A Fast Mode Decision Algorithm of Adaptive Inter-layer Prediction in Scalable Video Coding

YANG Dawei[1], Baochun Hou[2], ZHAO Chunhui[1]

[1]College of Information and Communication,
Harbin Engineering University,
Harbin, P.R. China
{yangdawei, zhaochunhui}@hrbeu.edu.cn

[2]School of Electronic, Communication and electrical Engineering,
University of Hertfordshire,
Hatfield, United Kingdom
b.hou@herts.ac.uk

***Abstract*—A fast mode decision algorithm is proposed to solve the computation complexity for enhancement Hierarchical-B pictures with adaptive inter-layer prediction in H.264/AVC scalable extension, scalable video coding. This method exploits the evaluated modes of the co-located reference macroblocks, which include the best and other abandoned modes, to compose a new candidate group for the current macroblock in adaptive inter-layer prediction of motion estimation according to their RD cost values. The candidate modes can be reduced by adjusting the size of the group in order to decrease the calculation. Experimental results show that the proposed algorithm reduces 35.49% on average with very limited decrease of encoding efficiency in terms of PSNR and bit rate.**

***Index Terms*—scalable video coding, fast mode decision, Hierarchical-B, inter-layer prediction**

I. INTRODUCTION

Scalable video coding (SVC) is the scalable extension of the advanced video coding (AVC) standard H.264 and developed by Joint Video Team (JVT) of the ITU-T VCEG and the ISO/IEC MPEG [1]. SVC combines temporal (frame rate), spatial (resolution) and fidelity (quality) scalabilities with one single bit stream to support the diverse devices of different display sizes and network bandwidth in multimedia communications.

SVC introduced a number of advanced features and technologies in terms of the encoding efficiency and robustness. For spatial scalability, the base layer is a reduced down-sampling resolution version of the sequence, while the enhancement layer are coded based on the predictions in both base layer frames and previous encoded enhancement layer frames by utilizing the inter-layer prediction tool. This feature obtains the better compression ratio than *simulcast* conditions [2]; on the other hand, it results in more calculation of macroblock (MB) modes than that of previous H.264/AVC standard. These macroblock modes are classified into two intra modes: {Intra_4×4 and Intra_16×16} and seven inter modes: {MODE_16×16, MODE_16×8, MODE_8×16, MODE_8×8, MODE_8×4, MODE_4×8, MODE_4×4}. When a macroblock is designated MODE_8×8, each MODE_8×8 block can be further split into submacroblocks: {8×4, 4×8 and 4×4}. In addition, special modes as so-called *direct modes* in B-frame and *skip modes* in P- and B-frame are provided. For spatial enhancement layer, another mode Intra_BL is added to cooperate with inter-layer prediction for enhancement layer [3]. As mentioned in H.264/AVC standard, the motion estimation and mode decision process in SVC is preformed by minimizing the rate distortion (RD) cost of the Lagrangian formulation [4], which is given in (1):

$$RDcost(s,c,Mode/QP, \lambda_{Mode}) = D(s,c,Mode/QP) + \lambda_{Mode} \cdot R(s,c,Mode/QP) \quad (1)$$

where s and c are the sample data of original and reconstructed block respectively, *Mode* is one of candidate macroblock modes illustrated above. *QP* is the quantization parameter. λ_{Mode} is Lagrangian multiplier for mode decision. D is the residual of motion estimation and R gives the number of encoding bits. This rate distortion operation is an exhaustive search process with traversing every candidate mode to find the best one for the current macroblock. It consumes more than 50% encoding time on rate distortion optimization in motion estimation [5]. Especially, the time extremely increases when multi-layers present in SVC. Therefore, fast mode decision is a very important and essential role to enhance the encoding speed.

II. THE INVESTIGATION OF MACROBLOCK CORRELATION

A. Inter-layer Prediction

Inter-layer prediction tools enable the usage of as much lower layer information as possible for improving the current encoding efficiency of the enhancement layer.

