MULTI-LOOP SCALABLE VIDEO CODEC BASED ON HIGH EFFICIENCY VIDEO CODING (HEVC)

Do-Kyoung Kwon, Madhukar Budagavi and Minhua Zhou

Systems and Applications R&D Center, Texas Instruments, Inc.
12500 TI Blvd., Dallas, TX 75243

ABSTRACT

A multi-loop scalable video coder for high efficiency video coding (HEVC) is proposed in this paper. A coding unit (CU)-level inter-layer sample prediction tool is proposed to exploit redundancy between enhancement-layer and up-sampled base-layer pictures. To reduce decoded picture buffer size and memory bandwidth in a multi-loop decoder, a hierarchical inter-layer prediction tool is proposed as well using two picture-level flags. The proposed solution requires minimum amount of changes relative to the single-layer HEVC codec to support HEVC scalable coding, and provides a good complexity and coding efficiency trade-off as revealed by the experimental results.

Index Terms— HEVC, SHVC, scalable video coding, inter-layer prediction

1. INTRODUCTION

High Efficiency Video Coding (HEVC), which is the successor of H.264/AVC [1], has been recently developed [2]. HEVC is reported to provide the same subjective quality at roughly half bitrate compared with H.264/AVC. While HEVC Version 1 was finalized in January 2013, the standardization of the scalable extension of HEVC (SHVC) has also started recently. The joint Call for Proposal (CfP) on scalable video coding extension of HEVC [3] was issued in July 2012 and the standardization started in October 2012 based on the responses to CfP submitted by participants.

Scalable video coding is a method to convey multiple videos of different qualities and resolutions into a single bitstream. Compared to simulcast, by which multiple videos are coded and sent in multiple independent bitstreams, the scalable video is more efficient because of the use of inter-layer prediction techniques.

Scalable video codec can be designed based on either multi-loop architecture or single-loop architecture. In multi-loop architecture, a full decoding loop takes place in every intermediate layer needed to decode a target layer. Both intra- and inter-coded blocks are fully reconstructed in all layers and the reconstructed samples from the intermediate layers are used as additional reference for the enhancement layer (EL). While the multi-loop scalable codec increases decoded picture buffer (DPB) size and memory bandwidth for motion compensation (MC) on the decoder side, its coding efficiency is better than the single-loop scalable codec since it can fully exploit high correlation between EL picture and reconstructed reference-layer picture, and does not require the reference layers to use constrained intra prediction.

In single-loop architecture, which H.264/SVC [4] is based on, a full decoding loop takes place only once in a target layer. Inter-coded blocks in intermediate layers are not reconstructed. The single-loop scalable codec has an advantage that the DPB size and memory bandwidth for MC remain the same on the decoder side regardless of the number of layers. However, it requires more sophisticated inter-layer prediction techniques such as residual prediction and motion prediction in order to achieve comparable coding efficiency as the multi-loop scalable codec. Therefore, it requires significant amount of changes to realize scalable coding on the top of existing single-layer codec. In addition, in order to avoid the need of doing MC in the reference layers, constrained intra prediction has to be forced in all layers except for the highest EL, which inevitably reduces video quality in intermediate layers, especially on pipelined hardware architectures.

When designing scalable video codec, an important issue to consider is the additional cost involved to support scalable coding on the top of the existing single-layer codec. For the cost-effective scalable coding solutions and fast adoption of scalable codec into market, scalable coding tools should minimize changes relative to the single-layer architecture. For this reason, we propose in this work a multi-loop solution that uses only inter-layer sample prediction. Discarding inter-layer residual prediction and motion prediction reduces complexity involved with them. Another advantage of our solution is that it can easily support a legacy base-layer (BL) codec conforming to standards other than HEVC since only the output of the BL codec is needed to decode EL. Considering that the H.264/AVC is the most popular video codec nowadays in many applications such as Smartphone, Tablet, video surveillance, video conferencing and broadcasting, this feature is very important to the success of SHVC.

