MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Bit-depth Scalable Coding for High Dynamic Range Video

Shan Liu, Woo-Shik Kim, Anthony Vetro

TR2007-078 April 2008

Abstract

This paper presents a technique for coding high dynamic range videos. The proposed coding scheme is scalable, such that both standard dynamic range and high dynamic range representations of a video can be extracted from one bit stream. A localized inverse tone mapping method is proposed for efficient inter-layer prediction, which applies a scaling factor and an offset to each macroblock, per color channel. The scaling factors and offsets are predicted from neighboring macroblocks, and then the differences are entropy coded. The proposed inter-layer prediction technique is independent of the forward tone mapping method and is able to cover a wide range of bit-depths and various color spaces. Simulations are performed based on H.264/AVC SVC common software and core experiment conditions. Results show the effectiveness of the proposed method.

SPIE Conference on Visual Communications and Image Processing, January 2008

Bit-depth Scalable Coding for High Dynamic Range Video

Shan Liu[a], Woo-Shik Kim[b], and Anthony Vetro[a]
[a]Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA USA 02139;
[b]Dept. of Electrical Eng., Univ. Southern California, Los Angeles, CA USA 90089[*]

ABSTRACT

This paper presents a technique for coding high dynamic range videos. The proposed coding scheme is scalable, such that both standard dynamic range and high dynamic range representations of a video can be extracted from one bit stream. A localized inverse tone mapping method is proposed for efficient inter-layer prediction, which applies a scaling factor and an offset to each macroblock, per color channel. The scaling factors and offsets are predicted from neighboring macroblocks, and then the differences are entropy coded. The proposed inter-layer prediction technique is independent of the forward tone mapping method and is able to cover a wide range of bit-depths and various color spaces. Simulations are performed based on H.264/AVC SVC common software and core experiment conditions. Results show the effectiveness of the proposed method.

Keywords: video compression, bit-depth scalable, high dynamic range, tone mapping, H.264/AVC

1. INTRODUCTION

Video contents represented by eight bits per pixel have been well accepted by industry and widely used in numerous applications such as HDTV, DVD, mobile/internet video streaming or communication, camcorder, surveillance system, etc. The eight-bit-per-pixel based video codec system has an advantage that it can compactly represent one pixel as a byte in prevalent memory chips such as dynamic or static random access memory (RAM). However, the 256 levels in one byte may sometimes not be sufficient to cover the details of luminance and color information captured in the real world. Such a limited representation could be a problem for professional applications such as digital cinema, medical imaging, and post-production. Moreover, as display technology improves to cover wider dynamic range and true color representations, it becomes a problem for the consumer applications, too.

To cope with this problem, professional applications have been using high dynamic range (HDR) videos which consist of bit-depths greater than eight bits-per-pixel, e.g. ten, twelve, and fourteen, etc. In addition, the technique for properly displaying an HDR video on a standard dynamic range (SDR, i.e. eight bits-per-pixel) display device has been studied, which is referred as tone mapping. Due to the different capabilities of end display devices, an HDR video source is expected to be displayed in various bit-depths. Hence, a bit-depth scalable video coding solution[1] is required. An application example was given by Gao and Wu [2].

In previous work, Ward and Simmons[3] proposed a JPEG based scalable image coding system where SDR images and the pixel-by-pixel ratios between SDR and HDR images are coded, respectively. Here (and in rest of this article, if not specified) the SDR image contains the same content as its HDR counterpart, but in lower bit-depth, i.e. 8 bits-per-pixel. It is generated from the HDR image via tone mapping; and the process of reconstructing the HDR image from the SDR image is called inverse tone mapping, or inter-layer prediction in the context of a scalable coding architecture. The method proposed by Ward and Simmons[3] is simple and can work compatibly with various forward tone mapping algorithms. However, the pixel-based prediction results in substantial bit rate increment. Mantiuk et al.[4] adopted a Moving Picture Expert Group (MPEG) based system, which codes the difference between input and predicted HDR video signals, and at the same time, the SDR video for a backward compatibility. They predicted the HDR video from SDR input using a global inverse tone map, which may encounter loss of coding gain when localized tone mapping is used to generate the SDR input from the HDR video source.

During the international standardization work of the Joint Video Team (JVT) of ISO/IEC JTC1/SC29/WG11 (MPEG) and ITU-T SG16 Q.6 (Video Coding Expert Group; VCEG), some bit-depth scalable coding solutions have been

[*] This work was performed while Woo-Shik Kim was with Mitsubishi Electric Research Laboratories.

