Chinese Journal of Electronics
Vol.18, No.4, Oct. 2009

A Novel Scene Cut Detection Method in H.264/AVC Compression Domain*

GAO Yu, ZHUO Li, WANG Suyu and SHEN Lansun

(*Signal and Information Processing Laboratory, Beijing University of Technology, Beijing 100124, China*)

Abstract — As an advanced video compression standard, H.264/AVC has been applied to various fields such as video surveillance, video conference, and wireless video communication. This paper presents a novel scene cut detection method in H.264/AVC baseline profile compression domain, which takes advantage of the available features from H.264/AVC bitstreams, including chroma prediction modes, motion vectors, macroblock types, and so on. Moreover, in this method, four new criterions used for scene cut detection have been proposed, i.e. the distribution difference of chroma prediction modes, the distribution difference of macroblock types, the accumulative motion amount, and the difference of motion vector angles. The thresholds of the criterions are mainly determined by the minimum error Bayesian decision. Experimental results show that the proposed method can detect the scene cuts at I-frames and P-frames correctly without the information of bi-directional prediction which is not available in H.264/AVC baseline profile.

Key words — Video, Scene change detection, H.264/AVC.

I. Introduction

Recently, more and more applications of H.264/AVC standard call for a set of new methods that can effectively organize, present, index, and search H.264/AVC bitstreams. Involved technologies include detection and segmentation of video scene changes, abstraction of key frames, video analysis and filtering, video indexing and retrieving, *etc.* Since a video scene is the elemental unit of video processing, video scene cut detection is the foundation of those technologies that are directly influenced by the result of detection.

In general, there are two kinds of scene change detection methods, *i.e.* pixel domain based and compression domain based method, where the latter one recognizes scene cuts from the compressed bitstreams without decoding or just with partial decoding. Compared to the method used in pixel domain, the compression domain based scene cut detection methods possess several advantages: less data to be analysed and stored, less time consumption, and easier abstraction of video features as well.

Some scene cut detection methods based on MPEG bitstreams have been proposed and already applied to some systems. The detection clues provided by MPEG bitstreams consist of DCT coefficients[1], motion vectors[2], macroblock prediction types[3], bit consumptions[4], and so on. However, the scene cut detection based on H.264/AVC standard, compared to MPEG, encounters many difficulties:

• Due to the intra prediction, the DCT coefficients of intra macroblocks cannot be directly abstracted. After transform and entropy coding of prediction residues, the computation of DCT coefficient reconstruction is nearly as complex as that of full decoding. Therefore, the classic method – MPEG DC picture fails to be applied to H.264/AVC bitstreams.

• H.264/AVC encoder adopts a prediction mode choosing strategy based on optimal rate-distortion function, and thus intra prediction is probable to be used in the smooth area of picture. According to that feature, the distribution of macroblock prediction types in H.264/AVC bitstreams is different from that in MPEG bitstreams, and hence the traditional methods using macroblock types and motion vector types of MPEG bitstreams can hardly applied to H.264/AVC bitstreams.

• When the scene change occurs, the bit-consumption variation of I-frames in H.264/AVC compression domain is less obvious than that in MPEG compression domain. Consequently, if the methods using bit rate in MPEG bitstreams are applied to the H.264/AVC bitstreams without any adjustment, the performance will be unacceptable.

• Since this paper focuses on H.264/AVC baseline bitstreams which have no B-frame, there is no information of temporal bi-directional correlation. Lack of such kind of information, it is impossible to detect scene cuts on I-frame if only using macroblock types and motion vector types. In mobile video application, in order to decrease coding complexity and storage space, a small number of reference pictures (often just one) are used. Also, in order to eliminate the influence of error propagation, the frequency of I-frame should be relatively high.

From above analysis, it can be seen that, it is quite challenging to detect the scene cuts directly in H.264/AVC bitstreams. In recent years, some researchers have begun to focus on this field. Liu *et al.*[5] counted the number of mac-

*Manuscript Received Oct. 2008; Accepted Dec. 2008. This work is supported by the National Natural Science Foundation of China (No.60772069) and the Beijing Novel Program (No.2005B08).

roblocks with different inter prediction modes as the features and exploited Hidden Markov Models to model different cases in which cuts occur, but this method which needs one forth data to train the Hidden Markov models reveals its inability in real time application. Zeng *et al.*[6] proposed a macroblock type analysis and the intra mode statistical constraint combined criterion to detect scene cuts at P-frame and B-frame, and used an intra mode histogram to measure the similarities among I-frames. But this method fails to detect the scene cuts at the boundary of two close-up frames, and causes false alarms by the sudden-appearance of new objects. Kim *et al.*[7] exploited intra prediction modes of macroblocks by partitioning a picture into several sub-blocks which makes the measure of similarity more accurate, and has the capabilities to eliminate interfere of object movement. Yet, this algorithm can only locate the scene cuts in the level of GOP (Group of Picture). In other words, it fails to tell us whether a frame is a cut or not.

