ENHANCED CTU-LEVEL INTER PREDICTION WITH DEEP FRAME RATE UP-CONVERSION FOR HIGH EFFICIENCY VIDEO CODING

Lei Zhao [†], *Shiqi Wang* [‡], *Xinfeng Zhang* [⋆], *Shanshe Wang* [†], *Siwei Ma* [†] *and Wen Gao* [†]

[†] Institute of Digital Media & Cooperative Medianet Innovation Center, Peking University, Beijing, China
[‡] Department of Computer Science, City University of Hong Kong, Hong Kong
[⋆] University of Southern California, Los Angeles, California, USA
[†] {lei_zhao, sswang, swma, wgao}@ pku.edu.cn, [‡] shiqwang@cityu.edu.hk [⋆] xinfengz@usc.edu

ABSTRACT

Inter prediction serves as the foundation of prediction based hybrid video coding framework. The state-of-the-art video coding standards employ the reconstructed frames as the references, and the motion vectors which convey the relative position shift between the current block and the prediction block are explicitly signalled in the bitstream. In this paper, we propose a high efficient inter prediction scheme by introducing a new methodology based on virtual reference frame, which is effectively generated with the deep neural network such that the motion data does not need to be explicitly signalled. In particular, the high quality virtual reference frame is generated with the deep learning based frame rate up-conversion (FRUC) algorithm from two reconstructed bi-prediction frames. Subsequently, a novel CTU level coding mode termed as *direct virtual reference frame* (DVRF) mode, is proposed to adaptively compensate for the current to-be-coded block in the sense of rate-distortion optimization (RDO). The proposed scheme is integrated into the HM-16.6 software, and experimental results demonstrate significant superiority of the proposed method, which provides more than 3% coding gains on average for HEVC test sequences.

Index Terms— Inter prediction, virtual reference frame, deep learning, video coding

1. INTRODUCTION

As the state-of-the-art video coding standard, High Efficiency Video Coding (HEVC)[1][2] adopts block based hybrid video coding framework, including block based intra/inter prediction and transform. The inter prediction, which aims to remove the temporal redundancy, serves as an indispensable part of the coding framework. In particular, inter prediction makes use of the temporal correlation between pictures to obtain the predicted version of the current to-be-coded block. HEVC employs multiple decoded frames as references, such that the motion information including motion vector (MV) and reference frame index are required to be signalled in the bitstreams in order to specify the predicted blocks. Generally speaking, motion data accounts for a large proportion of the total bitstream. In order to represent the motion data in a more efficient way, HEVC adopts two coding modes, i.e., advanced motion vector prediction (AMVP) mode and merge mode. More specifically, when AMVP mode is chosen, the reference index, motion vector predictor (MVP) and motion vector difference (MVD) need to be coded in the bitstream to restore the motion data. Regarding the merge mode, only merge index is needed to reuse the motion data from neighboring blocks. Although AMVP and merge modes have effectively reduced the bitrate consumptions during motion data coding process, there is still a heavy burden to signal motion information in the bitstreams, especially in low bitrate coding scenarios. As such, there is a high demand to generate a high quality virtual reference frame which can be directly used to predict the current one. In this manner, the motion information can be implicitly conveyed in the virtual reference frame generation process with the reconstructed reference frames.

The *Hierarchical B Coding Structure* is adopted in HEVC, and in comparison with the classical B picture coding, the coding efficiency can be improved by up to 1.5 dB [3]. A typical hierarchical B structure with 4 temporal levels in Random Access (RA) configuration is depicted in Fig.1, where I0 and B8 belong to temporal level_0, which provide high quality reference for subsequent frames. Once frames in level_0 are reconstructed, level_1 frame B4 can be bi-predicted by I0 and B8. Regarding level_2 frames B2 and B6, both reconstructed frames of level_0 and level_1 can be used as references. Besides, level_3 contains B1, B3, B5 and B7 which reference all the three lower level frames. Generally speaking, each B picture can be predicted using the nearest pictures of the lower temporal levels in forward and backward directions. Furthermore, as the temporal distance between two reference frames is getting closer for higher level frames, the prediction of the intermediate frame becomes more reliable.