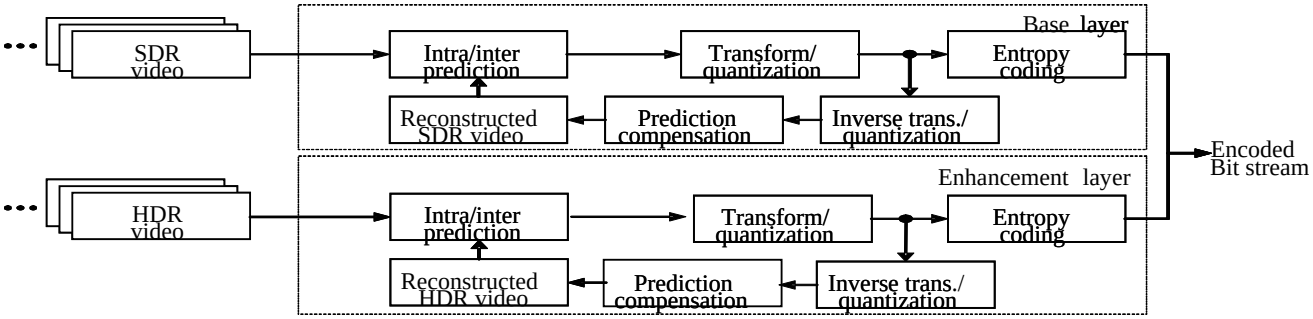
proposed. Winken et al.[5] presented a system which shared the similar philosophy as Mantiuk et al.[4] by using a global inverse tone map for inter-layer prediction; Gao and Wu[6] simplified the system by restricting the transform to a scaling of SDR video with a fixed factor over the sequence. These two methods inherit the disadvantage of Mantiuk et al's system[4], i.e. inefficiency for locally tone mapped videos. Segall and Su[8] proposed to use a scaling factor for each 4x4 pixel block and code the residues between the HDR and scaled SDR videos. This system overcomes the limit of previous works; however, the restrictions on scaling factors and chroma limit it from effectively covering a wider variety of dynamic ranges and color spaces.

In this paper, we propose a bit-depth scalable coding system, in which the spatial-adaptive inverse tone mapping method is used for inter-layer prediction from SDR base layer to HDR enhancement layer. The proposed method does not require any pre-knowledge of the forward tone mapping algorithm, nor metadata such as a tone map. The proposed approach is block-based, so that it works well with videos generated by localized tone mapping methods. With independent inverse tone mapping model for each color channel, the proposed method provides flexibility to applications in various color spaces such as XYZ, RGB, YUV444, etc. Furthermore, there are no constraints on the inverse tone mapping model parameters, thus it can cover a wide variety of dynamic ranges.

The rest of the paper is organized as follows. A brief overview of the reference bit-depth scalable coding system is given in Section 2. Section 3 illustrates the proposed inter-layer prediction scheme in detail. Section 4 provides experimental results with analysis and discussions. Section 5 concludes the paper.

2. BIT-DEPTH SCALABLE CODING SYSTEM OVERVIEW

The straightforward solution for including both HDR and SDR compressed representations in one bit stream is simulcast-like encoding, as shown in Fig. 1. That is, the SDR and HDR video inputs are compressed in base layer and enhancement layer independently, and then the bit streams from two layers are multiplexed to form a single output bit stream. The obvious advantage of this approach is its simplicity. However, the redundancies between two layers result in higher bit rate consumption. Moreover, the motion estimation and compensation processes bring extra computational complexity to the enhancement layer.



The bit-depth scalable video encoding system shown in Fig. 2 incorporates inter-layer prediction, where the HDR enhancement layer can be predicted from the SDR base layer. This significantly reduces the redundancies between two layers, and thus bits are saved. In fact, the enhancement layer can still be predicted from intra and inter predictions, as indicated by the switch in Fig. 2, which allows the enhancement layer to take advantages from intra prediction, motion compensation and inter-layer prediction. Similar to the intra/inter prediction residues, the inter-layer prediction residues

are calculated by subtracting the predicted HDR video from the original HDR video. These residues are then coded into bit stream via transform, quantization, and entropy coding. The bit streams from both layers are multiplexed to generate the scalable bit stream.

