Convolution Neural Network based Video Coding Technique using Reference Video Synthesis

Jung Kyung Lee†, Nayoung Kim†, Seunghyun Cho‡, and Je-Won Kang†
†Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul, Korea
‡Realistic AV Research Group, Electronics and Telecommunications Research Institute
E-mail: †jungkyong1204@gmail.com, †12skdud21@naver.com, ‡shcho@etri.re.kr, †jewonk@ewha.ac.kr

***Abstract*—In this paper, we propose a novel video coding technique that uses a virtual reference (VR) video frame, synthesized by a convolution neural network (CNN) for an inter-coding. Specifically, an encoder generates a VR frame from a video interpolation CNN (VI-CNN) using two reconstructed pictures, *i.e.*, one from the forward reference frames and the other from the backward reference frames. The VR frame is included into the reference picture lists to exploit further temporal correlation in motion estimation and compensation. It is demonstrated by the experimental results that the proposed technique shows about 1.4% BD-rate reductions over the HEVC reference test model (HM 16.9) as an anchor in a Random Access (RA) coding scenario.**

I. Introduction

Deep neural networks are emerging as important coding techniques for High Definition (HD) and Ultra High Definition (UHD) videos. Both the Video Coding Experts Group (VCEG) and Moving Picture Experts Group (MPEG) have launched a project to work on a new video coding standard, *i.e.*, Versatile Video Coding (VVC) [1], aiming to achieve substantial bit-rates saving around 50% over High Efficiency Video Coding (HEVC) standard. In the Call-for-Proposal (CfP) responses of VVC, there have been several coding methods employing the neural networks to improve coding efficiency [2]. At this moment, the VVC reference software advances the conventional hybrid video coding framework of HEVC with more enhanced coding tools. However, it has been also observed that the neural network based coding tools could replace the existing tools by improved coding performance in the following standardization activities.

Convolution Neural Networks (CNN) have been successfully applied in the field of low-level image processing such as image denoising [3] and image super-resolution [4], and the success is spreading into the video coding techniques. As ringing artifacts and blocking artifacts incurred by quantization process tend to have unique patterns of the distortions, the CNN can obtain useful features in the training to restore the visual quality of the original video frame. Following this idea, there have been several research works to integrate the trained in-loop filters into HEVC. In [5], Park *et al.* develop a CNN-based in-loop filter, replacing the conventional SAO(Sample Adaptive Offset) filter, to remove visual artifacts in video frames. In [6], Kang *et al.* develop a multi-scaled CNN in-loop filter, exploiting coded video parameters to further alleviate blocking artifacts. There are also a few studies to apply adaptive deep learning filters to an intra-coding. Li *et al.* develop a CNN-based block up-sampling filter, so that a down-sampled block is coded, and then restored to the original size in [6]. In [8], Li *et al.* propose a fully-connected neural network to create an intra-prediction blocks from multiple reference lines in the boundaries of the current block.

Most of the previous works have only focused on applying the deep neural networks for enhancing spatial information of video signals to improve coding efficiency. The works have been used for intra-coding or frame interpolation to have rich spatial correlation among pixels.

In this paper, we propose an inter-coding algorithm that exploits enhanced temporal information of a video by the CNN. Our method considers a synthesis of a reference frame, namely a virtual reference (VR) frame, that approximates the current frame as close as possible in the same time. To be specific, given previously coded frames in the decoded picture buffer (DPB), the CNN generates the VR frames of high quality by estimating kernels in an end-to-end manner. Then, the VR frames are included into the reference picture lists for motion estimation and compensation. The VR frame has an obvious advantage as it shows higher temporal correlation with the current video frame than the conventional reference frames of different moments. It is demonstrated by the experimental results that the proposed technique provides about 1.4% BD-rate reductions over the HEVC reference test model (HM 16.9) as an anchor in an Random Access (RA) coding scenario.

The rest of the paper is organized as following. Section II shows the background. Section III explains the proposed technique. The experiment results are shown in Section IV. Conclusion and future works are remarked in Section V.