Based on the hierarchical B coding structure, we further investigate the generation of the virtual reference frame that can be directly applied for inter prediction without the de-

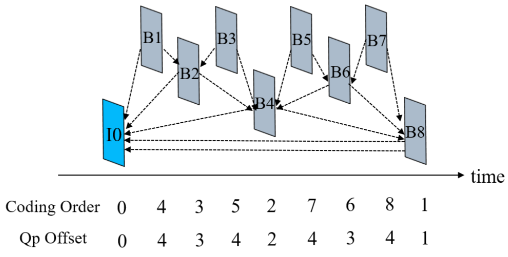


Fig. 1. Hierarchical B coding structure in HEVC.

mand of motion information signalling. In particular, based on the natural relationship between frame rate up-conversion (FRUC) and hierarchical B frame prediction, we propose to generate high quality virtual reference frame for high level B frames with the-state-of-the-art deep FRUC approach Adaptive Separable Convolution [4]. In order to take fully advantage of the virtual reference frame, a novel CTU level coding mode *direct virtual reference frame* (DVRF) mode is designed to adaptively select the best reconstruction method in the sense of rate-distortion optimization (RDO). Experimental results on HM-16.6 verify the performance of the proposed scheme on HEVC test sequences.

The rest of this paper is organized as follows. Section 2 reviews the Adaptive Separable Convolution algorithm. Our proposed scheme for better coding efficiency is presented in Section 3. Experimental results and analyses are given in Section 4, and finally the paper is concluded in Section 5.

2. REVIEW OF DEEP LEARNING BASED FRUC

The rapid development of deep learning has greatly facilitated the development of FRUC algorithms. Several works have been proposed to explore deep learning for better FRUC performance. In particular, Zhou *et al.* [5] trained a convolutional neural network (CNN) to predict appearance flows, which was then used to reconstruct the target view. In [6], the deep voxel flow approach generated dense voxel flows to optimize frame interpolation results with deep neural network. Among the existing deep learning based methods, Adaptive Separable Convolution shows considerable superiority in term of both interpolation quality and complexity cost.

Traditional FRUC methods interpolate the target frame in two steps: dense motion estimation and pixel interpolation. Niklaus *et al.* [7] formulated pixel interpolation as a local convolution process over patches in the input images, and provided an Adaptive Convolution approach which combines motion estimation and pixel synthesis in a single step. As shown in Fig. 2, for each individual output pixels *(x,y)*, the deep convolutional network takes receptive field patches $R_1(x, y)$ and $R_2(x, y)$ as input, and outputs a convolution kernel with size of N × N. This kernel then convolves with two patches P_1 and P_2 centered at *(x,y)* to produce the target pixel. A major drawback of Adaptive Convolution lies in its memory cost. To generate the kernels for all pixels in a 1080p video frame, the output kernels alone will require 26 GB of memory, which makes it impractical in real time applications. In view of this, Niklaus *et al.* [4] then proposed Adaptive Separable Convolution algorithm which approximates 2D convolution kernels with a pair of 1D kernels. In this way, an N × N convolution kernel can be encoded using only 2N variables, showing considerable superiority than 2D convolution version. In this paper, Adaptive Separable Convolution is employed to generate the virtual reference frame due to its superior performance.

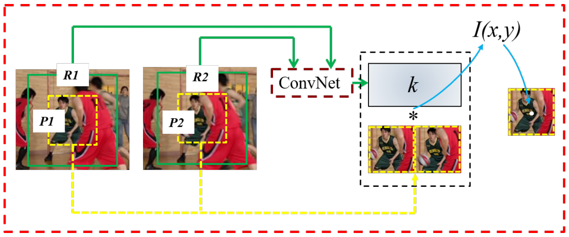


Fig. 2. Illustration of the Adaptive Convolution [7] method.

3. THE PROPOSED SCHEME

In order to improve the coding efficiency based on HEVC, we propose to generate the virtual reference frame for B frame prediction, such that the motion information is implicitly encoded into the virtual reference frame during the FRUC process. More specifically, for B frame compression, we first generate a high quality reference frame through Adaptive Separable Convolution algorithm. Subsequently, a CTU level coding mode is devised to adaptively select the interpolated frame as reconstructions. The details of these two processes are elaborated as follows.

3.1. Virtual Reference Frame Generation

In general, both FRUC and B frame prediction target for interpolating the intermediate frame with the frames in forward and backward directions. Ideally, if FRUC provides comparable interpolation quality as normal B frame motion prediction process, the signalling of additional MV information is not necessary. Since deep learning approach has achieved superior performance in various video processing tasks, we propose to adopt the FRUC method Adaptive Separable Convolution in the virtual frame generation, which can be utilized in the future inter prediction process.

