SCALABLE EXTENSION OF HEVC USING ENHANCED INTER-LAYER PREDICTION

Thorsten Laude, Xiaoyu Xiu, Jie Dong, Yuwen He, Yan Ye, Jörn Ostermann**

InterDigital Communications, Inc., San Diego, CA, USA
* Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany

ABSTRACT

In Scalable High Efficiency Video Coding (SHVC), inter-layer prediction efficiency may be degraded because much high frequency information can be removed during: 1) the down-sampling/up-sampling process and, 2) the base layer coding/quantization process. In this paper, we present a method to enhance the quality of the inter-layer reference (ILR) picture by combining the high frequency information from enhancement layer temporal reference pictures with the low frequency information from the up-sampled base layer picture. Experimental results show that on average 3.9% weighted BD-rate gain is achieved compared to SHM-2.0 under SHVC common test conditions.

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

1. INTRODUCTION

The recent years have seen explosive growth of smart phones and tablets in terms of their screen resolution and computational capability. New video applications, such as video streaming and video conferencing, require video transmission in heterogeneous environments with different screen resolutions, computing capabilities and varying channel capacity. In these scenarios, scalable video coding can provide an attractive solution by coding different representations (temporal resolution, spatial resolution, fidelity, etc.) of a video into layers within one bitstream and providing the possibility to decode only a subset of these representations according to the specific device capabilities and/or available channel bandwidths.

Recently the Joint Collaborative Team on Video Coding (JCT-VC) of ISO/IEC MPEG and ITU-T VCEG developed the new video compression standard called High Efficiency Video Coding (HEVC) [1], which offers twice as much as the compression efficiency of the predecessor standard AVC [2], [3]. The first version of HEVC was finalized in January 2013. Like previous standards, HEVC is built upon the hybrid coding framework, thus motion compensated prediction followed by transform coding of the prediction error. Upon the completion of the single-layer HEVC, scalable extensions of the HEVC standard, called Scalable High Efficiency Video Coding (SHVC), are currently under development [4], [5]. The current SHVC design is built upon the high level syntax only based framework, where all the scalable coding technologies operate on the slice level, picture level, or above, whereas all block level operations of the enhancement layer (EL) remain identical to those of a single-layer HEVC codec. Compared to the simulcast solution that simply compresses each layer separately, SHVC offers higher coding efficiency by means of inter-layer prediction. In SHVC [4], inter layer prediction is implemented by inserting inter-layer reference (ILR) pictures, generated from reconstructed base layer (BL) pictures, into the EL decoded picture buffer for motion-compensated prediction of the collocated pictures in the EL. If the EL has a higher resolution than that of the BL, the reconstructed BL pictures need to be up-sampled to form the ILR pictures.

Given that the ILR picture is generated based on the reconstructed BL picture, its inter layer prediction efficiency may be limited due to the following reasons. Firstly, quantization is usually applied when coding the BL pictures. Quantization causes the BL reconstructed texture to contain undesired coding artifacts, such as blocking artifacts, ringing artifacts, and color artifacts. Such coding or quantization noise reduces the quality of the ILR pictures. Secondly, in case of spatial scalability, a down-sampling process is used to create the BL pictures. To reduce aliasing, the high frequency information in the video signal is typically removed by the down-sampling process. As a result, the texture information in the ILR picture lacks certain high frequency information; again this reduces the effectiveness of the ILR picture.

In contrast to the ILR picture, the EL temporal reference pictures contain plentiful high frequency information, which could be extracted to enhance the quality of the ILR picture. To further improve the ILR picture quality, a low pass filter may be applied to it to alleviate the quantization noise introduced by the BL coding process. In this paper, an enhanced ILR picture (EILR) is proposed. The proposed EILR picture is formed by combining the high frequency information extracted from the EL temporal reference pictures together with the low frequency information extracted from the conventional ILR picture. The proposed EILR picture can improve the inter layer prediction efficiency in SHVC. Experimental results using the Common Test Conditions of SHVC [6] showed that the

proposed method on average provides weighted BD-rate (BL+EL) gains of 3.4%, 3.4% and 5.0% for Random Access (RA), Low-delay B (LD-B), and Low-delay P (LD-P), respectively, in comparison to the performance of the SHVC reference software SHM-2.0 [5]. Other approaches to enhance the quality of the ILR pictures were previously proposed in [7] and [8] to restore high frequency information based on the differential signal between EL temporal and BL temporal reference pictures. The drawback of these methods is the necessity to access the BL temporal reference pictures; since the proposed method does not require such memory access, it incurs much lower increase in memory bandwidth and complexity compared to the previous methods. Additionally, other ILR enhancement approaches, such as inter-layer sample-adaptive offset, inter-layer filtering and cross-plane filtering, have also been studied in [9] and [10]. These methods only use the ILR picture and the collocated BL picture in inter layer processing; in contrast, by also using information from the EL temporal reference pictures, the proposed method more effectively improves the ILR quality and achieves higher coding efficiency.