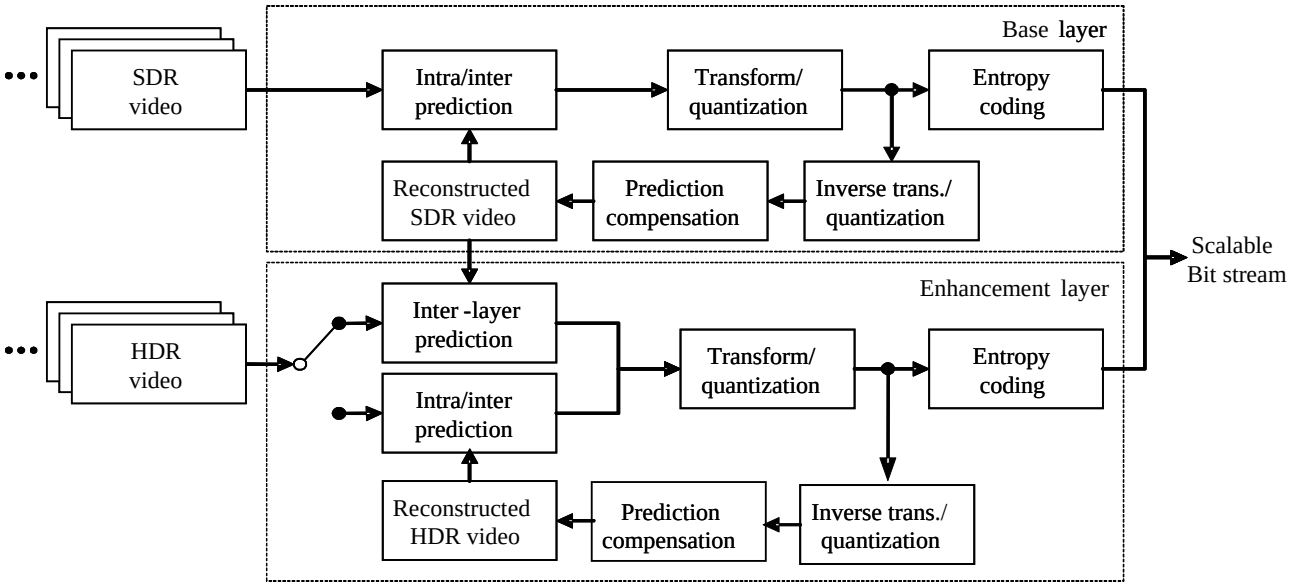


Fig. 2. Bit-depth scalable encoding system with motion compensation in enhancement layer.

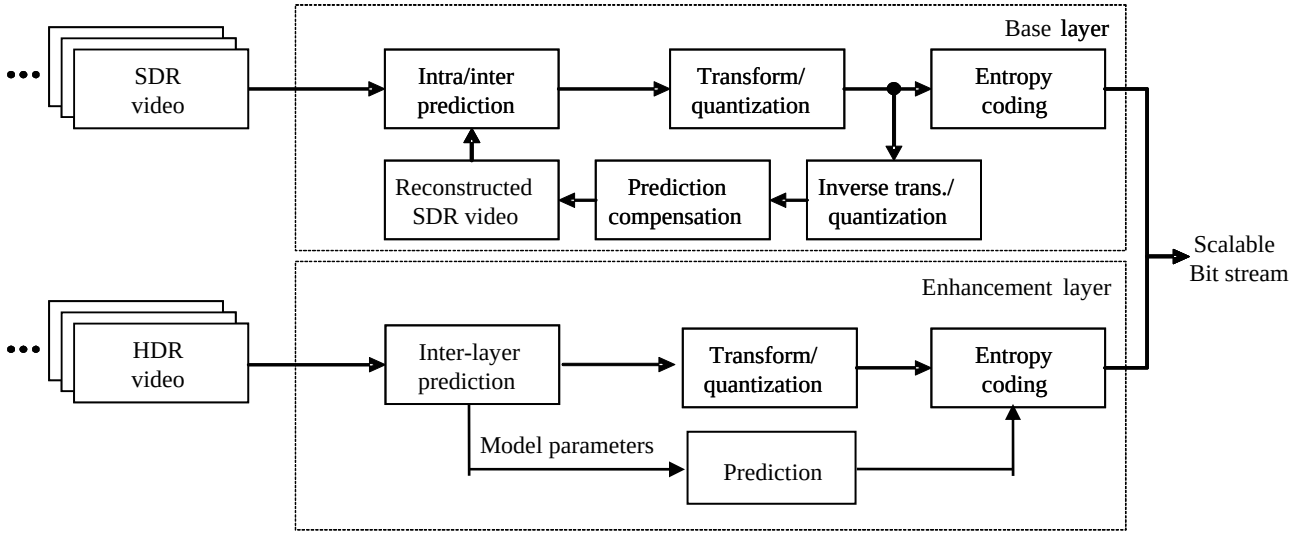


Fig. 3. Bit-depth scalable encoding system with motion compensation in base layer only.

The encoding system shown in Fig. 2 follows the coarse-grain scalability (CGS) coding approach of SVC[1] which requires motion compensation in the enhancement layer. If inter-layer predictions are restricted, single-loop decoding is

still possible; however, the implementation of motion compensation at higher bit-depths in the enhancement layer will obviously incur higher complexity, and might not be suitable for consumer electronics devices. Therefore, we adopt the bit-deth scalable coding architecture shown in Fig. 3, which utilizes motion compensation in the base layer only and inter-layer prediction for the enhancement layer. Intra and motion-compensated inter predictions are eliminated in the enhancement layer, which greatly reduces the encoding and decoding complexity. In the meantime, the structure is flexible and can integrate with intra and inter predictions easily.

The inverse tone mapping is performed in the inter-layer prediction process to make a prediction of the HDR video samples from the reconstructed SDR video samples. Different from forward tone mapping, which emphasizes the improvement of subjective SDR quality, the main purpose of inverse tone mapping is to improve the prediction performance, i.e., to reduce the numerical distance between the predictor and the original HDR video. The proposed inverse tone mapping for the inter-layer prediction is described in the following section.

3. PROPOSED INTER-LAYER PREDICTION SCHEME

3.1 Inverse tone mapping model

In this paper, we propose a linear inverse tone mapping method to perform the inter-layer prediction described in the previous section. The proposed inverse tone mapping is block-based, thus it works well with localized forward tone mapping methods, as well as global ones. It does not require any pre-knowledge of the forward tone mapping method used to generate the SDR input, nor any training for generating tone maps such as in [5]; thus it is flexible and of low-complexity. The proposed inverse tone mapping model is independent for each color channel. Therefore, it promotes compression quality of chroma, as shown in experimental results in the next section, and is especially suitable for some professional applications which utilize formats such as RGB, YUV444, etc.