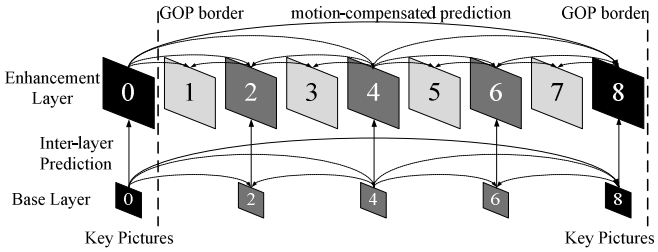

Figure 1. Multi-layer structure with inter-layer prediction.

A two-spatial-layer example with inter-layer prediction is shown in Fig. 1 [6]. GOP (group of picture) size is 8. The 0th and 8th picture are key pictures, others are Hierarchical-B pictures. All inter-layer prediction mechanisms are switchable for enhancement layer. These mechanisms are not suitable for base layer because it is the lowest layer in encoder. When the inter-layer prediction is off, multi-layer signals are equal to

multiple independent sequences transmission as same as *simulcast* does. In order to improve the encoding efficiency, the inter-layer prediction is normally set to *adaptive mode*. With this condition, the inter-layer prediction signal is either formed by motion compensated prediction inside the same layer, or by upsampling the reconstructed lower layer signal. Adaptive inter-layer prediction chooses the best mode using the rate distortion optimization function with neglecting the encoding time which spends tremendous coding time. Consequently, this motivates the fast mode decision algorithm and it is also the target of this paper.

B. Macroblock Correlation

Usually, the successive frames exhibit strong correlation in temporal and spatial domain. By employing that characteristic, the temporal and spatial redundancy from the video sequence can be removed during motion prediction to achieve the further compression.

First, regroup all the candidate modes for macroblock: {MODE_16×16, MODE_16×8, MODE_8×16, MODE_8×8, Intra_4×4, Intra_16×16 and Intra_BL} and for submacroblock: {8×8, 8×4, 4×8 and 4×4} [7]. Then, record and sort all the modes for macroblock and submacroblock respectively by its rate distortion cost after encoding any macroblock. Let positive integer N be the macroblock mode quantity and positive integer SN be the submacroblcok quantity for choosing from the forward and backward reference lists, where $N \in [1,7]$, $SN \in [0,4]$. Their relationship is in (2):

$$SN = \lfloor \alpha \cdot N + \beta \rfloor \tag{2}$$

where $\lfloor \cdot \rfloor$ operator indicates round down to the nearest integer. And α and β are the influence factors which adjust the value proportion between N and SN. The submacroblock is valid only when macroblock mode is MODE_8×8.

From the experiment of four standard sequences: BUS, FOOTBALL, CREW and SOCCER with 65 frames, we found that there is an obvious correlation between all the evaluated modes of the co-located reference macroblocks and the best mode of current macroblock among Hierarchical-B pictures. Table I gives the relationship at N=3, 4 and 5 (excluding submacroblock statistics):

TABLE I. MACROBLOCK CORRELATION

Sequence	N=3		N=4		N=5	
	Probability (%)	AC^a (%)	*Probability* (%)	AC^a (%)	*Probability* (%)	AC^a (%)
Bus	81.75	56.39	91.95	70.40	99.19	82.25
Football	80.63	57.01	91.38	71.45	97.22	81.98
Crew	72.47	55.13	81.40	69.46	93.60	82.11
Soccer	74.39	52.87	85.06	66.58	95.02	79.68

a. AC is the proportion of actual computation over to original exhaustive computation.

The actual mode approaches the current best mode with increasing the quantity N for Hierarchical-B pictures. There 81.75% macroblocks can acquire their best modes calculated by rate distortion optimization function in sequence BUS at N = 3, while it is only equal to calculate 56.39% of original macroblocks compared to the original exhaustive mode decision algorithm. In next section, a detailed algorithm will be demonstrated.

