DCC 2021 Bi-prediction Enhancement with Deep Frame Prediction Network for Versatile Video Coding

Hao Tao, Jian Qian, Li Yu*, Hongkui Wang

School of Electronic Information and Communications Huazhong University of Science and Technology, Wuhan, China

Abstract

Bi-prediction is a fundamental module of inter prediction in the blocked-based hybrid video coding framework. Block-based motion estimation (ME) and motion compensation (MC) with simple models are adopted in bi-prediction process. Unfortunately, this MEMC-based algorithm can't guarantee the prediction performance when it comes to videos with irregular motions. In this paper, a novel inter prediction scheme based on deep frame prediction network (DFP-net) is proposed to enhance bi-prediction accuracy, especially in complicated scenes. The DFP-net can precisely extract and fuse motion features in various scales and completely exploit temporal and spatial correlation to generate the prediction frame in a data-driven manner. Moreover, the DFP-net is integrated into VTM-6.2 to provide an additional prediction frame for bi-prediction. Since the prediction generated by DFP-net is more similar with the to-be-coded frame in the sense of temporal distance and texture, it can be appended to reference list to improve the content diversity of references.

Introduction

The block-based hybrid coding framework has been adopted in the exiting video coding standards, such as High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC). The general scheme of inter prediction is to perform motion estimation and motion compensation on current frame block by block to obtain the prediction frame. Nevertheless, the aforementioned inter prediction algorithms adopt coarse-grained motion estimation and motion compensation with low-order motion model to obtain the prediction block, which makes it unreliable when processing videos with irregular motions.

In this paper, a novel bi-prediction scheme based on deep convolution neural network (CNN) is proposed to generate a high quality prediction frame for compression performance enhancement. Specifically, a deep frame prediction network, namely DFP-net, is devised to generate an accurate prediction with two reference pictures of current frame. The proposed DFP-net consists of multi-scale motion alignment, fusion with temporal and spatial correlation and frame synthesis module, which can take full advantage of motion features and temporal correlation to locate and reconstruct the target frame. Since the CNN is of superb nonlinear fitting capabilities, complicated motion can be well modeled. The precise motion correspondence will further lead to better motion alignment and frame reconstruction. To improve the compression performance of VVC, the well-trained deep frame prediction model is integrated into VTM-6.2 to generate a better prediction, which is then utilized as an additional reference frame (ARF) to extend current reference lists. Since the ARF is a fusion of the original reference pictures, motion features will be implicitly encoded into it during the forward derivation process of DFP-net. Consequently, the ARF can be employed as a better hypothesis to obtain a more accurate prediction.

Design of the DFP-net

For improving bi-prediction performance, a deep frame prediction network, namely DFP-net, is devised to generate high quality prediction in challenging scenes. As shown in Fig.1, the proposed DFP-net consists of multi-scale motion alignment, fusion with temporal and spatial correlation and frame synthesis module. The proposed alignment algorithm extract features from up to down and fuse features in a bottom-up manner, motions of various grain can be effectively tackled and precise motion registration can be achieved in this way. In order to take full advantage of temporal and spatial correlation between consecutive frames, temporal attention (TA) and spatial attention (SA) layers are designed to estimate the correlation in a data-driven manner. Having obtained the well aligned and weighted features, a frame synthesis network is devised to reconstruct the target frame. Since residual learning can deepen the model and exploit hierarchical information, numbers of residual blocks are cascaded to synthesize the intermediate frame. Owing to the delicate design of DFP-net, motion features and spatial-temporal correlation can be fully used to synthesize the target frame.

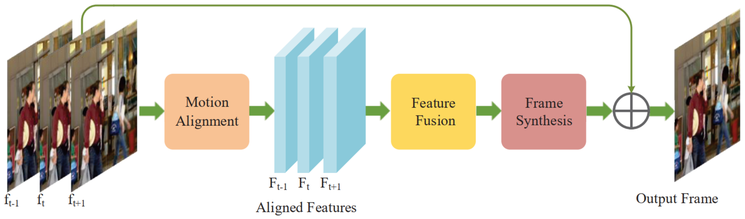

Figure 1 The framework of the proposed DFP-net.