The remainder of this paper is organized as follows. In Section 2, we present the proposed method, including how to generate the EILR picture and how to adaptively enable or disable the proposed method at picture level. We then present experimental results in Section 3. We conclude the paper in Section 4.

2. IMPROVED INTER-LAYER PREDICTION FOR SCALABLE VIDEO CODING

For ease of explanation, a scalable system with two layers (one BL and one EL) is used in this section to describe the proposed method, although it can be extended to a system with more than two layers.

2.1. Generation of the EILR picture

In [11], a method was proposed to reconstruct a skipped EL picture (a skipped EL picture contains no coded data) by applying motion compensation using the mapped BL motion information obtained through the motion field mapping process [12] in SHVC to the EL temporal reference pictures. In this paper, we refer to the picture generated according to [11] as the inter-layer motion compensation (ILMC) picture. The ILMC picture contains desirable high frequency information in the EL that is used to generate the proposed EILR picture. We next briefly review the ILMC picture.

For each block $B_{ILMC,t}(x,y)$ in the ILMC picture located at position (x,y) at time t, let (mvx, mvy) denote the mapped BL motion vector $MV_{IL,t-n}$ pointing to the reference picture at time $t-n$. When the corresponding BL block is uni-predicted, the block $B_{ILMC,t}(x,y)$ is generated by motion compensating the matching block in the EL temporal reference picture EL_{t-n} at time $t-n$ with (mvx, mvy), according to (1).

$$B_{ILMC,t}(x,y) = B_{EL,t-n}(x+mvx, y+mvy) \quad (1)$$

When the corresponding BL block is bi-predicted, the block $B_{ILMC,t}(x,y)$ is generated by combining two prediction components obtained from two EL temporal reference pictures. When the corresponding BL block is intra-coded, $B_{ILMC,t}(x,y)$ is directly copied from the collocated block in the conventional ILR picture.

Since the ILMC picture is built using EL texture information that contains high frequency information absent from the conventional ILR picture [4] (due to the down-sampling and up-sampling process), a high pass filter may be applied to extract such high frequency components from the ILMC picture to enhance the quality of the ILR picture. On the other hand, in order to reduce the quantization noise in the conventional ILR picture introduced by the BL coding, a low pass filter may be applied to the ILR texture samples. The proposed EILR picture is therefore generated by combining the low frequencies from the ILR picture and the high frequencies from the ILMC picture. Figure 1 shows how to generate the luma component of the EILR picture.

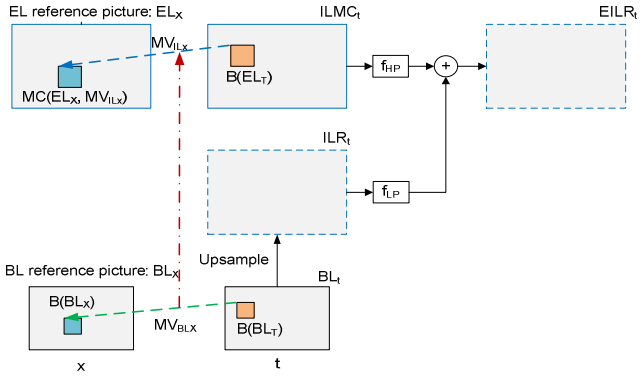

Figure 1 Generating the luma component of the EILR picture

At time t, denote the ILR picture as ILR_t and the ILMC picture as $ILMC_t$. The corresponding EILR picture $EILR_t$ is generated by applying a high pass filter f_{HP} to $ILMC_t$, a low pass filter f_{LP} to ILR_t, and then adding the two filtered signals as in (2), where ⊗ represents 2-D convolution.

$$EILR_t = f_{LP} \otimes ILR_t + f_{HP} \otimes ILMC_t \quad (2)$$

Unlike the luma component, the chroma components of the EILR picture are copied directly from the ILMC picture without additional filtering, because the filtering process in (2) incurs non-negligible computational complexity increase. Simulations show that, compared to the ILR picture, the chroma components in the ILMC picture contain more useful high frequency information. Therefore, directly copying the chroma components from the ILMC picture without the filtering process in (2) provides a good trade-off between performance and complexity.