The proposed inverse tone mapping model is illustrated in (1) where $\hat{p}_{HDR}$ is the predicted HDR pixel value, p_{SDR} is the reconstructed base layer SDR pixel value, SF is the macroblock-based scaling factor, and Os is the macroblock-based offset. For each macroblock, each color channel (e.g. Y, Cb, Cr; or R, G, B; etc.) one set of model parameters is assigned, which consists of the scaling factor, offset, and a prediction mode. The prediction mode is used to indicate the prediction direction of the scaling factor and the offset. The scaling factor and the offset of the current macroblock can be predicted from either its top or its left neighbors, if applicable, depending on which neighbor generates the minimum offset. The prediction and reconstruction of the scaling factor are illustrated in (2) and (3); those of the offset are given in (4) and (5).

$$\hat{p}_{HDR} = p_{SDR} \times SF_{curr} + Os_{curr}, \tag{1}$$

$$SF_{curr} = SF_{pred} + SF_{diff}, \tag{2}$$

$$SF_{pred} = (pred == ABOVE)\,?\,SF_{above} : SF_{left}, \tag{3}$$

$$Os_{curr} = Os_{pred} + Os_{diff}, \tag{4}$$

$$Os_{pred} = (pred == ABOVE)\,?\,Os_{above} : Os_{left}, \tag{5}$$

3.2 Estimation of model parameters

We first determine the scaling factor among the three parameters. The straightforward approach is to search for the optimal scaling factor in a range, which yields the minimum R-D cost. We adopted the calculation of R-D cost and the scaling factor search range suggested in the common software used by [5-11]. This approach guarantees to find the optimal scaling factor in a given range, e.g. [1,21] for 10-bit and [1,32] for 12-bit enhancement layer. However, the searching process is time consuming. In order to expedite the process, we explore a fast search scenario. Instead of sequentially scanning all the scaling factor candidates, the search starts from the SF_{start}, as defined in (6). The R-D cost of using SF_{start} is calculated, and then compared with the R-D costs of using $SF_{start} \pm 1$. If the R-D cost of $SF_{start} + 1$

or $SF_{start}-1$ is smaller, the R-D costs of $SF_{start}\pm 2$ are calculated and compared; so on and so forth, until the R-D cost stops decreasing, as in (7).

$$SF_{start} = 2^{(bit\_depth\_HDR - bit\_depth\_SDR)}, \tag{6}$$

$$Cost(SF_{start\pm k}) \le Cost(SF_{start\pm k\pm 1}), \tag{7}$$

The drawback of using fast search is that it may fall to a local minimum, which results in inaccurate prediction, then greater residues and thus degraded coding efficiency. However, in the number of simulations we performed, the fast search found the same optimal scaling factors as the exhaustive search. Hence, it is used to generate the experimental results shown in the next section.

Given the optimal scaling factor SF_{opt}, the offset is calculated as in (8),

$$Os = \frac{1}{N}\sum_{i=1}^{N}(p_{HDR} - p_{SDR} \times SF_{opt}), \tag{8}$$

where N denotes the total number of pixels in the block, e.g. 256 for Y and 64 for Cb, Cr in YUV420 format; p_{HDR} and p_{SDR} are values of enhancement layer HDR and base layer SDR pixels.

Another alternative method for estimating scaling factors and offsets, without searching, is to use the least squares (LS) fitting[12]. Let X be the pixel values of SDR video in the given region, i.e. macroblock in this article, and Y be those of HDR video. The matrix notation is,

$$Y = X \times SF + Os, \tag{9}$$

The LS solution for the scaling factor and the offset can be obtained by

$$\hat{SF} = \left(X^T X\right)^{-1} X^T Y, \tag{10}$$

$$\hat{Os} = \mu_Y - \hat{SF} \times \mu_X, \tag{11}$$

with μ_X and μ_Y denoting the means of X and Y, respectively.

3.3 Coding of model parameters

If the video is represented in N color channels (e.g. N=3 for YUV and RGB) then for each macroblock, there are N sets of model parameters to be conveyed through the bit stream. This could result in considerable cost of bit rate if the model parameters are not properly compressed. Here, we apply predictive and entropy coding to these model parameters. As illustrated in (2)-(5), the scaling factor and offset of the current macroblock are predicted from their top or left neighbors, and only the prediction differences, SF_{diff} and Os_{diff} are entropy coded and written into the bit stream. Realizing that the range of the offsets is normally much greater than that of the scaling factors, which results in higher bit usage, the prediction direction is determined by which generates smaller Os_{diff}. The scaling factor follows the same prediction direction as the offset. The prediction direction is also written into the bit stream, which costs no more than 1 bit per macroblock.