As shown in Fig. 3, assuming that B1 is the frame to be compressed, and given two reconstructed frames of I0 and B2, a virtual reference frame $\hat{B}1$ can be directly synthesised with deep learning based FRUC [4]. In particular, let $F(\cdot)$ denote FRUC process, the derivations of the predictions for level_3 frames in this group of pictures (GOP) can be expressed as

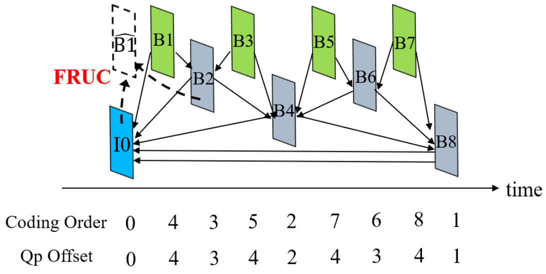


Fig. 3. Illustration of generating virtual reference frame with hierarchical B coding structure.

follows:

$$\begin{cases} \hat{B}1 = F(Rec(\mathrm{I0}), Rec(\mathrm{B2})) \\ \hat{B}3 = F(Rec(\mathrm{B2}), Rec(\mathrm{B4})) \\ \hat{B}5 = F(Rec(\mathrm{B4}), Rec(\mathrm{B6})), \\ \hat{B}7 = F(Rec(\mathrm{B6}), Rec(\mathrm{B8})) \end{cases} \quad (1)$$

where $Rec(\cdot)$ represents reconstructed frame, and $\hat{B}1$, $\hat{B}3$, $\hat{B}5$ and $\hat{B}7$ denote the generated virtual reference frames.

An intuitive idea to take advantage of the virtual reference frames is directly applying them as the reconstructed frames. However, this strategy is less efficient since it is still an open problem to interpolate the intermediate frames in complex motion scenarios. In order to adaptively enable the usage of the virtual reference frame, an adaptive mechanism in the sense of rate-distortion optimization (RDO) is developed, which will be presented in the next subsection.

3.2. CTU-Level Direct Virtual Reference Frame Mode

In order to take advantage of the virtual reference frame in a more efficient way, we propose a novel CTU level coding mode-*direct virtual reference frame (DVRF)* mode . The proposed method selects the virtual reference frame as the reference in a RDO manner, such that the RD performance is consistently improved with the expense of mode flag signalling. Compared with the traditional inter coding methods in HEVC, the proposed DVRF mode does not need to signal any motion information, and the residuals data are also skipped without transform, quantization and entropy coding.

3.2.1. DVRF Mode

Here, it is worth mentioning that the proposed DVRF coding mode applied to B frames with bi-predicted references, such that the application is restricted to the Random Access (RA) configuration. Specifically, the DVRF mode is enabled as a potential coding mode when the current frame has bi-directional reference, and the temporal distances between the current to-be-coded frame and the two references are identical.

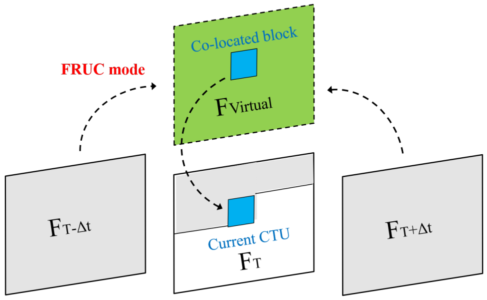


Fig. 4. Illustration of the proposed DVRF coding mode. The co-located block in the virtual reference frame is directly copied to the current to-be-coded CTU.

As is shown in Fig. 4, for the current to-be-coded frame F_T, $\mathrm{F}_{T-\Delta t}$ and $\mathrm{F}_{T+\Delta t}$, which are the nearest bi-directional reference frames in the reference list, are adopted to generate a high quality virtual reference frame $\mathrm{F}_{Virtual}$. Subsequently, for each CTUs in the current frame, a DVRF mode flag is signalled in the bitstreams to indicate whether the DVRF mode is chosen. In particular, when DVRF flag is true, the co-located block in the virtual reference frame is treated as the reconstruction block. Otherwise, traditional HEVC encoding process is conducted to encode the current CTU.