II. Background

A. Hierarchical B-Picture Coding Structure

The HEVC specifies a layer identifier (layer ID) in the network abstraction layer units to allow temporal prediction in an RA coding scenario [9]. B slices are hierarchically coded with the forward and backward references frames in the higher temporal level. Fig. 1 shows hierarchical B-picture coding structure, in a group of pictures (GOP) 8. The colors of the pictures indicate temporal sub-layers corresponding to temporal layer ID values of 0, 1, 2, and 3 respectively. The arrows show which frame can be used for reference frames in a hierarchical manner. For example, the current frame with

the picture of count (POC) 3 uses a frame with POC 2 and POC 4 as the temporal layer ID of the current frame is lower than those of the reference frames.

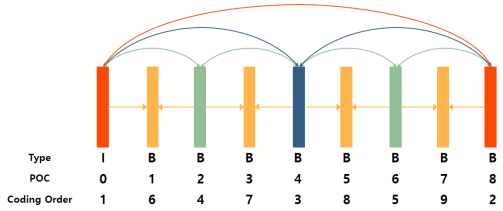

Fig. 1. Hierarchical B-Picture Coding Structure.

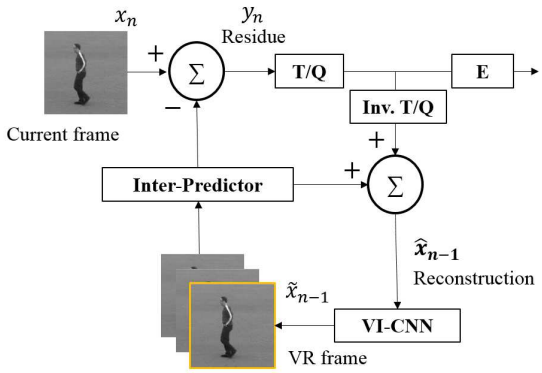

Fig. 2. Encoder blockdiagram of the proposed technique.

B. Video Interpolation by Convolutional Neural Network

Owing to the recent developments of CNNs, video interpolation is actively studied even though it is a classic image processing problem. A conventional video interpolation method uses optical flows. Yet, there are many studies about CNN architectures to generate an intermediate frame in an end-to-end fashion and significantly improve the visual quality. Dosovitskiy *et al.* propose to learn optical flows using CNNs to estimate motion information [10]. Pixel-based prediction by Voxel Flow is developed in [11]. Though the Voxel Flow network is trained without labeling, it can synthesis pixels in the interpolated frame, efficiently. Adaptive Separable Convolution Neural Network (ASCNN) [12] is the state-of-the-art video interpolation algorithm using CNN. The work uses two pairs of two 1-D kernels to synthesize pixels, in which the two kernels are used for the vertical prediction and the other two kernels are used for the horizontal prediction.

III. The Proposed Technique

A. Overview of the Proposed Technique

The overall idea of the proposed algorithm is shown in Fig. 2. As shown, the proposed technique integrates a CNN that can perform video interpolation (VI-CNN) by generation and synthesis into the HEVC framework. The virtual reference (VR) frame $\tilde{x}_{n-1}$ is newly generated from the synthesis process, using previously coded pictures $\hat{\mathbf{x}}_{n-1} = \{\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{n-1}\}$ stored in the decoded picture buffer (DPB). $\tilde{x}_{n-1}$ is added into the reference picture lists, so that it can be used for motion estimation and compensation for coding the current frame x_n. Fig. 2 depicts an encoder blockdiagram. In the decoder side, the same VI-CNN is used for generating the same VR frame to form reconstructed frame. For the generation of the VR frame that can approximate to the current frame as close as possible, we adopt the ASCNN [12] with only the minor modification as the VI-CNN to get high Peak Signal-to-Noise ratio (PSNR) quality at the low complexity. We use the pre-trained model in [12], and, thus none of the trained videos are included in the evaluations of coding performance.

B. VI-CNN for VR Frame Generation

The VI-CNN synthesises an intermediate video frame by using two frames, *i.e.,* one is from forward reference frames, and the other is from backward reference frames on the hierarchical B-picture coding structure. Fig. 3 shows the VI-CNN using the Adaptive Separable Convolution Neural Network. The two reconstructed frames from the DPB go through the network to create the VR frame as an intermediate frame. It is noted that the VR frame can be efficiently used for B-slices in the RA coding scenario as the Low-Delay coding scenario or P slices are restricted to perform the video interpolation.