To address DPB size and memory bandwidth issues in multi-loop solution, we also propose a hierarchical inter-

layer prediction structure to trade off coding efficiency with DPB size and memory bandwidth. Two flags are signaled for each picture, one to indicate whether a picture is used as a reference for EL pictures and the other to indicate whether a picture is used as a temporal reference for other pictures being referred to by ELs. Two flags are signaled based on hierarchical structure.

This paper is organized as follows. The proposed inter-layer sample prediction and hierarchical inter-layer prediction structure are described in Sec. 2 and Sec. 3, respectively. Experimental results are provided in Sec. 4, followed by the concluding remarks in Sec. 5. This paper is based on our standard contributions [5, 6]. Our proposal in [5] was selected as a basis of initial software model for the block-level approach in SHVC standardization [7, 8] since it require minimum amount of changes relative to the single-layer HEVC codec and provide a good complexity and coding efficiency trade-off.

2. PROPOSED SCALABLE HEVC CODEC ARCHITECTURE

Without loss of generality, the proposed solution for the scalable extension of HEVC is described assuming that two layers (i.e. BL and EL) are coded into a single bitstream.

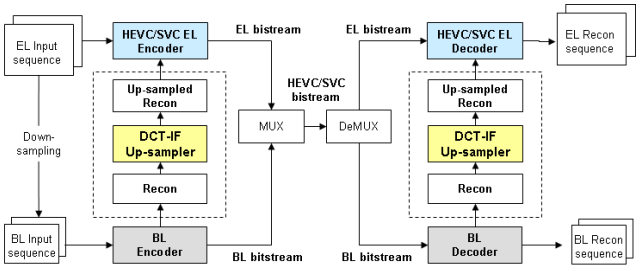

Fig. 1. Proposed architecture for the scalable extension of HEVC

Fig. 1 illustrates the structure of the proposed solution. An up-sampler is the only block newly added to scale reconstructed BL pictures since the reconstructed BL picture is the only information needed additionally for EL encoding and decoding. A DCT-IF up-sampling filter [9] is employed to scale the reconstructed BL picture to the size of EL in case of spatial scalability. Luma and chroma up-sampling filter coefficients for 2x and 1.5x spatial scalability are shown in Table 1.

Table 1. DCT-IF up-sampling filter for luma and chroma pictures.

Phase	Luma filter (8-tap)	Chroma filter (4-tap)
1/3	{ 1, 4, -11, 52, 26, -8, 3, -1}	{ -5, 50, 22, -3 }
1/2	{ -1, 4, -11, 40, 40, -11, 4, -1}	{-4, 36, 36, -4 }
2/3	{ -1, 3, -8, 26, 52, -11, 4, 1}	{ -3, 22, 50, -5 }

The up-sampled reconstructed BL picture is used as a reference for inter-layer sample prediction (ILP) in EL. The

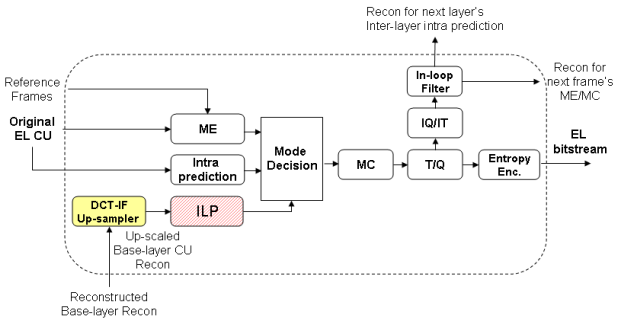

Fig. 2. CU encoding flow in enhancement layer

ILP is performed at CU level in the proposed solution. Fig. 2 shows the CU encoding flow in EL. For an input CU, ILP mode cost with respect to the collocated up-sampled base-layer CU is calculated in addition to regular intra and inter mode costs. Then the best CU mode is determined among best intra mode, best inter mode and ILP mode. If ILP mode is chosen as the best mode, *ilp_flag* is signaled after *skip_flag* and quantized transform coefficient are coded by the same procedure as in base layer. When ILP is not the best mode, *ilp_flag* is set to 0 and the CU is coded by the same procedure as in base layer. Fig. 3 shows the corresponding CU decoding flow in EL.