Hong *et al.*[8] determined the cut-candidates using the percentage of intra macroblocks, and then generated an edge histogram of each candidate with eight prediction directions of intra prediction mode in H.264/AVC. However, it is difficult for the method to accurately recognize the scene cut at I-frames. Bruyne *et al.*[9] adopted a similar idea as Kim's to identify the scene cuts at I-frame and concluded that the scene cuts happened at P-frame or B-frame can be detected using the information of macroblock prediction types. Furthermore, Kuo *et al.*[10] compared the statistical distribution of intra prediction modes with an adaptive threshold to detect the shot change, and a prediction mode similarity for luma component is defined.

This paper presents a new scene cut detection method in H.264/AVC baseline profile compression domain. Four novel criterions used for scene cut detection have been proposed, and the thresholds of the criterions mentioned above are mainly determined by the minimum error Bayesian decision. Experimental results show that the proposed scene cut detection method can achieve high performance.

II. Scene Cut Detection Criterions in H.264/AVC Compression Domain

This paper mainly focuses on the scene cut detection for H.264/AVC baseline bitstreams where B-slice is not available. For I-frame (in this paper, we assume that a frame is encoded as one slice, so we refer to I-slice and P-slice coded pictures as I-frame and P-frame), four new criterions have been proposed to detect the scene cut; for P-frame, the percentage of intra macroblock has been adopted. The whole procedure of detection is illustrated in Fig.1.

As can be seen in Fig.1, the proposed scene cut detection method adopts a classic criterion the percentage of intra macroblock D_P[3,9] and other four new criterions to determine if a scene cut occurs, *i.e.* the distribution difference of chroma prediction modes D_1, the distribution difference of macroblock types D_2, the accumulative motion amount D_3, and the difference of motion vector angles D_4. The definitions and functions of D_1, D_2, D_3, and D_4 will be explained as follows.

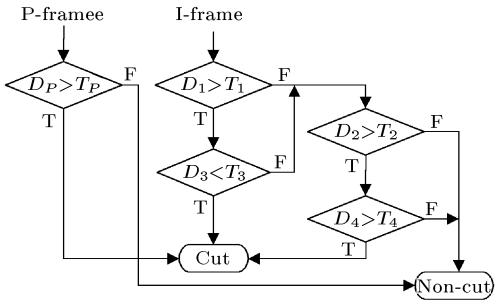


Fig. 1. Framework for proposed algorithm

1. D_1 - Distribution difference of chroma prediction modes

According to H.264/AVC standard, there are two kinds of intra prediction macroblock of luma samples: 16×16 block which has 4 prediction modes and 4×4 block which has 9 prediction modes, while intra prediction macroblock of chroma samples has only one kind of size: 8×8 block which has 4 prediction modes. Due to the informational distraction caused by the total 13 prediction modes of luma samples, this paper presents a method which takes intra prediction mode of chroma samples into consideration. Furthermore, chroma samples are less sensitive to the change of light than luma samples, which is important for the scene cut detection.

Consequently, we proposed the distribution difference of chroma prediction modes between previous I-frame and current I-frame – D_1, the quantitative difference of each chroma prediction modes in corresponding sub-block of neighbouring I-frames. D_1 can be acquired by the following steps: firstly, partition each I-frame into sub-blocks, and the sub-block S_k consists of $N_{MB}^{S_k}$ macroblocks; secondly, calculate the normalized results of quantitative differences divided by $N_{MB}^{S_k}$; finally, after summing all the normalized results, D_1 can be formulated as:

$$D_1(i) = \sum_{\forall k} \frac{1}{N_{MB}^{S_k}} \sum_{m=0}^{3} |NC_m^{i,S_k} - NC_m^{j,S_k}| \tag{1}$$

where i, j are the current I-frame, the previous I-frame respectively and NC_m is the quantity of certain intra chroma prediction mode m in a sub-block S_k. If there is no cuts at P-frames between two consecutive I-frames at all, and if D_1 is higher than the threshold (which will be discussed in Section III), the proposed method could preliminarily assert that a scene cut has been detected on the current I-frame.