III. PROPOSED FAST MODE DECISION ALGORITHM

A. Temporal Degressive Coefficient

The temporal scalability of Hierarchical-B structure with a GOP size of 16 is depicted in Fig. 2 [8]. The frames of index 0 and 16 are key pictures, the others are B pictures. The temporal resolution is 4 (= $\log_2 GOPsize$, $GOPsize$ = 16). It is obvious that the picture B^0 is the only picture in temporal Level 0 so that the proposed fast mode decision algorithm is not exploited for it, which has the significant effect to the successive B pictures on encoding quality. The incorrect encoding of B^0 can decrease the quality of GOP extremely.

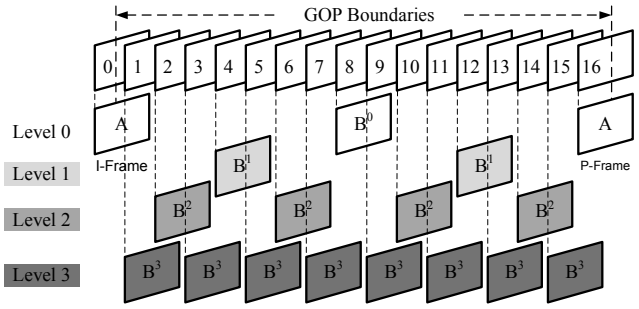

Figure 2. Temporal scalability with a GOP size of 16.

The temporal level is decided by the parameter *TemporalLevel*, whose maximum value is 6 when the maximum GOP size is 64. The temporal distance between two adjacent pictures in same temporal level is much closer when the *TemporalLevel* value is increasing. In case of the frame rate is 30 frames per second in Fig. 2, the sampling interval is 4/15 second for Level 1, 2/15 and 1/15 for Level 2 and Level 3, respectively. If the N is given the maximum value 7, which assuring the quality of B^0, the value of N for B^1 may be less than the maximum value because the correlation is much stronger in Level 2 than that in Level 1. Hence, a fixed value of N is not necessary for all the temporal level. The value should be much smaller when the *TemporalLevel* is larger.

This proposed algorithm adopts a temporal degressive coefficient method in co-located reference macroblock modes in order to guarantee the fast mode decision without decreasing the negligible encoding quality. These parameters are designated with k_{B0} = 7, k_{B1} = 6, k_{B2} = 5, k_{B3} = 4, k_{B4} = 3 and k_{B5}= 3 from Level 0 to Level 5.

B. Motion Speed

In most cases, the values of motion vectors provide information on the speed and the direction of movement of objects. Smaller values are obtained for relative static background and smoothly moving objects, and larger values indicate the speedy motion [4]. The high-speed movement of objects involved from time to time in the real video sequences. In these cases, the displacement of motion vectors is taken into consideration for co-located reference macroblock. A threshold TH is introduced to judge the status of movement in (3):

$$MV_x >= TH \text{ or } MV_y >= TH,\ TH \in \text{ positive integer} \tag{3}$$

where subscript x indicates horizontal direction and y indicates vertical direction. TH is a predefined parameter and it is set to 4

for this proposed algorithm. For simplicity, the best motion vectors for MODE_16×16 are checked only. If equation (3) is true, the motion speed of objects contained in the co-located reference macroblock is regarded as fast movement and all modes will be enabled for the current macroblock.

C. Proposed Algorithm

Record all the search modes for each macroblock in the whole encoding process and the pictures of base layer are not applied this proposed algorithm. Fig. 3 shows the flowchart of the selection of co-located reference macroblocks and Fig. 4 gives the flowchart of mode decision module of Fig. 3.

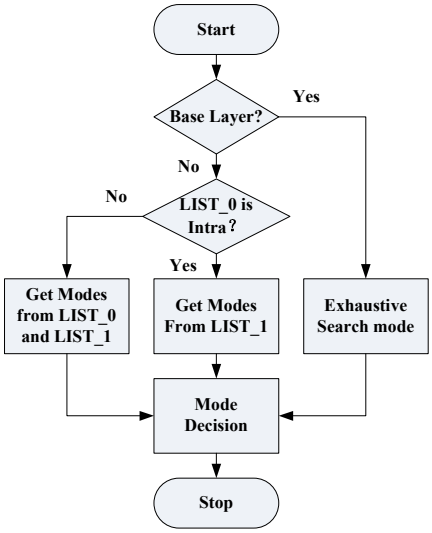

Figure 3. Flowchart of the selection of co-located reference macroblocks.

