Improving SHVC Performance with a Block based Joint Layer Prediction Solution

Xiem HoangVan
VNU-University of Engineering and Technology, Hanoi
xiemhoang@vnu.edu.vn

***Abstract*—Considering the need for a more powerful scalable video coding solution beyond the recent Scalable High Efficiency Video Coding (SHVC) standard, this paper proposes a novel SHVC improvement solution using the block based joint layer prediction creation. In the proposed improvement solution, the temporal correlation between frames is exploited through a motion compensated temporal interpolation (MCTI) mechanism. The MCTI frame is adaptively combined with the base layer reconstruction using a linear combination algorithm. In this combination, a weighting factor is defined and computed for each predicted block using the estimated errors associated to each prediction input. Finally, to achieve the highest compression efficiency, the fused frame is treated as an additional reference and adaptively selected using the rate distortion optimization (RDO) mechanism. Experiments conducted for a rich set of test conditions have shown that significant compression efficiency gains can be achieved with the proposed improvement solution, notably by up to 4.5 % BD-Rate savings on enhancement layer regarding the standard SHVC quality scalable codec.**

Keywords— *HEVC, SHVC, best prediction, joint layer mode*

I. Introduction

Modern multimedia applications such as video streaming, video conferencing, or video broadcasting has been asking for a more powerful video coding solution, which provides not only the high compression efficiency but also the scalability capability. To address this requirement, a scalable video coding [1] paradigm has been introduced as an extension of the most popular H.264/AVC standard [2]. However, with the recent development of video coding technology and the compression achievement of the newly High Efficiency Video Coding (HEVC) standard [3], a scalable extension of HEVC (known as SHVC) has been introduced in 2014 [4]. As reported, SHVC significantly outperforms the relevant benchmarks including SVC standard and HEVC simulcasting [5] while providing a larger scalability features, including spatial, temporal, quality, bit-depth, and color-gamut.

Similar to the prior SVC standard, SHVC is also designed with a layered coding structure where one base layer (BL) and one or several enhancement layers (ELs) constitute the scalable bitstream [4]. However, to make implementation and deployment easier as well as to make flexible backward compatibility with other standards like HEVC and H.264/AVC, SHVC does not follow the same conceptual approach adopted in the SVC where new macroblock-level signaling capabilities are defined to indicate whether the EL macroblock is predicted from the BL or the EL layers. Instead, in SHVC, the BL reconstructed picture is taken as an inter-layer reference (ILR) picture to be included in the EL prediction buffer, eventually after some inter-layer processing. In this context, the standard HEVC reference index signaling capabilities are enough to identify whether the EL block level prediction comes from the BL or the EL [4]. The main advantage is that this different scalable coding approach requires changes only in the HEVC high level syntax (HLS) and no changes in terms of the HEVC block level coding process.

SHVC has recently gained attentions due to its important role in video transmission and streaming. Most of the recent researches focus on the Inter-layer processing component since it directly affects to the compression efficiency of the SHVC standard. The main goal is to offer better predictions for EL Coding Units (CU). In [6], a generalized inter-layer residual prediction is proposed to improve the EL prediction accuracy by combining the obtained EL prediction with a new residual derived from both the BL and EL. Afterwards, in [7], a combined temporal and inter-layer prediction solution is presented in which the EL temporal prediction, BL collocated reconstruction and the BL residue are liner combined. However, the combination presented in [6, 7] employed fixed weights; thus, less adaptive to the variation of video content. In [8, 9], a different approach is followed as adaptive filters are proposed to be directly applied to the BL reconstructed samples. They simply tried to minimize the difference between the filtered BL reconstructions and the collocated pixels at EL. Following another way, the authors in [10] introduced an enhanced merge mode prediction, which also combines BL and EL available information [10]. However, this approach may significantly increase the coding complexity due to the use of pixel level – based merge prediction creation approach. Instead of employing the available EL prediction created with the guide of original information, the authors in [11] has presented a decoder side motion compensated temporal interpolation solution for improving the SVC performance. In this proposal, the EL decoded information is employed to create an additional prediction, which is expected to complement for the existent EL prediction.