It is also highlighted that the two input frames can change the visual quality of the VR frame, which would affect video coding efficiency. Therefore, in the proposed technique, we choose each of the forward and the backward reference frames, which has the lowest picture of counts (POC) difference between the current frame, from the reference picture lists of the conventional HEVC codecs. As the POC differences are the same in the forward and the backward reference frames, the VR frame displays the same moment of the current frame in principle.

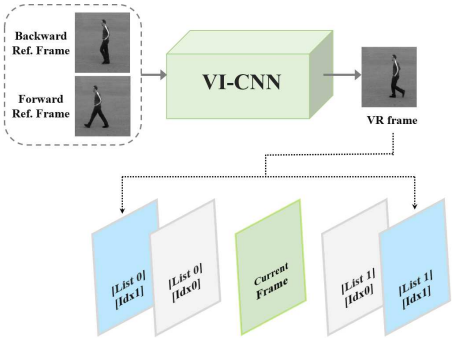

Fig. 3. VI-CNN using Adaptive Separable Convolution Neural Network in the proposed technique.

C. Reconfiguration of the Reference Picture Lists in HEVC

The VR frames are included into the reference picture lists to be used for motion estimation and compensation. For this, the reference pictures in the two reference picture lists denoted by List0 and List1 [14] in HEVC need to be reconfigured in the RA coding scenario. As the RA configuration in HEVC uses the hierarchical B-picture coding structure, we configure the both reference picture lists for the bi-directional motion predictions as follows.

First, the proposed technique reconfigures the lists only when coding a picture of non-reference picture type [14] residing in the lowest temporal layer, so that the other frames in the higher layers can be decoded properly. For example, in Table I, the current frames whose POC numbers are 1, 3, 5, and 7 can use the VR frames in the reference picture lists. Second, the reference pictures corresponding to the first indices in the two lists remain the same as the original HEVC. However, the reference pictures corresponding to the second indices of the lists are replaced with the VR frames.

The specific changes of the reference picture lists of the HEVC reference software model (HM16.9) are shown in Table I. The reference picture index (Idx) points a reference frame specified by the corresponding POC number. For example, the current frame whose POC number is 2 is coded with the backward and the forward reference frames whose POC numbers are 0 and 4, respectively, as in HEVC. However, the current frame whose POC number is 1 replaces the conventional reference frames in the second indices with the VR(0,2). VR(0,2) denotes the VR frame interpolated by using the reference frames 0 and 2 for an example. The VR frames are generated using temporally adjacent reference frames, so the lists manages only the short-term references. As the maximum number of the reference frames being active in the RA mode is two [14], the coding gain comes from the replacements of the conventional reference frames with the VR frames, if any.

TABLE I
RECONFIGURATION OF REFERENCE PICTURE LISTS IN HM16.9

POC	Reference Picture List 0		Reference Picture List 1	
	Idx 0 (POC)	Idx 1 (POC)	Idx 0 (POC)	Idx 1 (POC)
8	0	0	0	0
4	0	8	8	0
2	0	4	4	8
1	0	VR(0,2)	2	VR(2,0)
3	2	VR(2,4)	4	VR(4,2)
6	4	0	8	4
5	4	VR(4,6)	6	VR(6,4)
7	6	VR(6,8)	8	VR(8,6)

IV. EXPERIMENTAL RESULTS

A. Experimental Configurations

In this section, we first evaluate the coding performance of the proposed technique as in the common test conditions (CTC) [15]. We implement the proposed technique in the HEVC reference software HM16.9, and also compare the performance to the anchor. The quantization parameters are 22, 27, 32, and 37. The RA configuration is the same as in the CTC, where the size of the group of pictures is set to 8. We use the pre-trained ASCNN model without any fine-tuning, and, thus the training/testing video sets are exclusively chosen.