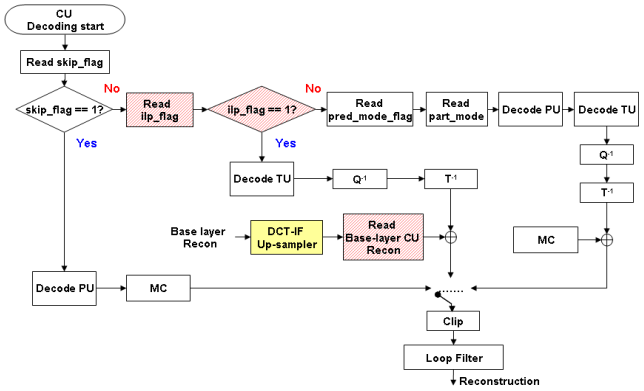

Fig. 3. CU decoding flow in enhancement layer

The *ilp_flag* is coded using context-adaptive binary arithmetic coding (CABAC) using three contexts. The first context is used when both top and left CUs are ILP mode, the second context is used when one of neighboring CUs is ILP mode and the third context is used when both neighboring CUs are not ILP mode. Finally, it should be noted that an ILP mode is treated as if intra DC mode for the purpose of determining transform type and in deblocking filter.

3. HIERARCHICAL INTER-LAYER PREDICTION

The proposed multi-loop scalable HEVC solution introduces marginal architectural changes into the single-layer HEVC codec as described in the previous section. However, DPB size and memory bandwidth in decoder increase because all

the pictures in intermediate layers should be reconstructed to decode a target layer. When a decoder tries to decode the highest EL among more than 2 layers, the required DPB sizes and memory bandwidth could be prohibitive.

To control DPB size and memory bandwidth, it is proposed to signal two picture-level flags: *ilp_ref_flag* that indicates whether a picture is used as a reference for an EL and *ilp_tref_flag* that indicates whether a picture is used as a temporal reference for other pictures with *ilp_ref_flag* equal to 1. It is possible to signal them in any high-level header including slice header. In this work, they are signaled in NAL unit header just for example and their values are assigned to each NAL unit based on a hierarchical structure.

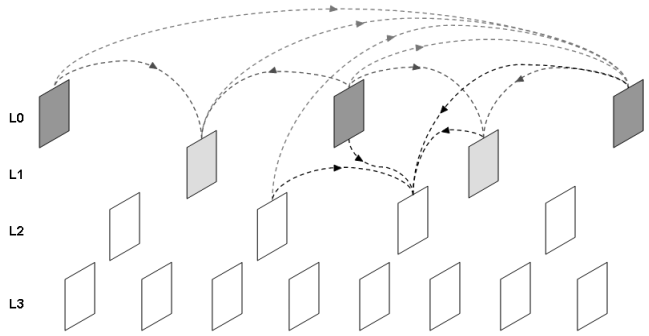

Fig. 4. Prediction structure for RA in CTC

Fig. 4 demonstrates the prediction structure in base layer for the random access (RA) common test condition (CTC) [10], where the dashed line shows temporal reference picture. Among the pictures in level 2 (L2) and level 3 (L3), the prediction arrows are shown for one picture to simplify figure. In the multi-loop decoder, assuming that inter-layer prediction is enabled for all access units, all pictures in all levels (L0, L1, L2, L3) in base layer should be fully decoded for inter-layer prediction in an enhancement layer.