It is also observed that the offsets can be quantized to some extent before prediction and entropy coding with little sacrifice in coding quality and efficiency. This is because that the inter-layer prediction residues are quantized in frequency domain, such that the accuracy of the offsets may be a waste, especially in low bit rates. Thus, we down-scale the offsets by a right shift as shown in (12); and apply a reverse left shifting process when computing the inter-layer prediction residues, as well as in decoder, as shown in (13).

$$Os_{enc} = Os >> [(bit\_depth\_HDR - bit\_depth\_SDR)/2], \tag{12}$$

$$Os_{rec} = Os << [(bit\_depth\_HDR - bit\_depth\_SDR)/2], \quad (13)$$

As previously mentioned, the proposed block-based inter-layer prediction method accommodates well with localized forward tone mapping methods. In case that the SDR base layer is generated by global tone mapping methods, the model parameters for each macroblock consume overhead bits and may degrade coding efficiency, compared with some global inverse tone mapping solutions such as [5]. However, on the other hand, in this case the correlations among neighborhood model parameters are stronger, so that the predictive and entropy coding can make up the coding efficiency loss. Experimental results of both global and local tone mapped sequences are given and discussed in the next section.

4. EXPERIMENTAL RESULTS

In this section, we show the experimental results of the proposed inter-layer prediction technique for bit-depth scalable coding. The proposed method is implemented based on the Joint Scalable Video Model version 8.12 developed by JVT. We followed the experimental conditions defined by JVT[14], where the SDR base layer is in eight bits-per-pixel and HDR enhancement layer is in ten and twelve bits-per-pixel, respectively. We generated results in both 4CIF, i.e. 704×576, and HD, 1920x1088 resolutions; in YCbCr 4:2:0 formats. The model parameters are determined by fast search.

Fig. 4 and Fig. 5 show the luminance rate-distortion performance of the proposed method in comparison with simulcast, 10-bit single layer encoding and scalable coding from related works, e.g. Winken et al.[5] and Segall and Su[8]. Note that all scalable solutions are single-loop, i.e. without motion-compensated inter prediction as well as intra prediction in the enhancement layer. It is shown that for the SVT sequence "InToTree", all scalable coding methods achieve very similar R-D performance, which is a bit worse than single-layer encoding, but significantly outperforms simulcast. This is because that SVT sequences use a simple bit-shifting tone mapping method to generate the 8-bit base layer from the 10-bit HDR video input. Thus, both global inverse tone map[5] and localized inverse tone mapping models[8,10] can provide an accurate approximation.

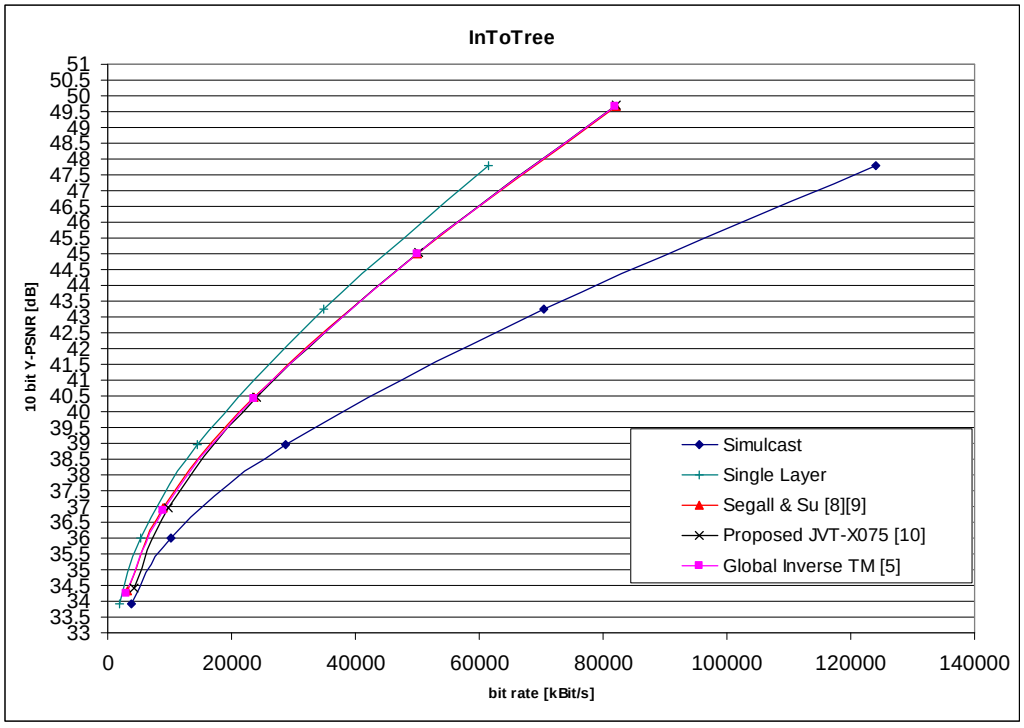


Fig. 4. Experimental results, Luminance, SVT sequence "InToTree", 8-bit base layer, 10-bit enhancement layer, 50Hz, 60 frames, 4CIF.