TABLE II
THE RATE-DISTORTION PERFORMANCE OF THE PROPOSED TECHNIQUE VERSUS THE ANCHOR HM16.9

Class	Sequence	BD-rate		
		Y	U	V
Class B (1920x1080)	BasketballDrive	-0.1%	0.3%	0.6%
	BQTerrace	-0.6%	-0.4%	-0.4%
	Cactus	-1.6%	-0.3%	-1.2%
	Kimono1	-2.0%	1.0%	1.0%
	ParkScene	-2.0%	0.0%	0.0%
Class C (832x480)	RaceHorses	-0.5%	-0.2%	-0.6%
	BQMall	-2.5%	-1.4%	-1.7%
	PartyScene	-1.4%	-0.8%	-1.0%
	BasketballDrill	0.0%	0.0%	-1.0%
Class D (416x240)	RaceHorses	-1.7%	-0.8%	-1.2%
	BlowingBubbles	-2.8%	-1.7%	-2.0%
	BasketballPass	-3.4%	-2.3%	-3.9%
	BQSquare	-3.3%	-0.2%	-0.8%
Class E (1280x720)	Johnny	-0.1%	0.1%	0.2%
	FourPeople	-0.9%	-0.3%	-0.2%
	KristenAndSara	-0.3%	0.1%	0.1%
Total Average		-1.4%	-0.4%	-0.8%

B. Rate-Distortion Performance Comparisons

Experimental results show that the proposed technique provides around 1.4% of the BD-rate reductions on average as compared to the anchor. Table II shows the performance comparisons between the anchor and the proposed technique. Fig. 4 shows the Rate-Distortion curve of "BasketBallPass". As shown, the proposed technique outperforms the anchor over the range of the bit-rates.

It is observed that the test video sequences with fast motions show higher coding gains as compared to the other test sequences. For example, "BQSquare" and "BasketBallPass" yield more than 3 $\%$ coding gains. As compared, the test videos with small motions such as "Johnny" and "KristenAndSara" show only minor coding gains. To be specific, we obtain the residual signals between the original frame and the prediction frames to figure out how the VR frame can be efficiently used for the motion prediction. Fig. 5 shows the visual comparisons using "Cactus" test video. The video frame in Fig.5.(d) shows the difference between the original frame and the VR frame, synthesized by the forward reference frame and the backward reference frame. Meanwhile, the differences between the original frame and the forward/backward reference frames are respectively shown in Fig.5.(e) and Fig.5.(f).

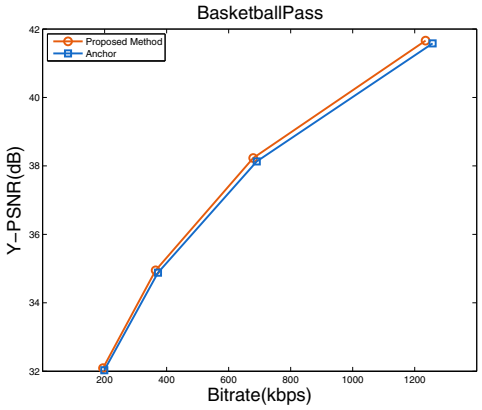

Fig. 4. The Rate-Distortion curve of BasketballPass.

As shown, the residues in Fig.5.(d) are much smaller than in Fig.5.(e) and (f). Actually, the residual signals are quite small, owing to the deep learning technique, when being compared to the difference between the original video and the its reconstruction after the coding. The results demonstrate the VR frames can be efficiently used for exploiting temporal correlation in the motion prediction.

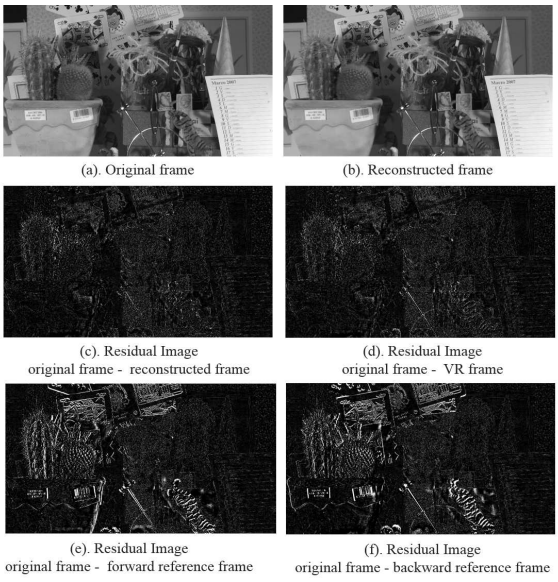

Fig. 5. Visual quality comparisons of the residues between the original frame and the various reference frames.