Inspired from the work in [11], we propose in this paper a novel SHVC improvement solution, which employs not only the decoded information derived from the EL but also from the BL through a so-called "joint layer prediction". To achieve the high joint layer prediction quality, a weighting factor indicating the contribution of each prediction candidate is adaptively computed for each predicted block using an error estimation model. Final, the joint layer prediction frame is treated as an additional reference and adaptively selected using a rate distortion optimization (RDO) mechanism.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

The rest of the paper is organized as follows. Section II describes the proposed method, JLP-SHVC. Section III evaluates and discusses the compression performance obtained with the proposed solution. Finally, section IV gives some conclusions and future works.

II. Proposed SHVC Improvement Solution

Our recent work [11] has shown that an important compression gain can be achieved for SVC standard with a decoder side information creation solution. However, the work in [11] considered only the available information from the enhancement layer, i.e., forward and backward references, to create an additional reference. Hence, it was not able to take all available information at the enhancement layer. In this context, we present in this session a novel joint layer prediction (JLP) scheme, which takes into account not only the available information from the enhancement layer (EL) but also from the base layer (BL) to improve the SHVC performance.

A. JLP-SHVC encoder architecture

Figure 1 illustrates the proposed JLP-SHVC architecture. JLP creation is actually a technique of combination the BL reconstruction and the EL motion compensated temporal interpolation. Those modules are highlighted in Fig.1. It should be emphasized that, SHVC standard usually demands for original information to perform the motion estimation and motion compensation in its main loop under the consideration of original data [4]. However, in our proposal, since only the decoded information and the block based weight computation are used, the computational complexity associated to the proposed modules is negligible.

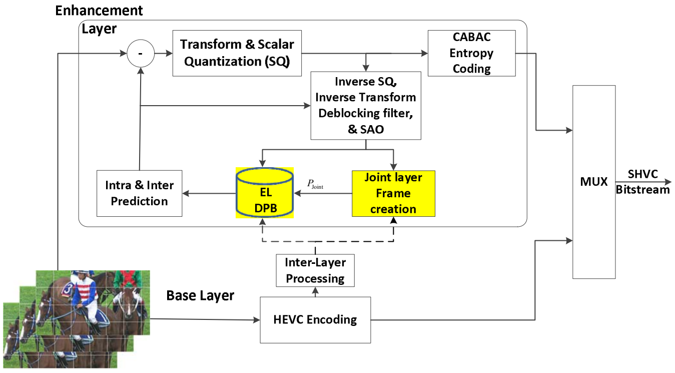

Fig.1. Proposed SHVC-DSFI architecture

In our proposal, the EL decoded frames are carefully studied and exploited to create a new reference for EL prediction. Hence, the quality of joint layer frame will directly affect to the final compression gain of the proposed SHVC.

B. Joint layer frame creation

Figure 2 illustrates the proposed joint layer frame creation. In the proposed joint layer frame creation, the well-known MCTI technique [12] is employed to create the EL side information while the frame fusion is employed to combine the EL side information and the BL reconstruction to form a high quality joint layer frame.

