuts (first frame of each shot is encoded as I frame), what should provide good prediction of following frames within one shot and lower bitrate needed by P or B frames. Thus we adapt GOP size to the video content. The principle of proposed GOP structure is shown in Fig. 4.

According to positioning frame types between two consecutive I frames, P and B frames are most commonly used. We have decided to use only P frames for simulation, because B frames are computationally high demanding and can cause unwanted delays.

In case of fixed GOP structure, the type of encoding frame is not adapted to the video content. Fixed GOP structure selects P frames for encoding of shot transitions very often. This decision causes the degradation of encoding efficiency and increase the number of bits used for encoding, because P frame will contains mainly I macroblocks.

We assume the false detection would increase needed bitrate, but improve the video quality. In contrast if we have a lot of missed detection it would lead to degrading of video quality and increasing the needed bitrate due to high number of I macroblocks in P frame at the place of abrupt cut. Therefore shot cut detection algorithm, which is able to operate in real time with the highest possible accuracy, is necessary requirement for proposed GOP structure.

III. Experimental Results

We confirmed the effectiveness of proposed method through a test experiment. For test purposes we created a video sequence (1989 frames) at CIF resolution (352 x 288 pixels) with 7 abrupt cuts sampled at rate of 30 frames per second.

The test video sequence consists of eight standard test sequences: akyio, foreman, hall, flower, mobile, mother-daughter, stephan and bus.

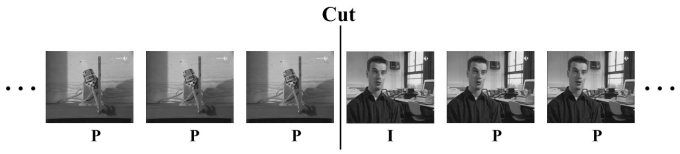

Fig. 4. The principle of proposed adaptive GOP structure.

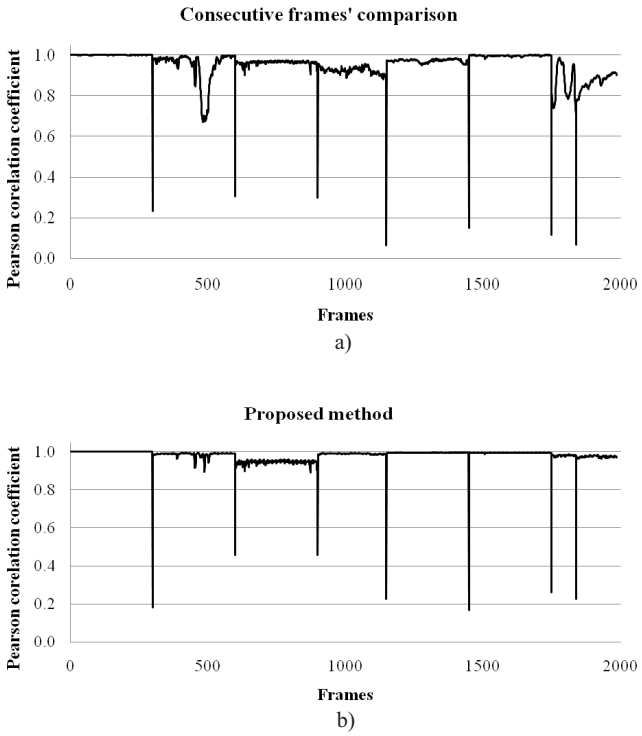

Fig. 5. Shot cut detection by Pearson correlation coefficient for (a) majority of published methods and (b) proposed method.

A. Detection of Video Cuts

The obtained results are compared with results of method using consecutive frames' comparison. For prediction of frames we have employed motion estimation scheme used in H.264 video encoding standard. The Pearson correlation coefficient is calculated for each component Y, U and V. The total value for YUV is computed as an average of components' values.

Fig. 5 shows results for abrupt cut detection. All shots were detected by both methods, they are represented by a significant decrease in the value of the Pearson correlation coefficient. Proposed method reached values higher than 0,8 for all non cuts frames. In addition the proposed method suppressed local extremes caused by huge motion in test sequences. This assures higher robustness to object or camera motion and decreases the probability of false detections.

To show the impact of threshold selection we ran simulation with the values of threshold from 0,001 to 1 with step 0,001. Every value under selected threshold is classified as shot cut. The dependency of F1 score is displayed on Fig. 6.