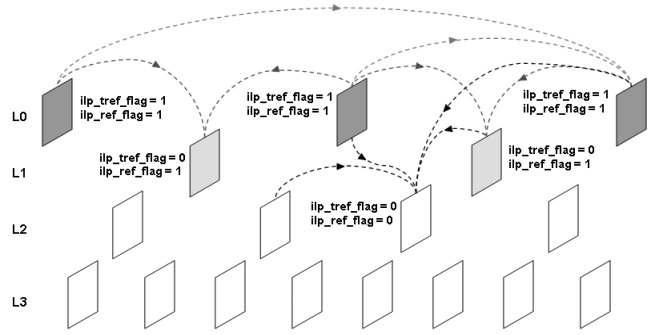

Fig. 5. Prediction structure for RA in CTC for 2-level ILP

Fig. 5 shows the 2-level hierarchical inter-layer prediction using the proposed two flags, where the pictures in only two lowest levels (i.e. L0 and L1) are used as references for inter-layer prediction. Such pictures (i.e. L0 and L1 pictures in Fig. 5) have *ilp_ref_flag* equal to 1 and are named ILP-REF picture. For the pictures not used for inter-layer prediction (e.g. L2 and L3 pictures in Fig. 5), *ilp_ref_flag* is set to 0. Moreover, the pictures in L0 in Fig. 5 are used as temporal references for other ILP-REF pictures in the same layer. Such pictures have *ilp_tref_flag* set equal to 1 and are named ILP-TREF picture. L1 pictures in Fig. 5 are not used in prediction of ILP-REF pictures, hence *ilp_tref_flag* is set to 0 for them. And *ilp_ref_flag* and *ilp_tref_flag* are set to 0 for all pictures in L2 and L3. In addition, the following two constraints are imposed on these flags: 1) ILP-REF pictures should have only ILP-TREF pictures as temporal references and 2) non ILP-REF picture also should be non ILP-TREF picture.

The prediction structure of Fig. 5 leads to coding efficiency loss in base layer since the number of reference pictures are reduced. However, with the proposed picture-level flags and constraints, it is possible to generate scalable video bitstream considering the allowed DPB size and memory bandwidth in decoder. The DPB size is determined by the pictures having *ilp_tref_flag* equal to 1 and memory bandwidth is determined by the pictures having *ilp_ref_flag* equal to 1. In Fig. 4, DPB should be as large as the size of four pictures. However, in Fig. 5, DPB needs to store only two pictures at the same time and memory bandwidth for MC is reduced roughly by 75%. By assigning two flags differently, we can trade off coding efficiency with DPB size and memory bandwidth. For example, in case of 3-level hierarchical inter-layer prediction, by assigning *ilp_ref_flag* equal to 1 to the pictures in L0, L1 and L2, we can reduce coding efficiency loss by increasing memory bandwidth twice compared with Fig. 5.

Finally, it is worth noting that all non ILP-REF pictures (which are also non ILP-TREF pictures) in base layer (i.e. pictures in L2 and L3) can be discarded in decoder without any effect on the successive pictures since non ILP-REF pictures are not used as temporal references and inter-layer references. And introducing the proposed flags does not affect the DPB management process. DPB is managed by the same reference picture set (RPS) syntax [2] without any change. Only modification in decoding and encoding process is that the reference picture list for ILP-REF picture is generated considering *ilp_tref_flag* of the pictures in DPB so that non ILP-TREF pictures are not used as temporal references.

4. EXPERIMENTAL RESULTS

The proposed scalable HEVC solution was implemented in HM6.1 software and evaluated using the test conditions in [3] with HEVC and H.264/AVC base-layer encoders separately.

