A NOVEL COMPRESSED DOMAIN SHOT SEGMENTATION ALGORITHM on H.264/AVC

Yang Liu[1]*, Weiqiang Wang*[2]*, Wen Gao*[1,2]*, and Wei Zeng*[1]

[1]Department of Computer Science, Harbin Institute of Technology, Harbin, 150001, China
[2] Graduate School of the Chinese Academy of Sciences, Beijing, 100080, China
liu_yang@hit.edu.cn, {wqwang, wgao , wzeng }@jdl.ac.cn

ABSTRACT

This paper presents a novel shot segmentation algorithm on the H.264/AVC video, which operates in the compressed domain. First, the algorithm exploits the intra prediction mode histogram to locate those potential GOPs, where shot transitions occur with great probability. Secondly, to further find shot boundaries at the frame level, we count the number of macroblocks with different inter prediction modes as the features and exploit HMMs to automatically model different cases in which shot transitions can occur among I, P and B frames. Since H.264/AVC provides more motion compensation modes, using HMMs can avoid the tediousness of manually tuning multiple thresholds simultaneously. The experimental results show that the algorithm is efficient and robust, and it can not only locate cuts, but also work for gradual shot transitions.

1. INTRODUCTION

The newest video compression standard H.264/AVC has been finalized. Due to its good compression performance, more video will be encoded with it in the future. Thus it is meaningful to study video analysis techniques in its compressed domain. Since shot boundary detection is the first step to high-level content analysis, many researchers have heavily focused on this problem in the past ten years and many approaches have been proposed. Zhang et al. [1] proposed a method based on dissimilarity between histograms, and two thresholds are used to detect abrupt and gradual shot transitions. A compressed domain approach [2], based on motion vectors of consecutive P frames, is further proposed to avoid mistaking camera motion as shot transitions. Yeo and Liu [3] directly extracted DC images in compressed domain and then their system identified shot transitions based on the difference of histograms between DC images. Pei et al. [4] used macroblock type information in compressed domain to analyze video. By means of analyzing and the pattern of MB types and counting the number of macroblocks with different patterns, the predefined thresholds were exploited to segment video into shots. Since H.264/AVC introduces many new encoding features compared with MPEG-1,2, the aforementioned compressed domain method cannot be directly applied to parse video encoded by the new standard. For example, intra prediction makes the method of extracting DC images infeasible. Hence the paper explores a new method to segment H.264/AVC video in compressed domain.

The paper is organized as follows. Section 2 presents the method employing intra prediction to locate those potential GOPs, where shot transitions occur with great probability. In Section 3 a scheme based on HMM is proposed to construct the models, which are used to further identify the exact shot boundaries. Section 4 describes the framework of the whole algorithm. Then the experimental results are presented in section 5. Section 6 concludes the paper.

2. LOCATE SHOT BOUNDARIES COARSELY USING INTRA PREDICTION

Compared with MPEG-1,2, intra prediction is a new technique adopted by H.264/AVC. It means, if a block or macroblock is encoded in intra mode, a prediction block must be formed based on previously encoded and reconstructed (but unfiltered) blocks. This prediction block should be subtracted from the current block before encoding. For the luminance component, the prediction block may be formed for each 4x4 subblock or for a 16x16 macroblock, denoted as Intra_4x4 and Intra_16x16 respectively. For Intra_4x4 type, Fig. 1 illustrates that the samples a-p are predicted by the samples A-M from eight directions. There are nine optional prediction modes for a 4x4 subblock and four optional prediction modes for a 16x16 luminance macroblock. More details can be found in [5]

Intuitively it is reasonable if two images are encoded into I frames and have very similar or the same content, they correspondingly have very similar or the same statistical distribution of the thirteen intra prediction modes related to each 4x4 subblock or 16x16 macroblock. Thus we calculate an intra prediction histogram with thirteen bins for the luma component of each I frame to

0-7803-8554-3/04/$20.00 ©2004 IEEE.

characterize its content, i.e. $h_k^s = \frac{r_k^s}{N}, k = 0,1,\cdots,12$, where N is the number of the subblocks when a frame is divided into 4x4 subblocks, r_k^s is the number of 4x4 subblocks with the k th intra prediction mode in the frame s. Note that for each Intra_16x16 macroblock 16 is voted into the corresponding bin. A distance is defined as follows

$$d(s_1, s_2) = \sum_{k=0}^{12} \left| h_k^{s_1} - h_k^{s_2} \right|$$

to measure dissimilarity of two consecutive I frames s_1 and s_2. If s_1 and s_2 belong to two shots, the distance $d(s_1, s_2)$ will be large. Therefore we can determine some potential GOPs containing shot boundaries with great probability by comparing $d(s_1, s_2)$ with a threshold.