The proposed EILR picture is added to the EL reference picture lists for EL coding in addition to the conventional

ILR picture, given that the ILR picture and the EILR picture have different characteristics. Specifically, if the EL slice is a P-Slice, the EILR picture is added as one additional reference picture after the ILR picture in reference list L0. If the EL slice is a B-Slice, the EILR picture is added as additional reference picture in both reference picture lists: in list L0 it is added after the ILR picture, whereas in list L1 it is added before the ILR picture.

2.2. Filter derivation process

The EILR picture is generated using two filters, f_{HP} and f_{LP}, as shown in (2). These two filters are derived jointly by optimizing both filters at the same time. The goal of the filter design is to derive the optimal filter coefficients f_{opt}, including the coefficients of the high pass filter and the low pass filter, which can minimize the distortion between the original EL picture $Org_{EL,t}$ and the EILR picture $EILR_t$. Non-separable 2-D filters are used in our design. The linear minimum mean square error (LMMSE) estimator method in [13] is used to derive the two filters with different characteristics that are applied to two different pictures, in order to reduce the distortion between the combination of these two filtered pictures and the original picture. Specifically, the optimal coefficients of f_{HP} and f_{LP} are jointly derived by solving the LMMSE problem formulated in (3), which minimizes the distortion between the EILR picture and the original EL picture.

$$f_{opt} = \arg\min\left[\left(f_{LP} \otimes ILR_t + f_{HP} \otimes ILMC_t - Org_{EL,t}\right)^2\right] \quad (3)$$

The filter coefficients derived from (3) are real numbers, and need to be quantized before they are signaled as part of the bitstream. Therefore, each real filter coefficient f_{float} is approximated by an integer, denoted as f_{int}. In this paper, a uniform quantizer is used to quantize the filter coefficients. The precision of the quantizer is chosen with respect to the dynamic range of the coefficients. In our implementation, each filter coefficient is represented by 6 bits. Thus, the dynamic range of the quantized filter coefficients is −32 to 31. In order to derive f_{float} from f_{int}, an additional factor k is used to make the integer filter coefficient approach the real valued filter coefficient, as shown in (4). Although k equals the quantization step in theory, the actual k may be slightly different from the quantization step due to the rounding operation of the quantization process. Therefore we calculate k as the inverse of the summation of all integer filter coefficients to ensure that the summation of the de-quantized filter coefficients is equal to one. Since k can be derived at the decoder, no signaling is needed. However, without any constraint k is a real number; this would in turn require floating-point multiplications to be used when generating the filtered samples in the EILR picture, thus severely increasing the computational complexity. In order to reduce the complexity of the filtering process, k is approximated by a computationally efficient multiplication with an integer number M followed by an N-bit shift to the right, as shown in (4).

$$f_{float} = f_{int} \times k \approx f_{int} \times \frac{M}{2^N} \quad (4)$$

The sizes of both low pass filter and high pass filter need to be determined. The filter size is proportional to the number of operations (multiplications and additions) and the overhead of transmitting the filter coefficients. A larger filter size can achieve a smaller distortion between the original EL picture and the EILR picture as shown in (3), which can translate to better inter layer prediction efficiency, but at the expense of increased computational complexity and increased overhead of representing the filter coefficients in the bitstream. Simulation results indicate that using a filter size of 3×3 for both filters provides a good trade-off between computational complexity, signaling overhead (108 bits per picture), and quality of the EILR picture. Therefore, it is adopted in our implementation. Simulations show that adaptive filters are more efficient than fixed filters due to a higher improvement of the EILR picture quality with minor signaling overhead.

2.3. Enabling/Disabling the Inter-Layer Reference Enhancement

The generated EILR picture is not always capable of improving the coding efficiency for all EL pictures, especially given the additional signaling overhead of filter coefficients. Therefore, we use a Lagrangian RD cost based method to adaptively enable/disable the use of the EILR picture on the picture/slice level. Specifically, the decision on whether to enable or disable the use of the EILR picture for a given EL picture is made based on comparing the RD costs of disabling the EILR picture (RD_{ILR}) and enabling the EILR picture (RD_{EILR}), according to equations (5) and (6), respectively.

$$RD_{ILR} = D_{ILR} \quad (5)$$

$$RD_{EILR} = D_{EILR} + \lambda \times R_f \quad (6)$$

where D_{ILR} and D_{EILR} denote the sum of squared errors (SSE) distortions between the conventional ILR and the proposed EILR picture and the original EL picture, respectively, R_f is the overhead of encoding the quantized filter coefficients, in number of bits, and λ is the Lagrangian weighting factor. In the proposed method, we use a 1-bit flag in the slice header to signal whether the proposed EILR picture is used or not. The proposed picture level selection method compares the approximate RD cost of using EILR and ILR pictures, but does not consider the actual RD cost of using these pictures at the prediction unit level.