Bi-prediction scheme based on DFP-net

Since the well designed DFP-net can represent motion in an efficient way and generate intermediate frame of high quality, it can be introduced to bi-prediction for compression performance enhancement. As depicted in Fig.2, the proposed bi-prediction method includes two major procedures: ARF generation and reference list extension. The proposed scheme is applied in the hierarchical B coding structure and the frames in low temporal layers will be reconstructed first and then serve as reference for frames in high layers. Consequently, the ARF generation process can be performed on frames with higher temporal layers. In order to obtain the ARF of high quality, two frames that are of identical temporal distance to current will be chosen from the current reference lists as the input of DFP-net. In order to take full advantage of the ARF, it will be appended to current reference list to improve the content variety of reference. Since the ARF is more similar with current frame in both temporal distance and texture, better prediction can be obtained from it by the MEMC-based inter prediction algorithms.

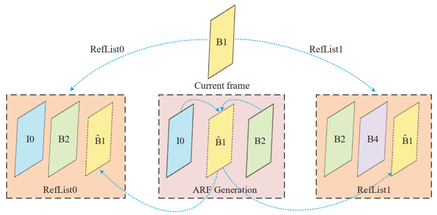

Figure 2 Illustration of the proposed bi-prediction scheme.

Result

The evaluation results are shown in Table.1. The proposed bi-prediction scheme has achieved on average 1.8% BD-rate saving with 29.7% complexity increase. Since the ARP generation process is executed just once a frame, the complexity overhead is reasonable considering the coding gain. Moreover, in order to further explore the reason for performance difference, the pick ratio of the learned additional reference picture. As shown in Table.2, the ARFs are selected by a considerable number of PUs as reference, which demonstrates that the DFP-net can provide a competitive reference quality in most scenes. It can be inferred that the compression performance is tightly associated with the pick ratio of ARF. The aforementioned results show the effectiveness of our method.

Table 1 Coding performance of the proposed method.

Sequence		Complexity	BD-rate	
Class A	PeopleOnStreet	37.9%	-5.4%	-3.8%
	Traffic	40.9%	-2.1%	
Class B	Kimono	38.2%	-1.0%	-0.6%
	ParkScene	30.6%	-1.5%	
	Cactus	20.5%	-0.7%	
	BasketballDrive	33.0%	+0.1%	
	BQTerrace	24.1%	+0.2%	
Class C	BasketballDrill	36.0%	-1.0%	-2.8%
	BQMall	25.6%	-5.5%	
	PartyScene	24.4%	-1.7%	
	RaceHorseC	45.2%	-3.1%	
Class D	BasketballPass	11.8%	-2.9%	-2.4%
	BQSquare	7.6%	-2.0%	
	BlowingBubbles	9.7%	-1.7%	
	RaceHorse	21.3%	-3.1%	
Class E	FourPeople	32.9%	-0.9%	-0.3%
	Johnny	24.2%	+0.1%	
	KristenAndSara	20.2%	-0.1%	
Average		29.7%	-1.8%	

Table 2 The pick ratio of ARF.

Sequence	Pick Ratio				
	27	32	37	42	Overall
RaceHorse	49.1%	63.3%	70.2%	80.3%	58.1%
BQMall	40.9%	49.1%	54.2%	64.8%	46.9%
BlowingBubbles	46.0%	53.4%	57.5%	69.3%	52.1%
PeopleOnStreet	54.5%	60.4%	63.8%	49.1%	57.1%
BQTerrace	32.9%	28.3%	31.3%	30.6%	31.5%
BasketballDrive	17.0%	17.3%	27.3%	40.9%	21.5%

Conclusion

In this paper, a novel deep p network-based bi-prediction algorithm is proposed to improve inter prediction performance in VVC. The novelty of the proposed scheme lies in that the DFP-net is introduced to bi-prediction to efficiently represent motions and generate a better prediction in a data-driven style. Owing to the delicate design of the network and the end-to-end training strategy, the DFP-net can effectively capture motions of various scales and fully exploit spatial and temporal correlation to synthesize a high quality target frame. Considering that the prediction generated by DFP-net is more similar with current frame in both temporal distance and texture, it can be employed as an addition reference picture to diverse the contents of references. In this manner, an average of 1.8% compression gain can be achieved over VTM-6.2, which demonstrates the superiority of the proposed method.