The Viper sequence “Waves” uses a non-linear tone mapping technique to generate the 8-bit base layer from the 10-bit HDR video input. Consequently, the three inverse tone mapping methods produce different results. It is shown that the global inverse tone mapping method[5] achieves the best performance. The proposed inter-layer prediction method losses about 0.2dB due to the overhead bits spent on coding the model parameters. The method proposed by Segall and Su[8] shows the lowest performance among three, possibly due to the constraints set on their inverse tone mapping model parameters. Overall, all three scalable solutions outperform simulcast.

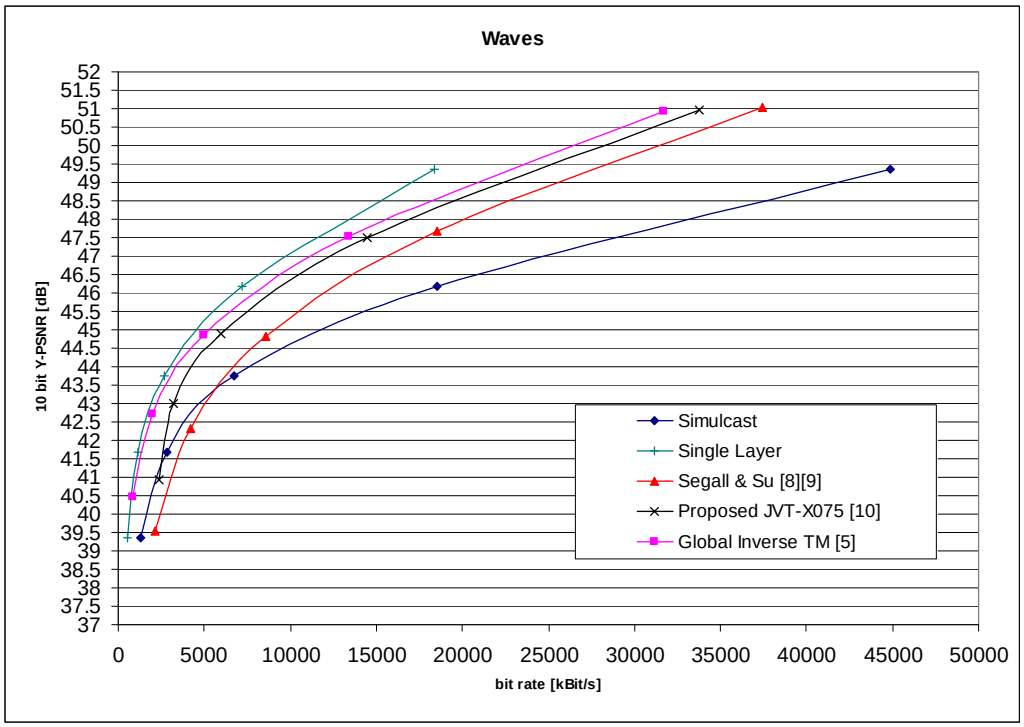
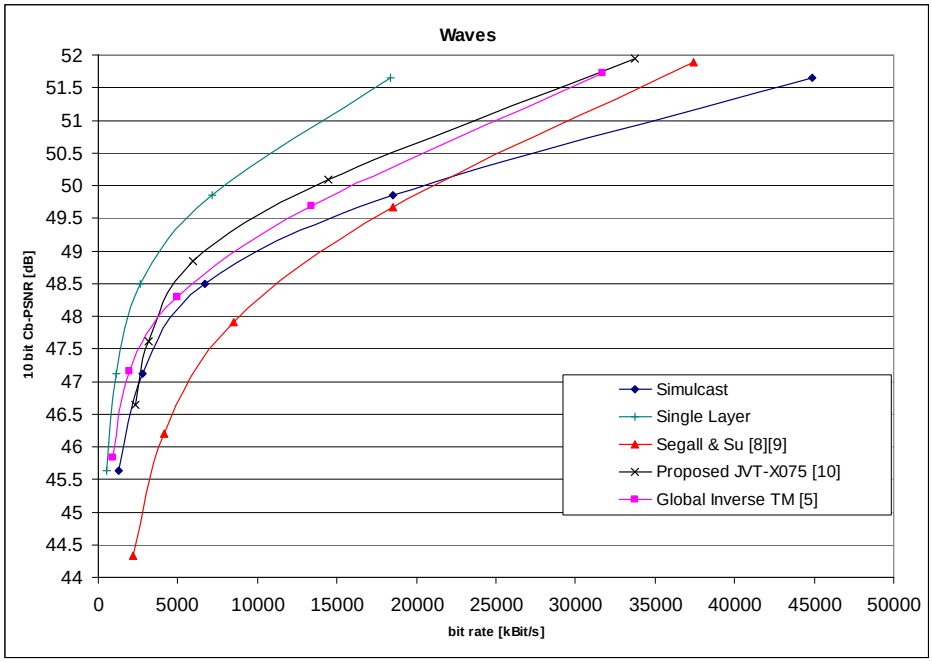


Fig. 5. Experimental results, Luminance, Viper sequence “Waves”, 8-bit base layer, 10-bit enhancement layer, 50Hz, 60 frames, 4CIF.