2. D_2 - Distribution difference of macroblock types

Another proposed criterion for I-frame cut detection is the distribution difference of macroblock types between two P-frames closely adjacent to an I-frame – D_2, which does not partition a picture into sub-blocks. The type t in this criterion includes Skip, Inter 16×16, Inter 16×8, Inter 8×16, and Inter 8×8. And, D_2 is represented as follows:

$$D_2(i) = \frac{1}{N_{MB}} \sum_{\forall k} \sum_{t=0}^{4} |NT_t^{i+1,MB_k} - NT_t^{i-1,MB_k}| \tag{2}$$

where $i, i+1$, and $i-1$ are the current I-frame, the next P-frame to the current I-frame, and the previous P-frame to the current I-frame respectively. NT_t represents the quantity of certain macroblock type t in a frame. The total num-

ber of macroblocks N_{MB} is required to normalize the result. MB_k represents the k-th macroblock. If D_2 is higher than the threshold, the proposed algorithm could also preliminarily assert that a scene cut has been detected at the current I-frame.

3. D_3 - Accumulative motion amount

It will cause a problem if merely employing the criterion D_1. In general, higher threshold will lead to increasing of precision and decreasing of recall, while lower threshold will give rise to increasing of recall and decreasing of precision. This is a contradiction. If we set a high value as the threshold when the values of D_1 are relatively low due to the contents of some videos, a number of cuts on I-frames will be ignored. In order to increase precision and prevent recall from large decreasing, a new criterion is proposed, the accumulative motion amount – D_3, which is defined as the sum of average motion vector magnitudes of P-frame in a GOP:

$$D_3(i) = \sum_{F_{GOP}} \frac{1}{N_{MV}} \sum_{\forall k} \sqrt{(X^{i,MB_k})^2 + (Y^{i,MB_k})^2} \qquad (3)$$

where X, Y are the horizontal, vertical component of a motion vector respectively. The total number of inter macroblocks in a frame – N_{MV} is required to normalize the result. If D_3 is relatively low but D_1 is still high, we can declare that the current I-frame is a scene cut. In short, this criterion helps to lower the threshold of D_1, and then ensure to get an acceptable recall value.

4. D_4 – Difference of motion vector angles

Also, for D_2, there is a problem, *i.e.* when the current I-frame is not a cut and the two neighbouring P-frames are so similar to each other that the quantity of Skip macroblocks will increase dramatically, and will result in the higher value of D_2 as well. As a result, the current I-frame is probably to be falsely recognized as a scene cut if only using the criterion D_2.

In that case, because of motionlessness, a number of motion vectors equal to zero and the angles of the non-zero motion vectors change slightly. The angles of the non-zero motion vectors, however, are likely to be changed in the P-frame after a cut. Therefore, the difference of motion vector angles – D_4 is introduced. D_4 is defined as the average difference of non-zero motion vector angles between two P-frames closely adjacent to an I-frame just as follows:

$$D_4(i) = \frac{1}{N_{MV'}} \sum_{\forall k} |A^{i+1,MB_k} - A^{i-1,MB_k}| \qquad (4)$$

where A is the motion vector angle between the motion vector and the horizontal axis. The total number of macroblocks whose motion vectors are not equal to zero – $N_{MV'}$ is required to normalize the result. If D_4 and D_2 are both relatively high, we can declare the current I-frame is a cut. In short, this criterion helps to get rid of false alarms made by only using D_2.

III. Determination of Thresholds Based on the Minimum Error Bayesian Decision

The determination of those criterions' thresholds is the key to the performance of scene cut detection. In this paper, the minimum error Bayesian decision is adopted to determine the thresholds of D_P, D_1, D_2, and D_4.

Suppose that there are C classes – $w_1, \cdots, w_C$ and each has a prior probability – $p(w_1), \cdots, p(w_C)$. x is the vector of samples. The conditional probability density functions of class j, k – $p(x|w_j)$, $p(x|w_k)$ are known, and then the minimum error Bayesian decision is formulated as:

$$\begin{cases} p(x|w_j)p(w_j) > p(x|w_k)p(w_k), x \in w_j, \\ p(x|w_j)p(w_j) < p(x|w_k)p(w_k), x \in w_k, \end{cases} k = 1, \cdots, C; k \neq j \qquad (5)$$

For video scene cut detection, there are two classes ($C = 2$): the frames that are not scene cuts – w_1 and the frames that are scene cuts – w_2. In above formula, x is a sample of certain video feature. $p(w_1)$ and $p(w_2)$ are the probabilities of w_1 and w_2 respectively, and they can be estimated by the percentage of each kind of frames from test videos. $p(x|w_1)$ and $p(x|w_2)$ are the probability density functions of x respectively in the condition that the type (cut or non-cut) of the current frame is known, and they can be estimated by the histogram of x.