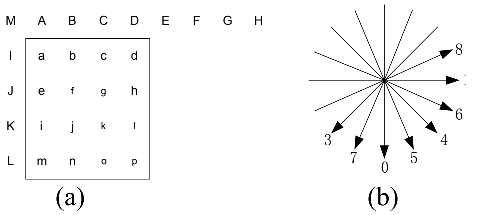


Figure 1. (a) samples a-p are predicted from samples A-M. (b) eight prediction direction. All for Intra_4x4.

3. CONSTRUCT HMM FOR SHOT BOUNDARY DETECTION

3.1.Feature selection

H.264/AVC supports more inter prediction modes. These modes include the following three important properties. First, the new standard offers more flexible macroblock partition modes. For example, the luma component of a macroblock can be divided into 16x16, 16x8, 8x16 and 8x8 blocks, in addition the 8x8 block can be further partitioned into 4x8, 8x4 and 4x4 blocks. Thus, the following statistics are all based on the minimum unit, 4x4 block. Secondly, improved “skipped” and “direct” motion inference methods are designed. The third property is multiple reference pictures (see Fig. 2). The first characteristic provides more accurate motion compensation and the latter two characteristics are heavily influenced by scene change. More related details can be obtained in [5] and [6].

When a shot change occurs, the information will exhibit different patterns, i.e. few blocks will be encoded with “skipped” or “direct” mode, simultaneously the number of blocks with multiple reference pictures will small. This characteristic helps us to detect shot change.

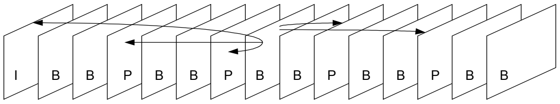


Figure 2. GOP structure and multiple reference pictures. The fifth B frame has multiple reference pictures.

In what follows, we will use HMM to segment video. Two reasons motivate us to adopt the model. One is that determining thresholds is a tedious thing, especially, when multi dimensional information are available. The second one is that HMM is suitable to recognize sequence patterns [7]. Thus, we handle the tedious work of determining thresholds by HMM.

3.2. Word structure

In practice, a video sequence is composed of GOPs, which consist of one I frame and a few P and B frames. The length of the GOP generated by our system is 15 and there are two B frames between two consecutive reference frames (P or I frames).

For detecting scene changes in video through HMM, we define a structure, referred as **word structure**. As Fig. 3 illustrates, the word structure is composed of the current P frame and the B frames between the preceding P (or I) frame and succeeding P (or I) frame. The word structure is regarded as a sequence. In this paper, the lengths of states sequence and observation sequence are both 5. The advantage of the word structure is that P frame and B frames are simultaneously taken into account.

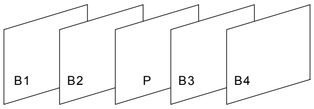


Figure 3. Word structure defined in this paper.

3.3. Building HMMs

- Define words for different patterns: For abrupt scene change, six words are defined according to the shot boundary position. The words are (100000), (010000), (001000), (000100), (000010) and (000001) respectively. For the pattern of no boundary in the structure, word (000000) is defined, and for long shot transition (111111) is defined. In these words, 1 represents the position where the scene change occurs. For example, (010000) represents the shot boundary locates between B1 and B2, and so on. For each word in the word sets, we build an HMM λ_i for it, which maximize the likelihood of observation vectors for the ith word.
- Observation sequence: The observation sequence of $(v_1 v_2 v_3 v_4 v_5)$ is composed of 5 feature vectors. Each

feature vector corresponds to a frame and contains 7 features, including the numbers of 4x4 blocks with forward, backward, and bidirection prediction respectively, the numbers of 4x4 blocks with "skipped" and "direct" modes, and the numbers of 4x4 blocks with forward and backward multiple reference pictures.

