A Fast Mode Decision Algorithm in Spatial and Temporal Scalable Video Coding

YANG Dawei
Harbin Engineering University
yangdawei@hrbeu.edu.cn

Baochun Hou
University of Hertfordshire
b.hou@herts.ac.uk

Reza Sotudeh
University of Hertfordshire
r.sotudeh@herts.ac.uk

Abstract

To decrease the computational complexity of adaptive inter-layer prediction and improve the encoding efficiency in spatial and temporal scalable video coding, a mode decision algorithm is proposed by exploiting the part of used modes in the co-located reference macroblock(s) for Hierarchical-B pictures of all layers. This scheme generates a reduced dynamic candidate list for the current macroblock according to the statistical information derived from the co-located reference macroblocks in their present temporal level. The experimental results show that this algorithm compresses approximately 29% in encoding time with the negligible loss in PSNR and slightly increasing in bit rate.

1. Introduction

Scalable video coding (SVC) has been finalized by the Joint Video Team (JVT) as an extension of H.264/AVC at March 2007 [1]. It inherits the original advantages from the previous video standard. Moreover, SVC introduces the scalable mechanisms of spatial (resolution), temporal (frame rate) and fidelity (quality) to fulfill the diverse requirement of different end-devices and flexible visual content adaptation in multimedia communications.

The block based encoding tool is used for inter-prediction, intra-prediction and quantization in SVC [2]. The unique block size is adopted after the encoder compares all the results in an exhaustive mode searching process. Flexible and smaller block size can encode macroblocks more accurate and effective; especially there are a large number of details in the video sequences. Usually, smaller block size leads to fewer residuals after motion prediction while introducing more motion vectors. As specified in H.264/AVC, rate distortion optimization enables a balance between the smaller residuals and the less accessory data to select a most suitable mode for each macroblock according to the rate distortion (RD) cost. This operation is an exhaustive search process with traversing every candidate mode to find the best one (minimum RD cost) for the current macroblock. It consumes more than 50% encoding time on rate distortion optimization in motion estimation [3].

In order to exploit the sample data from the different layers further, SVC adopts the tool of adaptive inter-layer prediction when multiple layers present. However, every new technology is a double-edged sword. To reduce the extremely encoding time becomes the main motivation of this fast mode decision proposed algorithm.

2. Investigation of macroblock correlation

2.1. Adaptive Inter-layer Prediction

The adaptive inter-layer prediction is a tool used for enhancement layers with applying motion estimation from the frames not only in temporal layer, but also in their base layer. And in high resolution and slow objects movement sequences, the mode decision of adaptive inter-layer prediction will obtain better encoding efficiency than original exhaustive search scheme.

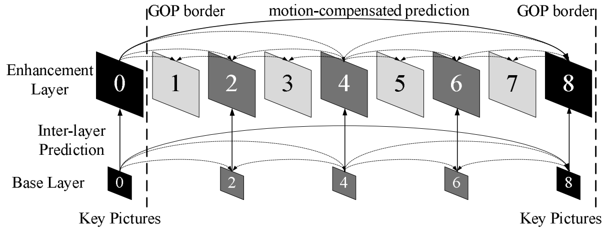

Figure 1. **The structure of Hierarchical-B**

An example of spatial and temporal scalable video coding is illustrated in Figure 1, where the size of GOP (group of picture) is 8. The 0th and 8th are key pictures, all the other frames are Hierarchical-B pictures. The lower layer is base layer with 15Hz frame rate and the higher one is enhancement layer with

978-0-7695-3859-4/09 $26.00 © 2009 IEEE
DOI 10.1109/IITA.2009.224

30Hz frame rate. It is obvious that the 2nd, 4th and 6th pictures of enhancement layer have their corresponding lower pictures in base layer, and they can utilize the adaptive inter-layer prediction tool to choose the appropriate macroblock from both base layer and enhancement layer motion estimation, but the other Hierarchical-B pictures of enhancement layer have to run the bi-directional prediction only in their temporal layer.

2.2. Macroblock Correlation

With the intensive experiment results, we found that the best mode of the current macroblock has a strong relationship with the used modes of the co-located reference macroblocks. Those used reference modes for motion prediction in reference frames compose the reference information for the current macroblock. The number of the lists is equal to the amount of reference frames of the current macroblock.