3.2.2. CTU Level Mode Decision

The selection of the DVRF mode is achieved with RDO. Let J_{HEVC} denote the RD cost of the traditional HEVC coding method for the current CTU, such that the rate-distortion cost of the traditional HEVC coding can be formulated as:

$$J_{HEVC} = D_{HEVC} + \lambda * R_{HEVC} \quad (2)$$

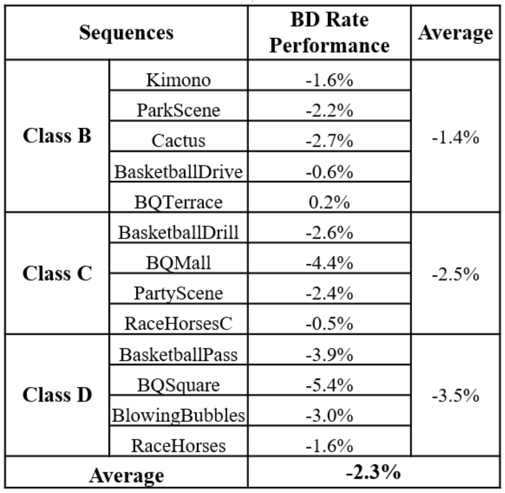
where D_{HEVC} and R_{HEVC} denote the distortion and rate of the HEVC coding, respectively. The parameter λ is the Lagrange multiplier, which controls the relationship between rate and distortion. For the proposed DVRF mode, let J_{DVRF} denote the RD cost of the DVRF mode, which can be expressed as:

$$J_{DVRF} = D_{DVRF} + \lambda * R_{DVRF} \quad (3)$$

where D_{DVRF} and R_{DVRF} are the distortion and rate of the DVRF mode. The flag is set to be true when $J_{DVRF} < J_{HEVC}$, and otherwise it is set to be false.

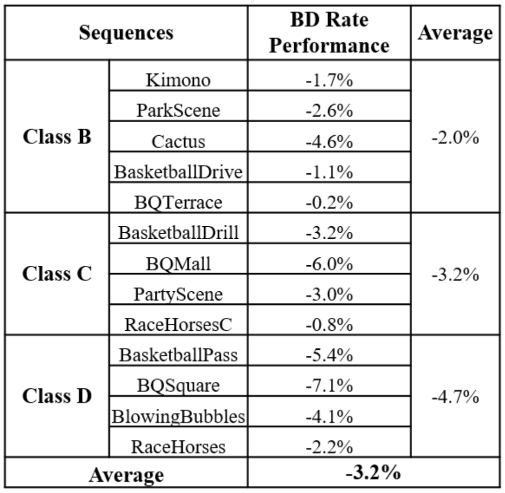
It should be noted that when DVRF mode is chosen, only DVRF flag needs to be singled in the bitstream. Regarding the traditional HEVC modes, motion data including merge index, MVP index, MVD and the quantized transform coefficients need to-be-coded. The effective reduction of coding bits makes DVRF mode a competitive one among all the candidate inter prediction methods.

Table 1. RD performance comparison when applying the proposed method to Level_3 B frames



Sequences		BD Rate Performance	Average
Class B	Kimono	-1.6%	-1.4%
	ParkScene	-2.2%	
	Cactus	-2.7%	
	BasketballDrive	-0.6%	
	BQTerrace	0.2%	
Class C	BasketballDrill	-2.6%	-2.5%
	BQMall	-4.4%	
	PartyScene	-2.4%	
	RaceHorsesC	-0.5%	
Class D	BasketballPass	-3.9%	-3.5%
	BQSquare	-5.4%	
	BlowingBubbles	-3.0%	
	RaceHorses	-1.6%	
Average		**-2.3%**	

Table 2. RD performance comparison when applying the proposed method to Level_2 and Level_3 B frames



Sequences		BD Rate Performance	Average
Class B	Kimono	-1.7%	-2.0%
	ParkScene	-2.6%	
	Cactus	-4.6%	
	BasketballDrive	-1.1%	
	BQTerrace	-0.2%	
Class C	BasketballDrill	-3.2%	-3.2%
	BQMall	-6.0%	
	PartyScene	-3.0%	
	RaceHorsesC	-0.8%	
Class D	BasketballPass	-5.4%	-4.7%
	BQSquare	-7.1%	
	BlowingBubbles	-4.1%	
	RaceHorses	-2.2%	
Average		**-3.2%**	

4. EXPERIMENTAL RESULTS

The coding performance of the proposed method is validated in this section. The proposed scheme is integrated into HM-16.6, and the Adaptive Separable Convolution model [8][4] is used to generate the virtual reference frame. Four quantization parameters (QP), i.e., 27, 32, 37, 42, are employed in the experiment. Since we focus on improving the quality of luma component, YUV400 format is adopted in our experiment. Moreover, the number of encoding frames is set to be twice of the frame rate, and all other experimental conditions follow common test conditions [9].