- Determine the number of states for every HMM: To learn the parameters of HMMs, we first decide the number of states of each model. Table 1 summarizes the number of states in different word models. In these HMMs, the models with the maximum number of states, four states, correspond to the words (000010) and (010000).

Table 1. The number of states in each model for different words (SN denotes the number of states).

Word	000001	000010	000100	001000
SN	3	4	3	3
Word	010000	100000	000000	111111
SN	4	3	2	2

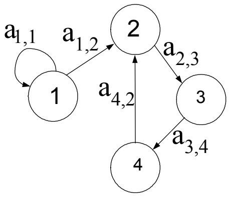
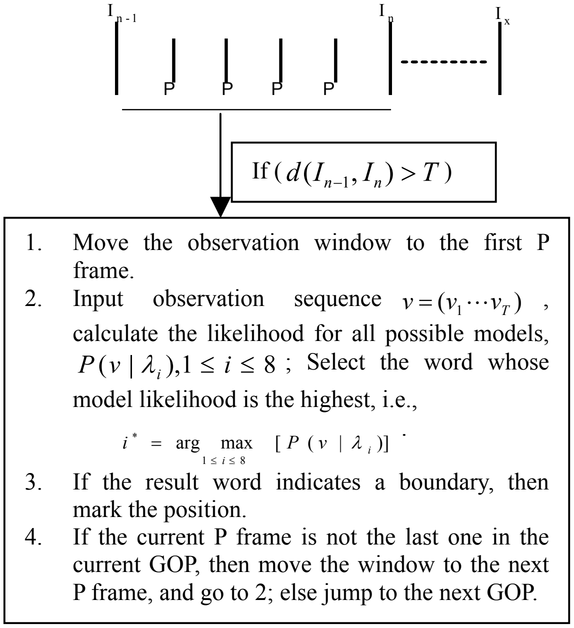


Figure 4. States transition graph of the HMM for the word (000010).

In Fig. 4, we illustrate the states transition relationship of the HMM for the word (000010), which has four states. The word (000010) represents that a scene transition occurs between B3 and B4 (Fig. 3). Four states are involved in the model. Intuitively the state 1 is used to characterize the motion compensation behaviour in B1 and frame B2 respectively. While the state 2 is used for P frame. Two states are specified to differentiate the frame B3 and B4, since they belong to different shots. Therefore most blocks in B3 are forward prediction and few blocks are backward or bidirection predictions, while most blocks in B4 are backward prediction and few blocks are forward or bidirection predictions. Simultaneously, they both have few blocks of direct mode, and the number of blocks with backward long-term in B3 and that of forward long-term reference blocks in B4 are both few.

4. FRAMEWORK for SHOT BOUNDARY DETECTION

In this section we describe the framework for detecting scene change, which integrates the two proposed Algorithms. The framework is shown in Fig. 5.



1. Move the observation window to the first P frame.
2. Input observation sequence $v = (v_1 \cdots v_T)$, calculate the likelihood for all possible models, $P(v \mid \lambda_i), 1 \le i \le 8$; Select the word whose model likelihood is the highest, i.e.,
$$i^* = \arg \max_{1 \le i \le 8} [P(v \mid \lambda_i)] \; .$$
3. If the result word indicates a boundary, then mark the position.
4. If the current P frame is not the last one in the current GOP, then move the window to the next P frame, and go to 2; else jump to the next GOP.

Figure 5. Framework for shot change detection.

Only if the dissimilarity distance of the two adjacent I frames is larger than a threshold T , the system will come into the potential GOP structure to find the shot boundary at frame levle. Otherwise next GOP will be coarsely detected.

5. EXPERMENTS AND ANALYSIS

Our algorithm has been tested on real video sequence. This test set is composed of two sequences, which are both coded by H.264 reference software JM7.3. The Spanish daily news sequence in the 17th CD of MPEG-7 is CIF size and consists of 10017 frames. The advertisement sequence acquired from CCTV is 720x576 size and composed of 2997 frames. The HMMs are trained are on one fourth data.