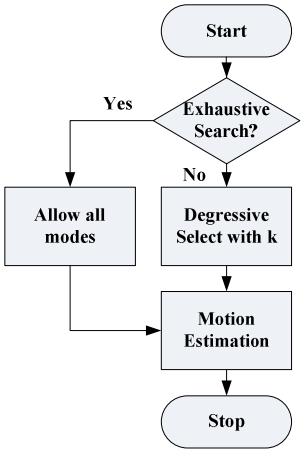

Figure 4. Flowchart of the mode decision of Fig. 3.

Step 1. Encode I and P pictures in enhancement layer with exhaustive search scheme.

Step 2. Position the co-located reference macroblocks at forward and backward reference pictures for current macroblock in B picture.

Step 3. Select certain amount of modes from the forward and backward reference macroblocks according to the parameter k_{Bn} in Section III-A when forward reference picture is not I-frame. Here, n = *TemporalLevel*.

Step 4. Select modes only from the backward reference macroblock if the forward reference picture is I-frame.

Step 5. Compose the modes with a new group for the current macroblock in mode decision.

Step 6. If the last macroblock in Hierarchical-B picture of GOP has been encoded, go forward to encode the next GOP. Otherwise, go back to Step 2.

IV. Simulation Results

The proposed fast mode decision algorithm is implemented in JSVM 9.1 encoder program [8]. The test platform used is Intel Pentium IV, 3.0 GHz CPU, 1G RAM with Windows XP professional operating system. The test condition is in Table II and the simulation parameters in configuration files are shown in Table III. Standard test sequences consist of the Bus, Football, Foreman, Mobile, City, Crew, Harbour and Soccer test data. The spatial resolution of the first four sequences is 352×288 luma samples per picture (CIF), while the last four sequences have a spatial resolution of 704×576 luma samples per picture (4CIF). All image frames used 4:2:0 color sampling and are downsampled by the linear filter kernel with tap values {-8, 0, 24, 48, 48, 24, 0, -8} (normalized by a 128 divisor with rounding) [2]. All enhancement layers are set to *adaptive* inter-layer prediction.

TABLE II. Simulation Condition

Sequence	Frame	GOP	Layers
Bus	65	32	2
Football	65	16	2
Foreman	65	32	2
Mobile	65	32	2
City	65	64	3
Crew	65	16	3
Harbour	65	64	3
Soccer	65	32	3

TABLE III. Simulation Parameters in Configuration File

Parameter	Value
SearchMode	4
SearchFuncFullPel	3
SearchFuncSubPel	2
SearchRange	96
BiPredIter	4
IterSearchRange	8
QP (base layer)	28
QP (enh. Layer 1)	32
QP (enh. Layer 2)	38

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 (3), (4) and (5):

$$\Delta Time = (Time_B - Time_A) / Time_A \times 100\% \quad (3)$$

$$\Delta PSNR = PSNR_B - PSNR_A \quad (4)$$

$$\Delta Bitrate = (Bitrate_B - Bitrate_A) / Bitrate_A \cdot 100\% \quad (5)$$

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 IV. EXPERIMENTAL RESULTS

Sequence	ΔPSNR /(dB)	ΔBitrate /(%)	ΔTime /(%)
Bus	-0.0121	0.42%	-22.95%
Football	-0.0048	0.34%	-21.05%
Foreman	-0.0180	0.22%	-29.02%
Mobile	-0.0318	-0.03%	-25.92%
City	-0.0127	0.44%	-47.36%
Crew	0.0603	1.98%	-41.01%
Harbour	-0.0176	0.39%	-51.28%
Soccer	-0.0131	0.27%	-45.34%
Average	-0.0062	0.50%	-35.49%

Table IV 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 35% on average of original JSVM original scheme. More time is saved with the number of sequence layers increasing. Moreover, this proposed algorithm achieves negligible loss in PSNR and increments in bit rate at a large bit rate range.