In order to obtain the correlation relationship between the current macroblock and the co-located macroblocks, here, we assume that $List(m_1(cost_1), m_2(cost_2), \ldots, m_N(cost_N))$ is the macroblock dynamic reference list, where $m_N(cost_N)$ is one of used macroblock mode with $cost_N$. N is a mode selection parameter, which is positive integer and $N \in [1,7]$, including seven macroblock modes: {MODE_16×16, MODE_16×8, MODE_8×16, MODE_8×8, Intra_4×4, Intra_16×16 and Intra_BL}. $ListSub(m_1(cost_1), m_2(cost_2), \ldots, m_{SN}(cost_{SN}))$ is for the sub-macroblock dynamic list, SN is positive integer and $SN \in [1,4]$, which including four macroblock modes: {8×8, 8×4, 4×8 and 4×4}, which is valid only when the MODE_8×8 is selected in *List*. These dynamic lists are sorted into *List'* and *ListSub'* by RD cost respectively when all the possible modes of each macroblock have been calculated in motion estimation. For one encoded macroblock, $m_1(cost_1)$ has always the lowest RD cost value and $m_N(cost_N)$ or $m_{SN}(cost_{SN})$ should be the largest RD cost value in *List'* or *ListSub'* after the sort order operation.

In order to restrict the selection of the sub-macroblocks, the proportional equation is given in (1):

$$SN = \lfloor \alpha \cdot N + \beta \rfloor \qquad (1)$$

where α and β are the influence factors which adjust the value proportion between N and SN, and $\lfloor \cdot \rfloor$ operator indicates round down to the nearest integer.

It is an actually exhaustive search calculation by comparing the RD cost of macroblocks one by one according to the rate distortion optimization function. However, it is feasible that there is a potential prediction method to get the current best mode before running the motion estimation. Figure 2 gives the ratio of the actual mode computation quantity over the theoretic computation quantity in motion estimation with different N values; the percentage equation is calculated in (2), (3) and (4):

$$Percentage = \frac{\sum_D \left(mode_n \times E(mode_n)\right)}{\sum_D mode_n} \times 100\% \qquad (2)$$

$$D = Q \times \left(Num_{Frame} - 2\right) \qquad (3)$$

$$E(mode_n) = \begin{cases} 1, & \text{if } mode_n \text{ is in } List' \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

where $mode_n$ is the best mode for macroblock, $n \in [0,D]$; D and Q are constant values; Q is equal to 99 for QCIF, 396 for CIF and 1584 for 4CIF picture; Num_{Frame} is the number of encoding frames; The dynamic *List' are* a truncated list by selecting a fixed number *N*. Given a specified values *N*, the *List'* reserves the first *N* modes from the sorted list and the sub-macroblock dynamic list *ListSub'* is still satisfied with (1). With the increasing of the selection parameter *N* in Figure 2, the ratio is approaching 100%. When the selected *N* is bigger than 4, which means getting four used macroblock modes from each reference macroblock, the percentage is nearly approaching above 90%.

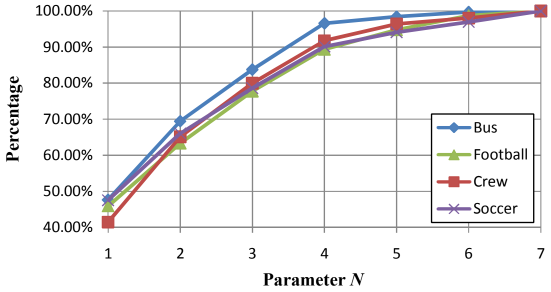

Figure 2. **The probability of obtaining the best modes for all macroblocks in Hierarchical-B pictures of 4 sequences with different *N*.**

3. Fast mode decision algorithm

The target of this algorithm is proposed to decreasing the encoding time with adaptive inter-layer prediction. According to the description above, the detailed of this algorithm will be described here.

3.1. Criterion for parameter *N*

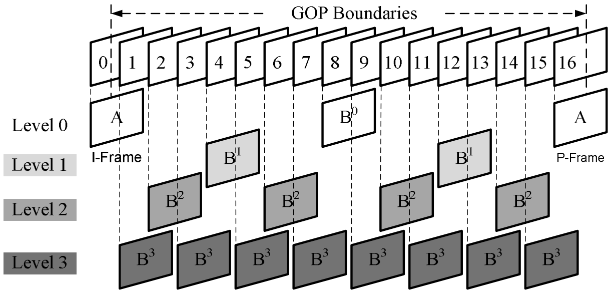

Figure 3. **Temporal levels in a GOP of size 16**

The temporal level structure of Hierarchical-B pictures is shown in Figure 3, where the GOP size is 16 and the temporal level resolution is 4. The maximum value of temporal level is satisfied with (5):

$$T_{MAX} = \log_2(GOPSize) \qquad (5)$$

where GOPSize is the size of GOP.