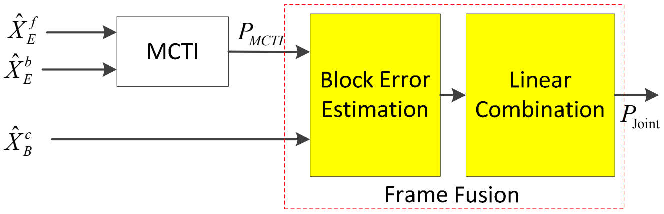

Fig.2. Proposed Joint Layer Prediction Creation

The proposed joint layer frame includes the following steps:

1) Motion compensated temporal interpolation: MCTI techniques work at the block level and extract a piecewise smooth motion field that captures the linear motion between the backward (past) and the forward (future) reference frames. The MCI SI creation solution consists in the following main steps inspired from [12]:

i) **Forward motion estimation**: This step aims to estimate the motion trajectory between the backward and forward frames. In this case, both reference frames are low pass filtered and used as references in a full search motion estimation algorithm using a modified matching criterion.

ii) **Bi-directional motion estimation**: To maintain the temporal consistency along frames, this step refine the motion information obtained from previous step. A hierarchical approach is used with block sizes of 16×16 and 8×8.

iii) **Weighted vector median filtering**: The weighted median vector filter improves the motion field spatial coherence by looking, for each interpolated block, for the candidate motion vectors at neighboring blocks, which can better represent the motion trajectory. This filter is also adjustable by a set of weights, controlling the filter strength (i.e. how smooth is the motion field) and depending on the block distortion for each candidate motion vector.

iv) **Motion compensation**: Finally, using the obtained motion vectors, the EL compensated frame is created by performing motion compensation on the forward and backward frames.

2) Frame fusion: The combination of the BL reconstruction, $\hat{X}_B^c$, and EL compensated frame, P_{MCTI}, is performed in this module. The higher quality of the fused frame, P_{joint}, the larger SHVC compression gain can be achieved. Therefore, it is proposed in this paper a linear combination for such fusion problem as the following formulation:

$$P_{\text{joint}} = w \times P_{MCTI} + (1-w) \times \hat{X}_B^c \qquad (1)$$

In this combination, a weighting factor, *w,* is defined and used to indicate the contribution from each predicted inputs. The weighting factor will directly affect to the final qulity of the joint prediction. In this case, the weighting factor computed for each block is performed using the block error estimation metrics as the following steps:

i) **Block error estimation:** First, the quality of each fusion frame candidate is assessed indirectly through the temporal correlation, E_{EL} , in MCTI and the spatial consistency, E_{BL} , in BL reconstruction candidate as the following metrics:

$$E_{EL} = \sum_{i-0}^{N} \left(\hat{X}_E^f(i,mv) - \hat{X}_E^b(i,mv) \right)^2 \tag{2}$$

$$E_{BL} = \sum_{i-0}^{N} \sum_{j=0}^{M} \left(\hat{X}_B^c(i) - \hat{X}_B^c(i,j) \right)^2 \tag{3}$$

Here, N is the total number of pixels in the current block, M is the number of surrounding blocks for assessing the spatial consistency. i is the pixel index and j is the surrounding block index.

ii) **Weight computation:** After computing the block error, a weighting factor is defined to indicate the contribution of each fusion candidate as follows:

$$w = \frac{E_{BL}}{E_{BL} + E_{EL}} \tag{4}$$

Finally, the weight computed in (4) is used for (1) to create the joint layer prediction. This prediction will be employed as an additional reference for the SHVC encoding process.

C. Joint layer frame manipultion

To take the full advantage of the proposed joint layer reference, the constructed P_{joint} frame is added to the Reference Picture Set (RPS) of EL, which is categorized as List0 and List1 as shown in Fig.3. Here, P_{joint} is inserted into the last position of both List0 and List1.