In the first test, both BL and EL sequences are coded by HEVC encoder and the proposed solution is evaluated against simulcast for SNR, 1.5x and 2x spatial scalability using all intra (AI) and RA configuration. BL was coded by HM-6.1 using QP values of 22, 26, 30 and 34. For spatial scalability, for each BL QP, EL was coded using BL QP – 2, 0, +2 and +4. For SNR scalability, EL QP was set to BL QP – 2, – 4, –6 and –8. In the second test, BL and EL sequences

are coded by AVC and HEVC encoders, respectively. The proposed solution is evaluated against simulcast for 1.5x and 2x spatial scalability using RA configuration. BL was coded by JM18.3 using QP values of 21, 25, 29 and 33 and EL QP was set to BL QP – 1, 1, +3 and +5 for each BL QP.

Two 4K×2K and five 1080p sequences were used as EL input sequences in the test along with corresponding BL input sequences provided in [3] (which were generated by down-sampling in case of spatial scalability). The coding efficiency by the proposed solution is measured using EL BD-Rate and 'BL + EL' BD-Rate. EL BD-Rate evaluates the two coding solutions (i.e. simulcast and the proposed) by comparing EL's bitrates at the same EL's PSNR. 'BL + EL' BD-Rate compares total bitrates of BL and EL at the same EL's PSNR.

4.1. HEVC BL – HEVC EL

Table 2 shows the average BD-rate gain by the proposed solution over simulcast for different configurations with HEVC base layer. Only inter-layer sample prediction tool is enabled for all pictures in this test. As shown in Table 2 where negative number means bit savings, the proposed solution improves the coding efficiency significantly. Even though the inter-layer sample prediction is simple, the strong correlation among the reconstructed BL picture and the original EL picture makes it quite effective in the multi-loop architecture.

Table 2. Average BD-Rate gain over simulcast when ILP is enabled for all pictures.

	EL BD-Rate (%)			'BL+EL' BD-Rate (%)		
	Y	Cb	Cr	Y	Cb	Cr
AI 2x	-37.4	-36.1	-36.5	-23.6	-22.7	-23.1
AI 1.5x	-57.3	-56.3	-56.4	-32.2	-31.7	-31.7
RA 2x	-27.3	-15.5	-14.6	-16.7	-6.9	-6.1
RA 1.5x	-46.6	-36.1	-33.9	-25.8	-17.2	-15.4
RA SNR	-35.4	-23.4	-21.0	-21.3	-11.5	-9.5

The effects of hierarchical structuring of inter-layer prediction are presented in Table 3 and Table 4, which shows the BD-rate saving of proposed 3-level and 2-level hierarchical inter-layer prediction, respectively. This test is done only for RA because the DBP size and memory bandwidth do not increase for AI and therefore it is reasonable to enable inter-layer prediction for all pictures.

For 3-level hierarchical signaling, *ilp_ref_flag* and *ilp_tref_flag* are set to 1 for L0 and L1 pictures, and *ilp_ref_flag* is set to 1 while *ilp_tref_flag* is set to 0 for L2 pictures, and both flags are set to 0 for L3 pictures. With these setting, as shown in Table 3, we can reduce memory bandwidth by half at the cost of luma BD-rate loss of 1.4% ~ 2.2%. For 2-level hierarchical signaling shown in Fig. 5,

Table 3. Average BD-Rate gain for 3-level hierarchical ILP (i.e. ILP is enabled for the pictures in L0, L1 and L2).

	EL BD-Rate (%)			'BL+EL' BD-Rate (%)		
	Y	Cb	Cr	Y	Cb	Cr
RA 2x	-24.9	-14.4	-13.6	-15.3	-6.7	-6.0
RA 1.5x	-42.8	-33.2	-31.3	-23.6	-16.1	-14.5
RA SNR	-32.4	-22.2	-19.9	-19.6	-11.3	-9.3

Table 4. Average BD-Rate gain for 2-level hierarchical ILP (i.e. ILP is enabled for the pictures in L0 and L1).