In temporal level 0, only one picture B^0 presents, whose index number is 8. There are 2 pictures in level 1, 4 and 8 pictures in level 2 and 3, respectively.

Parameter N is defined different values in different levels to improving the encoding time and decreasing the candidate modes for the current macroblock. In order to get the appropriate parameter N, the average PSNR of the frames in same temporal levels has been considered. It is defined as (6):

$$\overline{PSNR(T_0)} = \sum\left(PSNR(T_0)\right)\Big/M \qquad (6)$$

where T_0 is the index number of temporal levels, whose maximum value is equal to $T_{MAX}-1$; M is the number of frames with same temporal index; $\overline{PSNR(T_0)}$ is the mean value of PSNR in T_0 temporal level. Furthermore, we use two similar procedures to adjust the parameter N. One is from the original motion estimation, the other one is from the iteration with N. Equation (7) gives the criterion to choose the corresponding parameter N:

$$\left|\overline{PSNR_0(T_0)} - \overline{PSNR(T_0)}\right| < \varepsilon. \qquad (7)$$

where ε is an arbitrarily small positive integer; $\overline{PSNR_0(T_0)}$ is the mean value of PSNR in temporal level T_0 with N; $\overline{PSNR(T_0)}$ is the mean value of PSNR with the same temporal level T_0 with N in exhaustive search mode.

3.2. Initializing parameter *N*

The flowchart of initializing parameter N is illustrated in Figure 4(a). We exploit the first GOP of a sequence to find the appropriate value for N by iteration until (7) is satisfied. First, the original result generated is obtained by reference model JSVM 9.1 [4]. Then, let $N_i = 1$, $i \in [1,6]$, $N_i \in [1,7]$, that gives all temporal levels in all layers with the same value. Compare these two results and decide whether continue to search next value for parameter N. If it is not satisfied (7), $N_i = N_i + 1$, return to the next comparison.

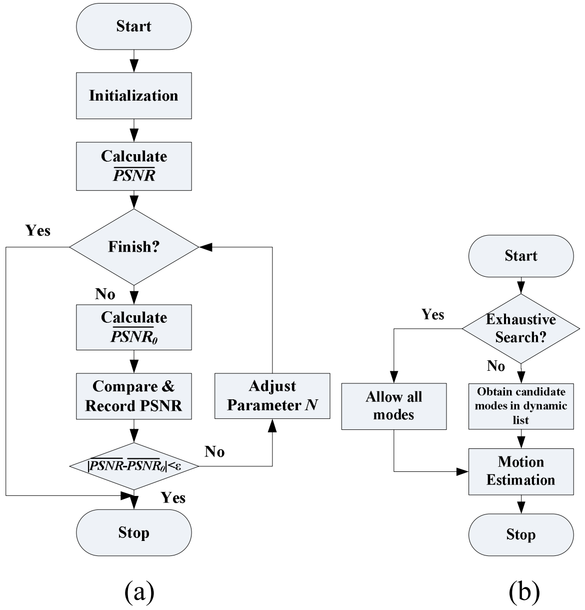

Figure 4. **The flowchart of proposed algorithm (a) Flowchart of initializing parameter *N* (b) Flowchart of mode decision**

Table I gives the initializing parameter N after calculated eight standard sequences in one GOP at $\varepsilon = 0.03$, where all test sequences use GOP size of 16 to run the simulation.

Table 1. **Parameter *N* in Different Temproal Levels with Different Sequences (ε = 0.03)**

	Temporal Level			
Sequence	***0***	***1***	***2***	***3***
Bus	4	6	3	6
Football	4	4	2	6
Foreman	5	5	5	4
Mobile	6	5	3	5
City	4	3	2	2
Crew	4	5	6	4
Harbour	4	4	4	4
Soccer	4	3	4	2

3.3. Initializing Mode Decision

Figure 4(b) is the flowchart of mode decision in proposed fast algorithm. The dynamic lists, *List'* and *ListSub'* as a reference group are set up for all macroblocks of Hierarchical-B pictures and their values should be hold until all of the frames are encoded in GOP. The so-called "co-located" reference macroblock is derived by the predictor of JSVM reference model before motion prediction.

4. Simulation results

The proposed fast mode decision algorithm is implemented in JSVM 9.1 encoder and all the tests below are applied this proposed algorithm in all layers, including base layer and enhancement layer(s). The test platform used is Intel Pentium IV, 3.0 GHz CPU, 1G RAM with Windows XP professional operating system. The simulation parameters in configuration files are illustrated in [5]. The spatial resolution of the first four sequences is 352×288 luma samples per picture (CIF) with 2 layers, while the last four sequences have a spatial resolution of 704×576 luma samples per picture (4CIF) with 3 layers. QP is set to 28 for base layer, 32 and 38 for enhancement layers.