5.1 Only HMMs are used

To test the performance of shot boundary detection by HMMs, we set the threshold $T = 0$ to eliminate the effect of two I frames' dissimilarity measured by intra prediction modes. The left side of table 2 shows the results. From the results we can find that when only using HMMs to detect shot boundary, the recall is high. Five and three frames within shots are detected as shot boundary in the news and the advertisement sequences respectively. This is because inconsistent movement results in more blocks predicted by backward mode for B2 (or B4), while more blocks with forward mode for B1 (or B3). Therefore it misleads to find a boundary.

5.2 Adding intra prediction information

The second experiment uses dissimilarity distance between two adjacent I frames to filter the GOPs with low probability of shot change. Fig. 6 illustrates the dissimilarity distances of I frames characterized by d. The results extracted from news sequence. 16 of the 78 distance values are larger than 0.3, in which there are 6 cut shot changes and 4 dissolve shot transitions. The right side of table 2 shows the detection results by adding two I frames' dissimilarity. The thresholds of dissimilarity distance are both 0.3 for the two sequences.

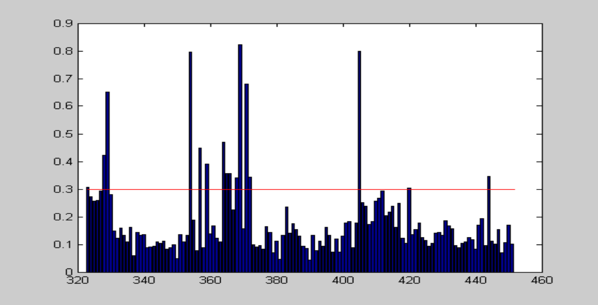


Figure 6. The dissimilarity distances of the news sequence, from I frame323 to I frame 451.

Table 2. Test results with different thresholds.

Shot boundary type	T=0		T=0.3	
	Advertise-ment	News	Advertis-e ment	News
Cut	69	48	69	48
Detected	67	48	67	48
False alarm	3	5	1	1
Dissolve	4	9	4	9
Detected	2	7	2	6
False alarm	2	2	0	0

When employing dissimilarity distance measured by intra prediction mode to select potential GOPs to further detect, the framework retains the recall and reduces the false alarm dramatically.

Table 3 gives out the total numbers of GOPs in the two sequences and the numbers of potential GOPs for further searching. The results indicate that the coarse detection by intra prediction histogram not only can improve the performance of the framework, but also can speed up shot change detection.

Table 3. The numbers of total GOPs and potential GOPs.

T=0.3 Sequence	Total GOPs	Potential GOPs
News	667	78
Advertisement	199	112

In addition, our method finds two of the three wipe type shot transitions in the news sequence. Meanwhile, none of twelve flashes is detected as shot boundary, because none of them is in the potential GOPs.

6. CONCLUSION

A framework is proposed to detect shot changes in H.264/AVC compressed domain. First, we exploit the thirteen intra prediction modes for luma component to measure the dissimilarity of two adjacent I frames. Second, we utilize more inter prediction modes and HMMs to segment video instead of manually tuning thresholds. Experimental results show that framework performs well on real video and achieves fast detection since only the potential GOPs are examined.

Acknowledgement

This work is supported by the Bairen Project of Chinese Academy of Science and the NEC sport video analysis project.

7. REFERENCES

[1] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst*., Vol. 1, July 1993. pp, 10-28.

[2] H.J. Zhang, C.Y. Low, and S.W. Smoliar, "Video parsing and browsing using compressed data," *Multimedia Tool Applicat.,* Vol. 1, no. 1, Mar. 1995, pp. 89-111.

[3] B.-L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. Circuits Syst*. Video Tech, vol. 5, Dec.1995, pp. 533-544.

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

[5] Thomas Wiegand, Gary J. Sulliva, et al, "Overview of the H.264/AVC Video Coding Standard," *IEEE Trans. On Circuits Syst. Video Tech*, Vol. 13, July, 2003, pp. 560-576.

[6] Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 /ISO/IEC 14 496-10AVC) Joint Video Team (JVT), Mar. 2003, Doc. JVT-G050.

[7] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Processing of the IEEE*, Vol. 77, no. 2, Feb. 1989. pp. 257-286.