Fig. 5 to Fig. 8 present the rate distortion curves with four different sequences: Bus, Football, Soccer and Crew. The rate points are declared in [9]. Sequence Bus and Football belong to video streaming application in a two-layer configuration, and Soccer and Crew are broadcasting application with three-layer application. From these figures, 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.

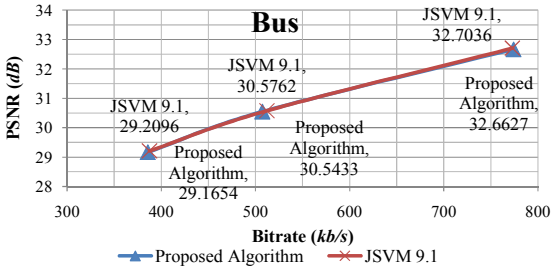

Figure 5. Rate distortion curve of Sequence Bus.

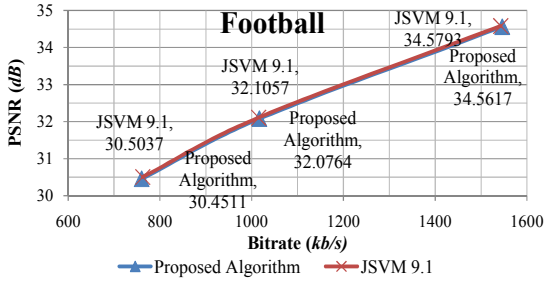

Figure 6. Rate distortion curve of Sequence Football.

V. CONCLUSIONS

In this paper, a fast mode decision algorithm based on the correlation between the evaluated co-located reference macroblocks and the current best mode has been proposed for enhancement Hierarchical-B pictures in spatial SVC. The algorithm can reduce the computational time nearly 35% with slight loss of PSNR and bit rate increase.

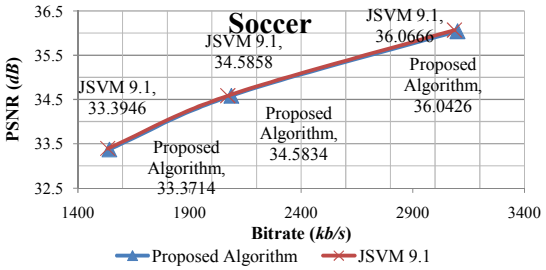

Figure 7. Rate distortion curve of Sequence Soccer.

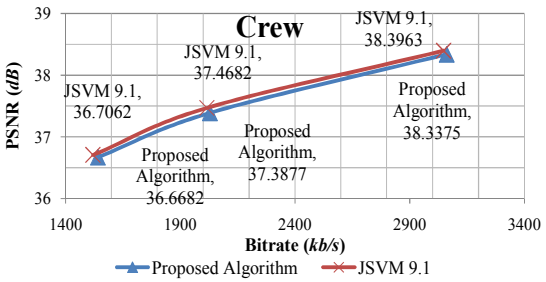

Figure 8. Rate distortion curve of Sequence Crew.

REFERENCES

[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] C. A. Segall and G. J. Sullivan, "Spatial Scalability Within the H.264/AVC Scalable Video Coding Extension," IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, pp. 1121-1135, 2007.

[3] H. Li, Z. G. Li, and C. Wen, "Fast Mode Decision Algorithm for Inter-Frame Coding in Fully Scalable Video Coding," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, pp. 889, 2006.

[4] 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.

[5] 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.

[6] H. Schwarz, M. Wien, and J. Vieron, "JSVM Software Manual," ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q, vol. 6, 2006.

[7] W. Tae-Shick, P. Chun-Su, K. Jun-Hyung, A. M.-S. Y. Min-Seok Yoon, and A. S.-J. K. Sung-Jea Ko, "Improved inter-layer intra prediction for scalable video coding," IEEE Region 10 Conference, 2007, pp. 1-4.

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

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