V. Conclusion

In this paper, we propose a new video coding algorithm in inter prediction using a deep learning technique. The proposed technique uses the virtual reference frame to reduce temporal redundancies, so that achieves 1.4% BD-rate saving on average. Our future work includes further developments of the VI-CNN to enhance the quality of the VR frames as well as improvements of the low-level inter-coding tools applied to the VR frames to enhance coding gains.

Acknowledgment

This work was supported by Institute for Information and communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (2017-0- 00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media)

References

[1] B. Bross, “Versatile Video Coding (Draft 1).” JVET-J1001, San Diego, USA, Apr. 2018.
[2] S. Liu, L. Wang, P. Wu, and H. Yang, “JVET AHG report: Neural Networks in Video Coding (AHG9).” JVET-J0009, San Diego, USA, Apr. 2018.
[3] C. Dong, Y. Deng, C. C. Loy, and X. Tang “Compression Artifacts Reduction by a Deep Convolutional Network” IEEE International Conference on Computer Vision (ICCV) 2015.
[4] J. Kim, J. Lee, and K. M. Lee “Deeply-recursive convolutional network for image super-resolution.” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016
[5] W. S. Park and M. Kim, “CNN-based in-loop filtering for coding efficiency improvement.” IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016
[6] Kang, Jihong, Sungjei Kim, and Kyoung Mu Lee. “Multi-modal/Multi-scale convolution neural network based in-loop filter design for next generation video codec.” in Proceedings of the International Conference on Image Processing, 2017.
[7] Li, Y., Liu, D., Li, H., Li, L., Wu, F., Zhang, H. and Yang, H., “Convolutional Neural Network-Based Block Up-sampling for Intra Frame Coding”IEEE Transactions on Circuits and Systems for Video Technology (2017).
[8] J. Li, B. Li, J. Xu and R. Xiong, “Intra prediction using fully connected network for video coding,” IEEE International Conference on Image Processing (ICIP), 2017.
[9] J.-W. Kang, Y.-Y Lee, C.-S. Kim, and S.-U. Lee “Coding Order Decision of B frames for Rate-Distortion Performance Improvement in Single-View Video and Multi-View Video Coding”, IEEE Trans. on Image Processing, vol.19, no.8, pp.2029-2041, 2010.
[10] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. “Flownet:Learning optical flow with convolutional networks”. In ICCV,2015
[11] Liu, Z., Yeh, R., Tang, X., Liu, Y. and Agarwala, A. “Video frame synthesis using deep voxel flow.” International Conference on Computer Vision (ICCV). Vol. 2. 2017.
[12] Niklaus, Simon, Long Mai, and Feng Liu. “Video frame interpolation via adaptive separable convolution.” arXiv preprint arXiv:1708.01692 (2017).
[13] Lin, J.L., Chen, Y.W., Huang, Y.W. and Lei, S.M. ”Motion vector coding in the HEVC standard.” IEEE Journal of selected topics in Signal Processing 7.6 (2013): 957-968.
[14] Sjoberg, R., Chen, Y., Fujibayashi, A., Hannuksela, M.M., Samuelsson, J., Tan, T.K., Wang, Y.K. and Wenger, S., “Overview of HEVC High-Level Syntax and Reference Picture Management.” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1858-1870, Dec. 2012
[15] F. Bossen, “Common test conditions and software reference configurations” JCTVC-L1100, 12th JCT-VC meeting, Geneva, CH, Jan.2013.
[16] Y. Dai, D. Liu, and F. Wu, “A convolutional neural network approach for post-processing in HEVC intra coding” in Proceedings of the International Conference on Multimedia Modeling, 2017.