Both Segall and Su[8] and the proposed inter-layer prediction schemes are block-based. One major difference of these two approaches is that they use one inverse tone mapping model for both Chroma (Cb and Cr), while we keep the two sets of parameters independent. Fig. 6 demonstrates the chroma R-D curves of these two methods compared to the global inverse tone mapping method[5], simulcast and single layer coding results as references. It is clearly shown that our proposed scheme outperforms both Segall and Su[8] and the global inverse tone mapping method[5] in Cb and Cr, which compensates the coding efficiency loss in low bit rates. Considering both testing sequences are in YUV420 format, we may expect more chroma coding gain for professional formats such as YUV444, etc.

Fig. 7 and Fig. 8 show the experimental results with 12-bit HDR video sources. Both sequences “Library” and “Sunrise” have an 8-bit base layer and 12-bit enhancement layer, in 1920x1088 HD resolutions. Coding gains are achieved by using our proposed method, compared with global inverse tone mapping[5] for both luminance and chrominance. It is observed that the performance of the global inverse tone mapping method drops as the bit-depth increases. This can be due to the limitation of the multiple-to-one look-up table used for global tone mapping and its inverse.



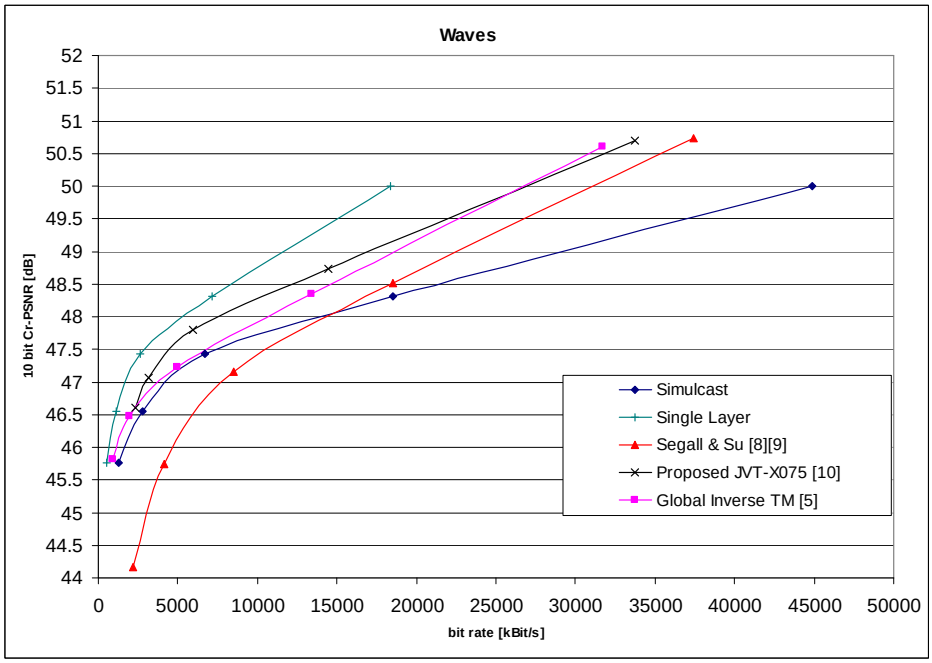


Fig. 6. Experimental results, Chroma, Viper sequence “Waves”, 8-bit base layer, 10-bit enhancement layer, 50Hz, 60 frames, 4CIF.

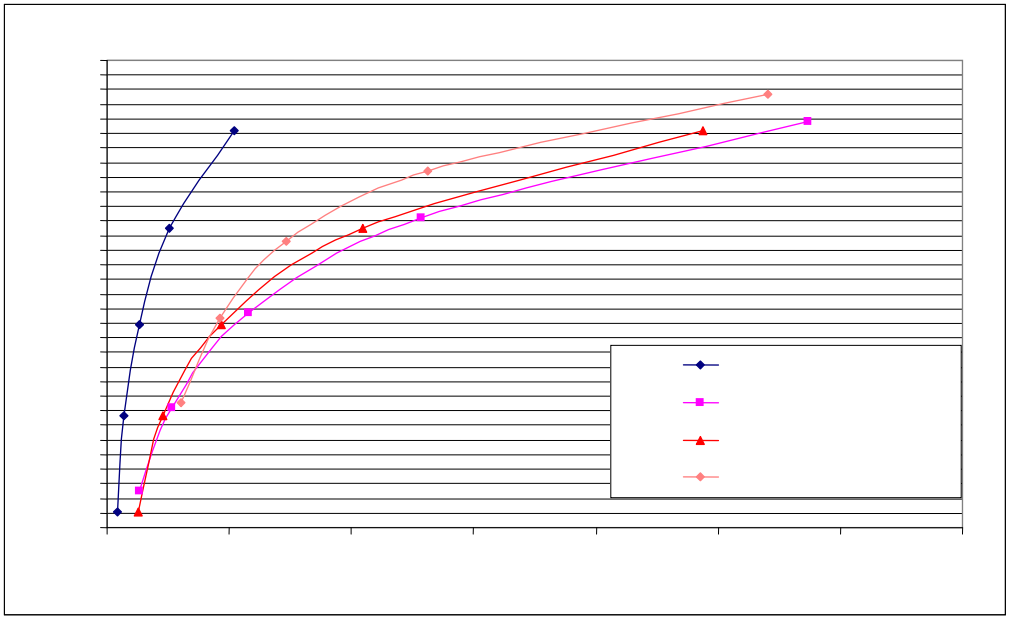


Fig. 7. Experimental results, 12 bit sequence “Library”, 8-bit base layer, 12-bit enhancement layer, 50Hz, 60 frames, 1920x1088.