In order to evaluate the proposed algorithm, three parameters including encoding time reduction rate, variation of PSNR and bit rate increase rate were defined in (9), (10) and (11):

$$\Delta Time = (Time_B\text{-}Time_A)/Time_A \times 100\% \tag{9}$$

$$\Delta PSNR = PSNR_B - PSNR_A \tag{10}$$

$$\Delta Bitrate = (Bitrate_{\mathrm{B}} - Bitrate_{\mathrm{A}})\ /\ Bitrate_{\mathrm{A}} \cdot 100\% \tag{11}$$

where subscript *A* indicates the result under the original exhaustive mode decision algorithm specified in JSVM 9.1, and subscript *B* is the result of proposed fast mode decision algorithm.

Table 2. **Experimental results**

Sequence	ΔBitrate /(%)	ΔPSNR /(dB)	ΔTime /(%)
Bus	0.50%	-0.0194	-13.98%
Football	2.43%	0.0117	-39.88%
Foreman	0.13%	-0.0215	-16.28%
Mobile	0.16%	-0.0191	-19.03%
City	0.06%	-0.0082	-54.66%
Crew	2.77%	-0.0275	-23.75%
Harbour	0.23%	-0.0132	-24.82%
Soccer	0.75%	-0.0174	-43.27%
Avg.	0.88%	-0.0143	-29.46%

Table 2 demonstrates the coding results of JSVM original algorithm and our proposed algorithm. Only the enhancement layer Y-PSNR and bit rate results are shown in the table. The results show that the proposed algorithm is very efficient in reducing the encoding time, which is about 29.46% on average of original JSVM original scheme. More time is saved with the number of sequence layers increasing.

Figure 5 and 6 present the rate distortion curves with two different sequences: Bus and Soccer. From these figures, we can see that there is little difference between the two rate distortion curves in each figure so that this proposed algorithm does not degrade the encoding efficiency and picture quality.

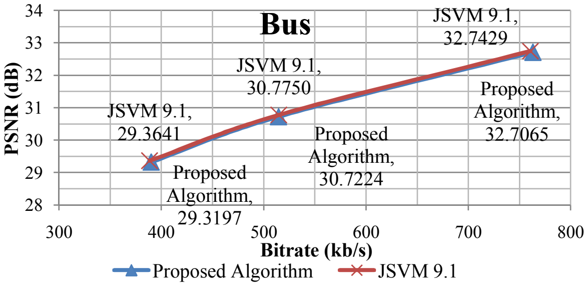

Figure 5. **RD curve of Sequence Bus**

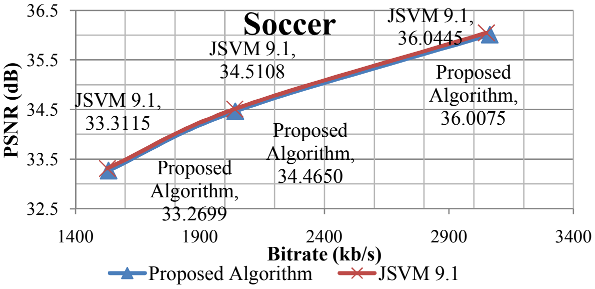

Figure 6. **RD curve of Sequence Soccer**

5. Conclusions

In this paper, a fast mode decision algorithm for Hierarchical-B pictures within spatial and temporal SVC has been proposed in order to decrease the computational complexity and shorten the loss on quality within SVC sequences. The algorithm can reduce the computational time nearly 29% with slight loss of PSNR and bit rate increase.

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1103-1129, 2007.

[2] Z. Bin, H. Baochun, and R. Sotudeh, "A Fast Intra/Inter Mode Decision Algorithm of H.264/AVC for Real-Time Applications," *IEEE International Conference on Communications*, 2008, pp. 510-514.

[3] Y. Liang, Z. He, and I. Ahmad, "Analysis and design of power constrained video encoder," *IEEE 6th CAS Symposium on Emerging Technologies*, Shanghai, China, 2004, pp. 57-60.

[4] J. Reichel, H. Schwarz, and M. Wien, "Joint Scalable Video Model Algorithmic Text Description," *JVT-X202*, Geneva, Switzerland, July, 2007.

[5] M. Wien, "Testing Conditions for Coding Efficiency and JSVM Performance Evaluation," *JVT-P205*, Poznan, Poland, July, 2005.