3. EXPERIMENTAL RESULTS

The proposed EILR method is implemented based on the SHVC reference software SHM-2.0 [5] and evaluated under a comprehensive set of simulations as defined by the SHVC CTC [6]. The SHVC CTC define four coding structures,

seven test sequences, a set of four BL QPs in combination with two different delta QPs (ΔQP) for the EL, and three different spatial ratios between BL and EL (2x, 1.5x and SNR). Given that the proposed method requires the mapped BL motion information to generate the EILR picture, it is not applicable to the "All Intra" configuration, which does not utilize temporal prediction for the BL and EL pictures. To evaluate the proposed EILR method, the performance of SHM-2.0 is used as the anchor. Table 1 presents the BD-rate performance [14] between the proposed method and SHM-2.0 based on actually decoded bitstreams. The average coding gain of the proposed method varies between 1.1% and 6.3% for the luma component, and 4.5% and 11% for the chroma components. Taking into account that the chroma components have a smaller impact on the overall bit rate than the luma component, a weighted average is more accurate to measure the overall performance of the proposed method. As suggested in [15], the BD-rates of luma and chroma components are averaged as a single value by applying the weighting factors as $BD_{avg} = (6 \times BD_Y + BD_U + BD_V)/8$. The resulting overall weighted BD-rate (EL + BL) gains are 3.4%, 3.4% and 5.0% for RA, LD-B and LD-P, respectively. The encoding and decoding times are increased by 3% and 103%, respectively.

Table 1 BD-rate gain of the proposed method compared with SHM-2.0 (%). Class A: EL resolution 2560×1600. Class B: EL resolution 1920×1080.

	RA 2x			RA 1.5x			RA SNR		
	Y	U	V	Y	U	V	Y	U	V
Class A	2.5	8.5	8.3				3.6	10.1	10.7
Class B	1.1	4.9	5.7	2.3	6.8	8.0	1.4	8.1	10.0
Average	1.5	5.9	6.4	2.3	6.8	8.0	2.0	8.7	10.2
	LD-B 2x			LD-B 1.5x			LD-B SNR		
	Y	U	V	Y	U	V	Y	U	V
Class A	2.5	8.9	8.6				3.9	9.8	10.4
Class B	1.5	4.5	5.4	2.5	6.4	7.2	1.7	7.2	8.8
Average	1.8	5.8	6.3	2.5	6.4	7.2	2.3	7.9	9.3
	LD-P 2x			LD-P 1.5x			LD-P SNR		
	Y	U	V	Y	U	V	Y	U	V
Class A	3.2	9.0	8.6				6.3	10.3	11.0
Class B	3.2	5.0	5.9	3.7	6.7	7.6	4.4	9.0	10.7
Average	3.2	6.1	6.6	3.7	6.7	7.6	5.0	9.4	10.8

Figure 2 illustrates the RD-curves for the sequence *People on Street* under the LD-P configuration with SNR scalability, where the y-axis represents the average peak signal-to-noise ratio (PSNR) of the reconstructed EL pictures and the x-axis represents the overall bit rate of encoding the BL pictures and the EL pictures. In Figure 2, the solid line represents the RD-curve for the proposed EILR method and the dashed line represents the RD-curve for the SHM-2.0 anchor. It can be observed that the EILR method provides significant coding gains compared to the SHM-2.0 anchor for all the bit rates. The gain is larger at high bit rates, with PSNR improvement of up to 0.5dB. There are two reasons for better performance at higher bit rates: 1) relative overhead of signaling the filter coefficients is negligible at high bit rates; 2) the quality of the BL motion information and BL reconstructed texture (both of which are used to generate the proposed EILR picture) is better at higher bit rates.