F1 score takes into account both the precision and recall measure. The proposed method holds the highest possible value (1) for about 50% of threshold's range in contrast to less

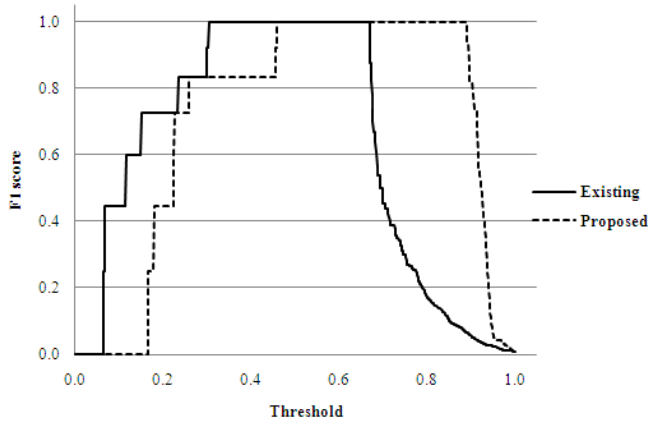

Fig. 6. F1 score for majority of published methods and proposed method.

TABLE I
THE RESULTS OF SHOT BOUNDARY DETECTION WITH PROPOSED ADAPTIVE THRESHOLDS

Threshold multiple	Method	Recall	Precision	F1 score
3	Existing	1	0.29167	0.45161
	Proposed	1	1	1
4	Existing	1	0.013807	0.027237
	Proposed	1	1	1

than 40% achieved by method based on the existing approach. It proves that the presented method is more stable according to threshold selection. However, too high or too low chosen value of threshold would cause decreasing of detection accuracy for both methods.

We have also proposed two versions of adaptive threshold to provide automatic shot boundary detection. The initial value of threshold is set to 0,5 and after detecting the first cut it is set to value of found cut multiplied three and four times respectively. Table I illustrates the results obtained for automatic shot boundary detection with proposed adaptive threshold measured by precision, recall and F1 score. Proposed method achieved value 1 for all measures for both version of adaptive threshold. The method based on existing frames' comparison reached F1 score 0,45161 for first version of adaptive threshold and 0,027237 for second one.

B. Adaptive GOP Structure

Obtained results for proposed GOP structure were compared with the situation, when whole sequence is coded only with one I frame at the beginning and the rest are P frames (fixed GOP structure IPPP with the length of entire video sequence). With effort to provide similar comparison as in [8], adaptive GOP structure was additionally compared with fixed GOP structures of size 4, 8, 12, 16, 32, 64 and 128 (it means I frame is followed by (size-1) P frames).

Each simulation was performed under following condition of H.264/AVC encoder:

- Profile: Main
- Total number of reference: 1
- Reference for P slices: 1

TABLE II
COMPARISON OF PSNR AND BITRATE ACHIEVED BY FIXED AND ADAPTIVE GOP STRUCTURE

GOP	PSNR [dB]	Bit rate [kbps]	Bit rate gain [%]
4	37.9	977.51	50.59
8	38.2	819.87	41.089
12	38.2	737.02	34.46
16	38.1	695.11	30.51
32	38.1	633.69	23.78
64	38.1	601.86	19.74
128	38.1	586.72	17.67
IPPP	38.1	570.50	15.33
Adaptive	38	483.03	

TABLE III
THE DIFFERENCE IN BIT USAGE FOR I AND P FRAMES AT THE PLACE OF CUTS

Cut	Fixed GOP IPPP, QP 26 P frame placed to position of cut			Adaptive GOP, QP 28 I frame in cut position	
	PSNR [dB]	Bits	Intra MB	PSNR [dB]	Bits
1	39	52328	396	39	49920
2	39.8	52000	395	39.9	48184
3	36.4	154496	396	36.6	145960
4	35.3	184496	390	35.3	183048
5	40.7	34256	396	40.5	33224
6	37	112464	396	37.1	112032
7	36.1	109928	385	36.2	109032

- Search range: 16
- Entropy coding method: CABAC

Our aim was to compare bitrates for different GOP structures under the same PSNR (or as close as possible). Thus for adaptive GOP structure and GOP of size 4 we have used QP 28 for both I and P frames and for the rest of simulated GOP structure is QP set to 28 for I frames and to 27 for P frames.

The obtained results are evaluated according to achieved final bitrate and video quality is evaluated by achieved value of PSNR.

Table II shows obtained results. The last column obtains the bit rate gain of proposed method in comparison to fixed GOP. Proposed method achieved a bit rate reduction from 15,33% to 50,59%, while providing the same (or nearly the same) PSNR. The highest bit rate reduction was achieved in comparison with fixed GOP structure of size 4. If we select small size of GOP structure, we force the encoder to use a lot of I frames and increase the bit rate. In the case of selected GOP with the size same as the length of entire video sequence (IPPP), we have forced encoder to use only one I frame for whole sequence. Despite this fact proposed method achieved 15,33% bit rate reduction. The bit rate of fixed GOP structure was increased due to large amount of I macroblocks used in P frames in positions of abrupt cuts. The reduction should become more significant for video sequences with more shot transitions.

Table III shows the difference in bit usage for I and P frames at the places of cuts for the same PSNR. With effort to provide as close PSNR as possible for adaptive and fixed GOP structure (of size of entire video sequence) in cuts' positions, we have used different quantization parameter for each GOP structure. The number of bits needed for encoding was higher if we used P frames in detected cuts for each cut. This was caused by a large number of used Intra macroblocks as it can be seen from third column for fixed GOP structure.