Fig.2 shows the curve of $p(x|w_1)p(w_1)$ and $p(x|w_2)p(w_2)$ when x represents the percentage of intra macroblock. In order to obtain the intersection, the two curves are the output results of Gaussian smoothing filtering. Since the probability of cuts $p(w_2)$ is relatively small and then $p(x|w_2)p(w_2)$ is small as well, logarithmic coordinate is utilized in Fig.2 for observing the intersection more clearly. According to Fig.2, the intersection is approximately at 0.8. Therefore, the estimated threshold of D_P is set as 0.8. The thresholds of D_1, D_2, and D_4 are all decided in the same way.

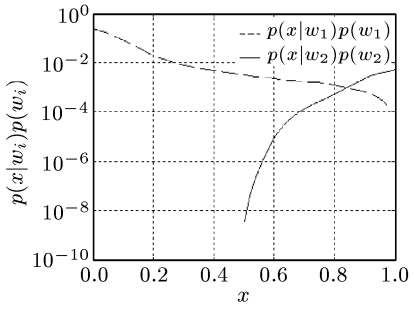


Fig. 2. Percentage of intra macroblock

The minimum error Bayesian decision makes effect to minimize the error probability. The Bayesian error probability marks the effectiveness of classification, and is defined as:

$$e_B = \min \sum_i \int [p(w_i)p(x|w_i)]dx \qquad (6)$$

Table 1 lists the estimated value of e_B and the determined thresholds. According to the table, we can find that being as the criterions for I-frame detection, the estimated values of e_B for D_1 and D_2 are lower than D_4, so D_1 and D_2 are used as the principal criterions for cut detection and D_4 as an auxiliary criterion. Moreover, due to lacking classification character, D_3 which equals to 18.0 in this paper is set by experiential evaluation.

Table 1. Estimated value of e_B and the determined thresholds

Criterions	Estimated value of e_B	Thresholds
D_1	0.1281	7.20
D_2	0.0800	0.92
D_3	none	18.0
D_4	0.1982	85.0
D_P	0.0073	0.80

IV. Experimental Results

1. Performance evaluation metric

In general, Precision and Recall are used as the metrics for the performance evaluation of scene cut detection. The two ratios are based on the number of correct, miss, and false detection which are represented as N_c, N_m, and N_f, and thus Precision and Recall are defined as follows:

$$Precision = \frac{N_c}{N_c + N_f} \times 100 \tag{7}$$

$$Recall = \frac{N_c}{N_c + N_m} \times 100 \tag{8}$$

Another performance evaluation metric which combines Precision and Recall together is $F1$ proposed by Qi[11]. Unless both Precision and Recall are relatively high, $F1$ cannot acquire a high value. The definition of $F1$ is as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

2. Experimental results

For the purpose of verifying the effectiveness of the proposed method, we use five types of video, *i.e.* sports, cartoon, news, movie, and "test" (a sequence joining standard test sequences together, including akiyo, coastguard, mobile, hall and so on), with QCIF (176×144) format for the experiment. The encoding mode is IPPP..., that is, the first frame is I frame, and the other 5 frames are encoded as P frame, and the bitrate is set as fixed 55kbps at 6 frames displayed per second. Table 2 shows the result of the proposed method which only adopts three criterions to perform scene cut detection: D_P, D_1, and D_2, while Table 3 shows the result of the improved method which adopts five criterions: D_P, D_1, D_2, D_3, and D_4. Comparing Table 2 with Table 3, we can find that, in the absence of D_3 and D_4 as the auxiliary criterions, recall is high but precision is comparatively low, after adding D_3 and D_4, however, precision and $F1$ is significantly improved while recall is nearly unchanged. The experimental results indicate that our proposed method can effectively apply to the scene cut detection.

Table 2. Performance with adopting three criterions: D_P, D_1, and D_2

Type of video	Number of cuts	N_c	N_m	N_f	Recall	Precision	$F1$
sports	15	13	2	25	86.7	34.2	49.1
cartoon	30	29	1	15	96.7	65.9	78.4
news	54	50	4	18	92.6	73.5	82.0
movie	55	53	2	55	96.4	49.1	65.1
test	200	198	2	23	99.0	89.6	94.1

Table 3. Performance with adopting five criterions: D_P, D_1, D_2, D_3, and D_4

Type of video	Number of cuts	N_c	N_m	N_f	Recall	Precision	$F1$
sports	15	12	3	2	80.0	85.7	82.6
cartoon	30	28	2	2	93.3	93.3	93.3
news	54	47	7	1	87.0	97.9	92.1
movie	55	53	2	4	96.4	93.0	94.7
test	200	196	4	23	98.0	89.5	93.6

Regarding the "test" video, the experimental results are very good no matter if adopting criterions D_3, D_4, because the scene changes of the "est" video which joins standard test sequences together are drastic, not as smooth as other types of video, so the correlation among scenes is low, which is in favor of detection.