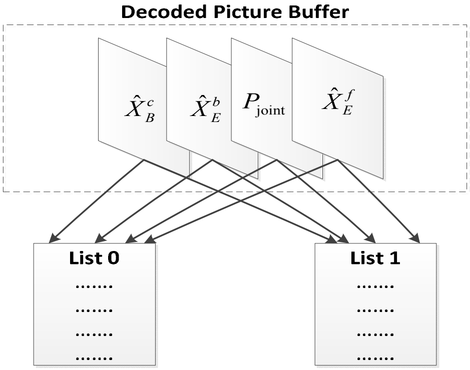

Fig. 3. Reference Picture List of EL

The long-term flag is chosen due to the signaling long-term reference in bitstream only needs least significant bits (LSB) of POC number. Furthermore, in SHVC, the Advance Motion Vector Prediction (AMVP) and Merge mode demand on searching up to 5 motion candidate blocks which belong to short-term and Inter-layer reference [4]. Since P_{joint} is actually a virtual reference frame containing only texture and no motion data, it guarantees that there is no conflict between signaling real reference frame and virtual reference frame in bitstream as well as preventing P_{joint} motion block data, which is not available, from adding to motion candidate of AMVP and Merge mode.

III. Experimental Results

This section presents the performance evaluation of the proposed SHVC-DSFI scheme, notably in comparison with the state-of-the-art SHVC standard, the most relevant benchmark for this evaluation.

A. Test conditions

In this work, three video sequences from class D (416×240) and two sequences from class C (832×480) from JCT-VC common test conditions [13] were selected due to their diversity of texture and motion characteristics. The SHM reference software version 12.1 [14] is used to develop our improvement solution. The test condition can be summarized as in Table 1 while first frame of each test sequence is illustrated in Fig.4.

RaceHorses BlowingBubbles BasketballPass

PartyScene BasketballDrill

Fig. 4. Illustration of the first frame of each testing sequence

Table 1. Summarization of the test conditions

Test sequence, spatial resolution, frame rate, and number of encoded frames	1. RaceHorses, 416×240, @30hz, 300 frames 2. BlowingBubbles, 416×240, @50Hz, 500 frames 3. BasketballPass, 416×240, @50Hz, 500 frames 4. PartyScene, 832×480, @50Hz, 500 frames 5. BasketballDrill, 832×480, @50Hz, 500 frames
Codec settings	Search range: 32 Intra period: -1 GOP size: 2 (IBPBPBP…)
Quantization parameters	{BL;EL} = {(38; 34), (34; 30), (30; 26), (26; 22)}

B. Joint layer prediction quality assessment

To assess the quality of new predicted frame, P_{joint} , the common objective video quality metric, PSNR (Peak Signal to Noise Ratio) was computed and compared for each sequence. The higher PSNR is received, the better image quality can be obtained. The joint predicted frame is compared with the MCTI compensated frame as created in [11]. Fig. 5 illustrates the PSNR comparison for several test sequences.

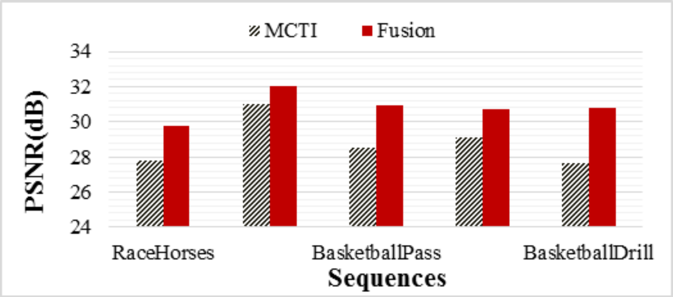

Fig.5. Joint layer prediction frame quality comparison

As shown in Fig. 5, the quality of joint prediction frame signficiantly outperforms the MCTI reference. This reflects the accuracy of the proposed fusion model.

C. SHVC-JLP rate-distortion performance assessment

As common in video coding evaluation, the compression efficiency is one of the most important factors indicating the efficiency of a new proposal. In this case, a popular Rate – Distortion (RD) performance assessment, the Bjøntegaard Delta (BD) [15] metric is computed. Since only EL is changed, the compression gain for EL frames is measured and presented in Table 2.