Generally speaking, the proposed DVRF mode can be applied to all the B frames with bi-directional references. However, the B frames in different levels have distinct characteristics, which may further influence the usage of the DVRF mode. On one hand, the DVRF mode tends to provide finer predictions for B frames with higher levels since the temporal distance of the two frames in FRUC is closer. On the other hand, as DVRF mode seeks optimal performance in terms of RD optimization, the reconstruction quality may get worse when DVRF mode is involved. Accordingly, applying the DVRF mode to lower level B frames is not preferred as the reference quality to the subsequent higher level frames should be taken into consideration. In view of this, we conduct two experiments in RA configuration to explore the performance of DVRF mode for different levels of B frames.

In the first experiment, we apply DVRF mode only to level_3 frames, and the simulation results are depicted in Table 1. As we can see, FRUC provides on average 2.3% BD rate gain on HEVC sequences, and up to 5.4% gain is achieved on *BQSquare*, which demonstrates the effectiveness of the proposed scheme. However, for some particular sequences, such as *RaceHorse* and *RaceHorseC*, DVRF mode only provides slightly better RD performance, as the quality of the virtual reference frames cannot be guaranteed in the extensive motion scenarios. Regarding *BQTerrace*, DVRF mode even has negative effect as Adaptive Separable Convolution may not be able to well handle the waterwave case.

In the second experiment, the DVRF mode is applied to both level_2 and level_3 B frames. As shown in Table 2, the compression efficiency achieves further improvement compared with the level_3 frames only. In particular, 3.2% coding gain is achieved, which demonstrates the robustness of the proposed method when the input frames of the FRUC algorithm have longer temporal distance.

5. CONCLUSION

In this paper, a novel CTU level inter prediction method with the advanced deep learning based FRUC method is presented. The novelty of the proposed method lies in that the virtual reference frame is adaptively generated with advanced deep learning based FRUC method, such that better prediction performance can be achieved without the expense of signalling any motion information. Experimental results show the significant superiority of the proposed method, and more than 3% coding gain has been achieved when compared to the state-of-the-art video coding standard.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (61571017, 61632001), in part by the National Postdoctoral Program for Innovative Talents (BX201600006), in part by the Top-Notch Young Talents Program of China and City University of Hong Kong under Grant 7200539/CS, High-performance Computing Platform of Peking University, which are gratefully acknowledged.

6. REFERENCES

[1] Sullivan G J, Ohm J, Han W J, et al., “Overview of the High Efficiency Video Coding (HEVC) Standard,” *IEEE Trans. Circuits Syst. Video Technol*, vol. 22, no. 12, pp. 1649-1668, 2013.

[2] High Efficiency Video Coding, document ITU-T Rec. H.265 and ISO/IEC 23008-2 (HEVC), ITU-T and ISO/IEC, 2013.

[3] Schwarz H, Marpe D, Wiegand T, “Hierarchical B pictures“, Joint Video Team (JVT), Document JVT-P014, Poznan, 2005

[4] Simon Niklaus, Long Mai, Feng Liu, “Video Frame Interpolation via Adaptive Separable Convolution,” *IEEE ICCV*, pp. 261-270, 2017.

[5] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, “View synthesis by appearance flow,“ *ECCV*, pp. 286-301, 2016.

[6] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala, “Video frame synthesis using deep voxel flow,“ *IEEE ICCV*, pp. 4473-4481, 2017.

[7] Simon Niklaus, Long Mai, Feng Liu, “Video Frame Interpolation via Adaptive Convolution,” *IEEE CVPR*, pp. 2270-2279, 2017.

[8] Github: “sepconv“, http://graphics.cs.pdx.edu/project/sepconv

[9] F. Bossen, “Common HM test conditions and software reference configurations,“ ISO/IEC JTC1/SC29/WG11, JCTVC-G1200, Geneva, CH, 2011.