V. Conclusion

Compared to the former methods of scene cut detection in H.264/AVC compression domain, the proposed method in this paper can achieve a good performance, especially applied to baseline bitstreams lacking B-frame as an important clue for scene cut detection. Therefore, the proposed method can be used in the abstraction of key frames, video analysis and filtering, video indexing and retrieving, *etc.* Otherwise, this paper is mainly wireless-communication-oriented and thus the test video is low in bit rate and frame rate. The complexity of this method is computationally less demanding than that of H.264/AVC decoding, for D_1, D_2, D_4, and D_P are computed without multiplication. Moreover, the algorithm delay of this method is just 1 frame, because the information of the next P-frame is needed to calculate D_2 and D_4 of current I-frame. Future work will focus on the high-quality video application such as wideband video communication and high definition field.

References

[1] B.L. Yeo and B. Liu, "Rapid scene analysis on compressed video", *IEEE Transactions on Circuits and System for Video Technology*, Vol.5, pp.533–544, 1995.

[2] A. Divakaran and H. Sun, "A descriptor for spatial distribution of motion activity for compressed video", in *Proc. SPIE Conference on Storage and Retrieval for Media Database*, Vol.3972, pp.392–398, 2000.

[3] S.C. Pei and Y.Z. Chou, "Efficient MPEG compressed video analysis using macroblock type information", *IEEE Trans. Multimedia*, Vol.1, pp.321–333, Dec. 1999.

[4] H. Li, G. Liu, Z. Zhang and Y. Li, "Adaptive scene-detection algorithm for VBR video stream", *IEEE Transactions on Multimedia*, Vol.6, pp.624–633, 2004.

[5] Y. Liu, W.Q. Wang, W. Gao and W. Zeng, "A novel compressed domain shot segmentation algorithm on H.264/AVC", in *International Conference on Image Processing*, Vol.4, pp.2235–2238, 2004.

[6] W. Zeng and W. Gao, "Shot change detection on H.264/AVC compressed video", *IEEE International Symposium on Circuits and Systems*, Vol.4, pp.3459–3462, 2005.

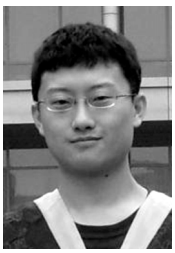
[7] S.M. Kim, J.W. Byun and G.S. Won, "A scene change detection in H.264/AVC compression domain", in *Advances in Multimedia Information Processing*, pp.1072–1082, 2005.

[8] B.Y. Hong, M.Y. Eom and Y.S. Choe, "Scene change detection using edge direction based on intra prediction mode in H.264/AVC compression domain", in *TENCON IEEE Region 10 Conference*, p.4, 2006.

[9] S. De Bruyne, D. Van Deursen, J. De Cock, W. De Neve, P. Lambert, and R.V. de Walle, "A compressed-domain approach for shot boundary detection on H.264/AVC bit streams", *Signal Processing: Image Communication*, Vol.23, pp.473–489, 2008.

[10] T.Y. Kuo and Y.C. Lo, "Detection of H.264 shot change using intra predicted direction", in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp.204–207, 2008.

[11] Y. Qi, A. Hauptmann and T. Liu, "Supervised classification for video shot segmentation", in *Proc. IEEE Conf. on Multimedia Expo* (*ICME*), Vol.2, pp.689–692, 2003.



GAO Yu was born in Beijing in 1983. He received M.S. degree in signal and information processing from Beijing University of Technology, China. His research interest is image and video coding. (Email: yugaobob@sohu.com)

ZHUO Li was born in Jiangsu Province, in 1971. She received Ph.D. degree in signal and information processing from Beijing University of Technology, China. Now, she is a professor of Beijing University of Technology. Her research interests are multimedia analysis, image and video coding, video communications.

WANG Suyu was born in Hebei Province in 1976. She received Ph.D. degree in signal and information processing from Beijing University of Technology, China. Now, she is a lecturer of Beijing University of Technology. Her research interests are super-resolution restoration from image/video sequences, intelligent video surveillance.

SHEN Lansun was born in Jiangsu Province, in 1938. He is a professor of Beijing University of Technology. His research interests are image and video processing, biomedical image processing system.