Table 2. BD-Rate Saving [%] with the proposed SHVC solution

Sequences		**Proposed SHVC vs. SHVC standard**	
		EL Only	**BL + EL**
Class D	RaceHorses	4.08	0.67
	BlowingBubbles	1.74	0.12
	BasketballPass	**4.54**	**0.89**
Class C	BasketballDrill	2.28	0.54
	PartyScene	1.98	0.29
Average		***2.92***	***0.50***

- As expected, the proposed JLP-SHVC always outperforms the SHVC standard with around 3% bitrate saving.
- The biggest gain, 4.5%, is from *BasketballPass* which typically associates with the high JLP frame quality as discussed in previous sub-section..
- Besides, though the compression improvement with JLP is lower than the proposed enhanced merger mode [10], the proposed DSFI is still important due to its block based architecture as well its easy integration.

IV. CONCLUSIONS

This paper proposes a novel low complexity SHVC improvement solution where decoder frame interpolation is created and exploited at enhancement layer, both at encoder and decoder sides. In contrast to the conventional SHVC standard, the novel JLP creation exploits only the decoded information; thus, no overhead bitrate is required to obtain the JLP frame without accessing to the original data. The proposed fusion model significantly improve the JLP frame quality. Experiments result have shown that the proposed JLP-SHVC framework significantly outperforms the traditional SHVC standard, notably by up to 4.5% bitrate saving while providing a similar joint layer frame quality. The future works may study better fusion model to create JLP frame with higher quality.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01- 2016.15.

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.,* pp. 1103-1120, vol. 17, no. 9, Sep. 2007.

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

[3] G. J. Sullivan, J. -R. Ohm, W. –J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1649-1668, vol. 22, no. 12, Dec. 2012.

[4] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable Extensions of the Highe Efficiency Video Coding Standard", *IEEE Trans. Circuits Syst. Video Technol.,* pp. 20-34, vol. 26, no. 1, Jan. 2016.

[5] X. HoangVan, J. Ascenso, and F. Pereira, "Designing an HEVC based scalable video coding extension", *Proc. of Conference on Telecommunications,* Castelo Branco, Portugal, May 2013.

[6] X. Li, J. Chen, K. Rapaka, and M. Karczewicz, "Generalized inter-layer residual prediction for scalable extension of HEVC," *in IEEE International Conference on Image Processing*, pp. 1559-1562, Melbourne, VIC, Australia, Sept. 2013.

[7] P. Lai, S. Liu, and S. Lei, "Combined temporal and inter-layer prediction for scalable video coding using HEVC," in *Picture Coding Symposium*, pp. 117-120, San Jose, CA, USA, Dec. 2013.

[8] M. Guo, S. Liu, and S. Lei, "Inter-layer adaptive filtering for scalable extension of HEVC," *in Picture Coding Symposium*, pp. 165-168, San Jose, CA, USA, Dec. 2013.

[9] P. Lai, S. Liu, and S. Lei, "Low latency directional filtering for inter-layer prediction in scalable video coding using HEVC*," in Picture Coding Symposium*, pp. 269-272, San Jose, CA, USA, Dec. 2013.

[10] X. HoangVan, J. Ascenso, and F. Pereira, "Improving enhancement layer merge mode for HEVC scalable extension," in *Picture Coding Symposium*, pp. 15-19, Cairns, QLD, Australia, Jun. 2015.

[11] Xiem HV, et al. "Improving scalable video coding performance with decoder side information", *Picture Coding Symposium,* San Jose, CA, USA, Dec. 2013.

[12] J. Ascenso, C. Brites, and F. Pereira, Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding, in: *Proceedings of the EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Slovak Republic, (2005).

[13] "Video test sequences," [Online]. Available: ftp://hevc@ftp.tnt.uni-hannover.de/testsequences/.

[14] SHVC Reference Software, version 12.1, https://hevc.hhi.fraunhofer.de/shm

[15] G. BjØntegaard, "Calculation of average PSNR differences between RD curves," Doc. VCEG-M33, 13th ITU-T VCEG Meeting, Austin, TX, USA, Apr. 2001.