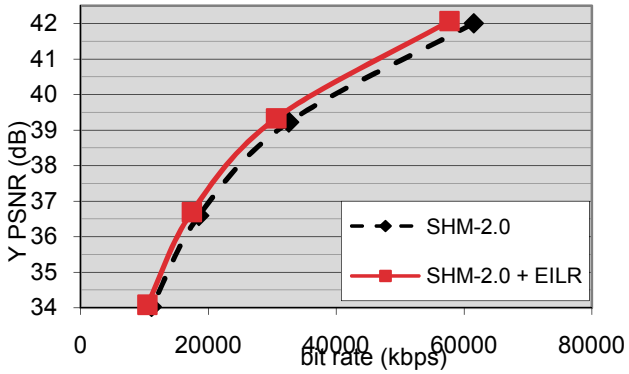

Figure 2 RD-curves of *People on Street*, where LD-P configuration and SNR scalability are applied with BL QP=26, ΔQP=−6

Figure 3 shows the exemplar frequency responses of the derived filters for the 128th frame of *People on Street* under the LD-P configuration with SNR scalability. The filters f_{HP} and f_{LP} derived by the LMMSE method according to (3) clearly show the characteristics of a low pass filter and a high pass filter, respectively. This verifies the fundamental assumption of the proposed method, which is inter layer prediction can be improved by combining the low frequency components from the ILR picture and the high frequency components from the ILMC picture.

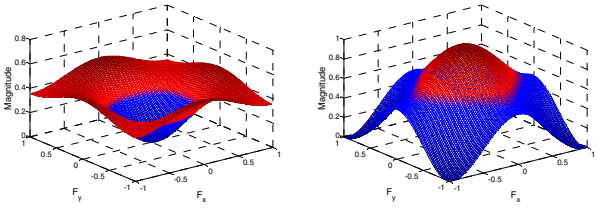

Figure 3 Frequency responses of the filter f_{HP} (left) and the filter f_{LP} (right)

4. CONCLUSION

In this paper, we present an improved inter-layer prediction method for SHVC. We combine the high frequency information from EL temporal reference pictures with the low frequency information from the inter-layer reference picture to generate the proposed enhanced inter-layer reference picture. Adaptive filters are derived to minimize the distortion between the enhanced inter-layer reference picture and the original EL picture and are signaled in the bitstream. Simulation results show that, under the SHVC common test conditions, the proposed method can achieve 3.9% weighted BD-rate (EL+BL) reduction on average compared to the SHM-2.0 anchors.

5. REFERENCES

[1] B. Bross, W-J. Han, J-R. Ohm, G. J. Sullivan, Y. K. Wang, T. Wiegand, "High Efficiency Video Coding (HEVC) Text Specification Draft 10," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-L1003, January 2013.

[2] P. Hanhart, M. Rerabek, F. De Simono, T. Ebrahimi, "Subjective Quality Evaluation of the upcoming HEVC Video Compression Standard," SPIE Optics and Photonics, in Proceedings of SPIE, vol. 8499, San Diego, August 2012.

[3] "Advanced Video Coding for generic audio-visual services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC), ITU-T and ISO/IEC JTC 1, May 2003.

[4] J. Chen, J. Boyce, Y. Ye, and M. M. Hannuksela, "SHVC Working Draft 2," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-M1009, April 2013.

[5] J. Chen, J. Boyce, Y. Ye, and M. M. Hannuksela, "SHVC Test Model 2 (SHM 2)," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-M1007, April 2013.

[6] X. Li, J. Boyce, P. Onno, Y. Ye, "Common SHM test conditions and software reference configurations," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-M1009, April 2013.

[7] Y. He, Y. Ye, X. Xiu, "ILR Enhancement with Differential Coding for SHVC Reference Index Framework," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-N0204, July 2013.

[8] A. Aminlou, J. Lainema, K. Ugur, M. Hannuksela, "Non-CE3: Enhanced inter layer reference picture for RefIdx based scalability," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-M0155, April 2013.

[9] J. Chen, A. Segall, E. Alshina, S. Liu and J. Dong, "SCE3: Summary Report of SHVC Core Experiment on Inter-layer Filtering." ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-N0033, July 2013.

[10] J. Dong, Y. Ye. Y. He, "Cross-Plane Chroma Enhancement for SHVC Inter-Layer Prediction," to appear in Picture Coding Symposium (PCS) 2013, December 2013.

[11] J. Boyce, X. Xiu, Y. Ye, "SHVC HLS: SHVC Skipped Picture Indication," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-N0209, July 2013

[12] X. Xiu, Y. Ye, Y. He and Y. He, "Inter-layer motion field mapping for the scalable extension of HEVC," Proc. SPIE 8666, Visual Information Processing and Communication IV, Feb. 2013.

[13] Y. Vatis, J. Ostermann, "Adaptive interpolation filter for H.264/AVC," IEEE Transactions on Circuits and Systems for Video Technology, Vol.19, No. 2, pp.179-192, Feb. 2009.

[14] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," document VCEG-M33, ITU-T SG16/Q6, Apr. 2001.

[15] G. J. Sullivan and J-R. Ohm, "Meeting report of the fourth meeting of the Joint Collaborative Team on Video Coding", ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-D500, January 2011.