	EL BD-Rate (%)			'BL+EL' BD-Rate (%)		
	Y	Cb	Cr	Y	Cb	Cr
RA 2x	-21.2	-12.4	-11.7	-13.2	-6.0	-5.4
RA 1.5x	-36.3	-27.7	-26.1	-20.3	-13.6	-12.2
RA SNR	-28.1	-19.8	-17.9	-17.3	-10.4	-8.8

ilp_ref_flag and *ilp_tref_flag* are set to 1 for L0 pictures, and *ilp_ref_flag* is set to 1 while *ilp_tref_flag* is set to 0 for L1 picture, and both are set to 0 for L2 and L3 pictures. The memory bandwidth is reduced by 75% with luma BD-rate loss of 3.5% ~ 5.5%. The amount of loss is higher as expected. However, it is useful when the memory bandwidth and DPB size are critical in decoder especially when there are more than 2 layers in a bitstream.

4.2. AVC BL – HEVC EL

When inter-layer prediction with AVC base layer is enabled for all EL pictures, the average BD-rate gain by the proposed solution over simulcast is shown in Table 5. When compared to the results with HEVC BL in Table 2, the coding gain was slightly reduced. However, the proposed solution still resulted in the significant improvement especially for 1.5x scalability.

Table 5. Average BD-Rate gain over simulcast with AVC BL.

	EL BD-Rate (%)			'BL+EL' BD-Rate (%)		
	Y	Cb	Cr	Y	Cb	Cr
RA 2x	-24.8	-13.1	-12.1	-16.1	-6.8	-6.1
RA 1.5x	-40.4	-31.4	-28.1	-22.5	-15.6	-13.0

5. CONCLUSIONS

Easy extension from the single-layer HEVC architecture is crucial to the success of SHVC in video market place. In this work, a simple but effective SHVC solution was proposed based on the multi-loop scalable codec approach. The proposed inter-layer sample prediction at CU-level provides high coding gains while requires marginal changes of the HEVC architecture. The idea of hierarchical inter-layer prediction structure was also proposed to reduce DPB size and memory bandwidth on the decoder side. It turned out that the proposed solution provides a good complexity and coding efficiency trade-off.

6. REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits and Syst. Video Technol.,* vol. 13, no. 7, pp 560-576, Jul. 2003.

[2] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 9," JCTVC-K1003, ITU-T SG16 WP3 and ISO/IEC JCT1/SC29/WG11, Oct. 2012.

[3] ISO/IEC and ITU-T, "Joint call for proposal on scalable video coding extension of high efficiency video coding (HEVC),"N12957, ISO/IEC JTC1/SC29/WG11, Jul. 2012.

[4] H. Schwarz, D. Marpe and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 17, no. 9, September 2007.

[5] D.-K. Kwon, M. Budagavi and M. Zhou, "Description of scalable video coding technology proposal by Texas Instruments," JCTVC-K0038, ITU-T SG16 WP3 and ISO/IEC JCT1/SC29/WG11, Oct. 2012.

[6] D.-K. Kwon, M. Budagavi and M. Zhou, "Hierarchical inter-layer prediction in multi-loop scalable extension of HEVC," JCTVC-K0264, ITU-T SG16 WP3 and ISO/IEC JCT1/SC29/WG11, Oct. 2012.

[7] J. Chen, V. Seregin, M. Zhou, B. Jeon, S. Lei, E. Nassor, K. Ugur, A. Segall, A. Tabatabai, H. Yu, J.W. Kang, Y. Ye, W. Zhang and P. Bordes, "A proposal for Scalable HEVC test model", JCTVC-K0348, ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Oct. 2012.

[8] G. Sullivan and J.-R. Ohm, "Meeting report of the 11th meeting of the Joint Collaborative Team on Video Coding (JCT-VC), Shanghai, CN, 10–19 Oct 2012", JCTVC-K_Notes_dD, ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Oct. 2012.

[9] J. Boyce, D. Hong and W. Jang, "Information for HEVC scalability extension", JCTVC-G078, ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Nov. 2011.

[10]F. Bossen, "Common HM test conditions and software reference configurations," JCTVC-K1100, ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Oct. 2012.