IV. CONCLUSION

In this paper we present new methods for abrupt cut detection and for adaptive GOP structure. The novelty of cut detection method is in use of different logic for frames' comparison. The majority of existing methods compares successive frames, our approach is to compare actual frame with its motion estimated prediction. We have chosen Pearson correlation coefficient for evaluating the similarity of compared frames. The novelty of method for improving coding efficiency of H.264/AVC by adaptive GOP is in adapting GOP structure to the video characteristics. Proposed method places I frames to the positions of detected abrupt cuts, the rest of sequence is encoded as P frames. Both methods were verified through test experiment and compared to commonly used methods.

The proposed method for video cut detection suppresses local extremes caused by motion activity, which are visible for existing methods, and could lead to false cut detections. We ran analyses for fixed threshold in the range of all values reachable by Pearson correlation coefficient. These analyses were evaluated by recall, precision and F1 score measures. The results show the proposed method is more stable and holds the maximal accuracy for more values of threshold. We also propose two versions of adaptive threshold. The proposed methods achieve significantly better results in comparison to commonly used technique.

We ran an experimental tests and proved efficiency of our approach in a comparison to fixed GOP structures of sizes 4, 8, 12, 16, 32, 64, 128 and length of the whole video sequence. Bit rate reduction obtained by method of adaptive GOP differs from 15,33% (fixed GOP structure with length of video sequence) to 50,59% (fixed GOP structure of size 4), while providing PSNR gain from 1,33% to 0,26% at the same time. In comparison to [17] we achieved higher bit rate reduction for the same GOP sizes 4, 8, 16 and 32. Furthermore our proposed method is not degrading PSNR.

The other advantage of using proposed adaptive GOP structure, in addition to improvement of video coding efficiency while providing the same quality, is the simplification of later video segmentation. There is no need to run shot cut detection process again, because cuts can be identified in positions of I frames.

For future work, we would like to examine the influence of various quantization parameters and of the different frames orderings in GOP structure on the video coding efficiency and video quality. Also it would be interesting to compare the efficiency of proposed methods with existing adaptive GOP structures implemented to available H.264 video encoders.

REFERENCES

[1] Z. Cernekova, "Temporal video segmentation and video summarization," Ph.D. dissertation, Comenius Univ., Bratislava, 2009.

[2] A. Amiri and M. Fathy, "Video shot boundary detection using qr decomposition and gaussian transition detection," *EURASIP Journal on Advances in Signal Processing*, 2009.

[3] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 90–105, 2002.

[4] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," in *Storage and Retrieval for Still Image and Video Databases IV*, 1996, pp. 170–179.

[5] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Storage and Retrieval for Still Image and Video Databases VII*, 1999, pp. 290–301.

[6] R. M. M. R. Naphade, "A high-performance shot boundary detection algorithm using multiple cues," in *Proc. IEEE Int. Conf. on Image Proc.*, vol. 2, 1998, pp. 884–887.

[7] C. Taskiran and E. J. Delp, "Video scene change detection using the generalized sequence trace," in *Proc. IEEE Int. Conf. on Image Proc.*, 1998, pp. 2961–2964.

[8] J. K. Y. Yusoff and W. Christmas, "Combining multiple experts for classifying shot changes in video sequences," in *Proc. 6th Int. Conf. on Multimedia Comp. and Systems (ICMCS)*, 1999, pp. 700–704.

[9] K. R. R. Dugad and N. Ahuja, "Robust video shot change detection," in *IEEE Workshop on Multimedia Signal Processing*, 1998.

[10] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 533–544, 1995.

[11] J. M. R. Zabih and K. Mai, "A feature-based algorithm for detecting and classifying production effects," *ACM Multimedia Systems*, vol. 7, no. 2, pp. 119–128, 1999.

[12] ITU-T, "Draft itut recommendation and final draft international standard of joint video specification (itu-t rec.h.264 —iso/iec 14496- 10 avc)," International Telecommunication Union, Tech. Rep. ITU-T Rec.H.264 —ISO/IEC 14496- 10 AVC, 2003.

[13] G. B. T. Wiegand, G. J. Sullivan and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[14] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[15] S. C. C. S. C. Hsia and C. L. Chen, "A real-time chip implementation for adaptive video coding control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 8, pp. 1098–1104, 2004.

[16] Y. K. Eugene and R. G. Johnston, "The ineffectiveness of the correlation coefficient for image comparisons," Los Alamos, Tech. Rep. LA-UR-96-2474, 1996.

[17] J. S. B. Zatt, M. Porto and S. Bampi, "Gop structure adaptive to the video content for efficient h.264/avc encoding," in *Proceedings of 2010 IEEE 17th International Conference on Image Processing*, SEPTEMBER 2010, pp. 3053–3056.