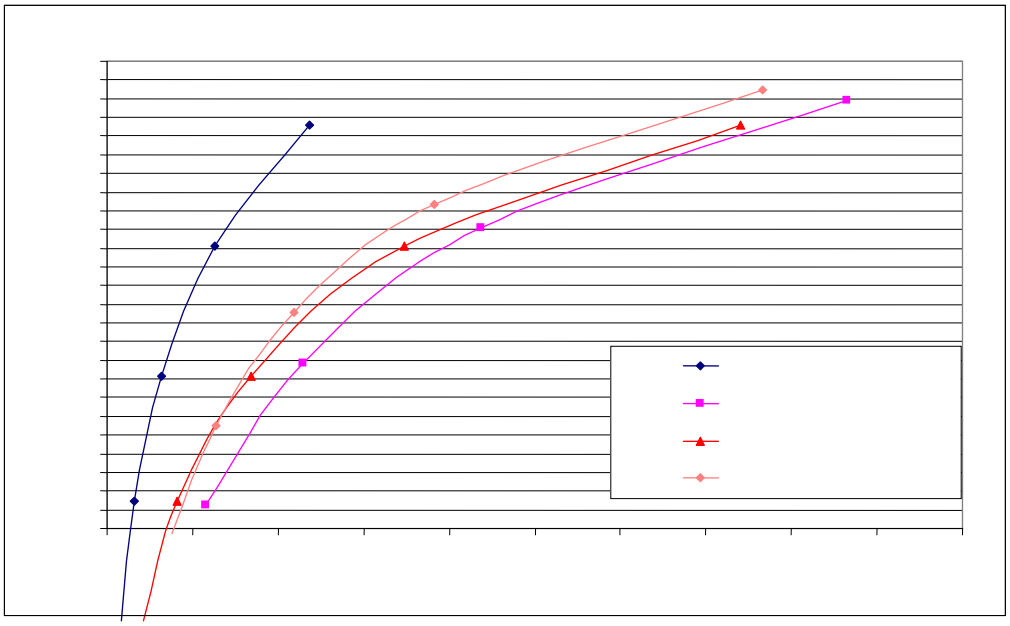


Fig. 8. Experimental results, 12 bit sequence “Sunrise”, 8-bit base layer, 12-bit enhancement layer, 50Hz, 60 frames, 1920x1088.

5. CONCLUSION

In this paper we proposed an inter-layer prediction technique for bit-depth scalable video coding. A block-based inverse tone mapping model is utilized to perform this prediction, which is independently derived for each color channel. The proposed solution is low-complexity and general, thus can be applied to various video inputs generated with different forward tone mapping algorithms. It is adaptive to a wide range of bit-depths and a variety of color spaces. The experimental results show the efficiency of the proposed method.

REFERENCES

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

[2] Y. Gao and Y. Wu, "Applications and requirement for color bit depth scalability," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-U049, Oct. 2006.

[3] G. Ward and M. Simmons, "JPEG-HDR: a backwards-compatible, high dynamic range extension to JPEG," *Proc. 13th Color Imaging Conf.*, 283-290, Nov. 2005.

[4] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel, "Backward compatible high dynamic range MPEG video compression," *Proc. ACM SIGGRAPH 2006*, 713-723, Jul. 2006.

[5] M. Winken, H. Schwarz, D. Marpe, and T. Wiegand, "SVC bit depth scalability," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-V078, Jan. 2007.

[6] M. Winken, D. Marpe, H. Schwarz, and T. Wiegand, "Bit-depth scalable video coding," *Proc. ICIP*, Volume 1, pp. 5-8, Sept. 16 2007-Oct. 19 2007.

[7] Y. Gao and Y. Wu, "Bit depth scalability," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-V061, Jan. 2007.

[8] A. Segall and Y. Su, "System for bit-depth scalable coding," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-W113, Apr. 2007.

[9] A. Segall, "Scalable Coding of High Dynamic Range Video," *Proc. ICIP*, Volume 1, pp. 1-4, Sept. 16 2007-Oct. 19 2007.

[10] S. Liu, W. Kim and A. Vetro, "Inter-layer prediction for bit-depth scalability," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-X075, Jul. 2007.

[11] A. Segall and Y. Su, "CE1: Inter-layer prediction for SVC bit-depth scalability," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-Y071, Oct. 2007.

[12] E. Weisstein, "Least Squares Fitting – Polynomial," *MathWorld – A Wolfram Web Resource*, available at http://mathworld.wolfram.com/LeastSquaresFittingPolynomial.html.

[13] A. Segall, "CE1: Bit-depth Scalability," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6*, Document JVT-X301, Jul. 2007.