Hierarchical B-frame with Conditional Video Coding

David Alexandre, Hsueh-Ming Hang, Wen-Hsiao Peng
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
{davidalexandre.eed05g, hmhang}@nctu.edu.tw, wpeng@cs.nctu.edu.tw

Abstract

We proposed a learning-based hierarchical bi-directional (B-frame) video coding, which uses the information from both past and future frames. We incorporate a conditional ANF coding scheme for encoding the motion information and the image residual. Our design includes a mask-merge-net that reconstructs the motion-compensated frame. Also, our model is scalable for variable bit-rates and targeted at MS-SSIM metric. The evaluation results show that our method produces good visual quality images and achieves competitive PSNR compared to other learning-based methods.

1. Introduction

Our proposed system performs hierarchical bi-directional (B-frame) video coding. We use the previously decoded past and future frames to do bi-directional prediction. In brief, our contributions can be summarized as follows. (1) We design an end-to-end hierarchical bi-directional coding with scaling variable rates that control the bit rate in the compressor to match the CLIC Challenge bit rate requirement. (2) We integrate the conditional coding scheme developed by [4] together with the B-frame coding framework. (3) We propose a mask-merge-net for the feature-pixel domain merge and image synthesizing mechanism to reconstruct the motion compensated frame from $\hat{x}_{t-1}$ and $\hat{x}_{t+1}$.

2. Proposed Method

Our proposed system is shown in Fig 1. We adopt the hierarchical bi-directional video coding concept and frame structure in the traditional video codec (such as H.265), but we replace the motion and residual components by the deep neural networks. The system input is RGB format. Inside the GOP (Group Of Pictures) structure, each GOP contains one I-frame (intra coding) in the first frame. Assuming the GOP size is 16, the first I-frame of the next GOP is used together with the frame 1 (I-frame) of the current GOP to perform the first-level bi-directional prediction for frame 5. Then, frame 1 and frame 5 are used to predict frame 3.

For the intra coded frame, we use the H.266/VVC coder [1]. The main components inside our system are motion estimator powered by SpyNet [2], motion compensation (mask-merge-net) inspired by [3], and conditional ANF (Augmented Normalizing Flows) motion/residual compressor from [4]. Our motion-compensation process includes the forward warping mechanism from [5].

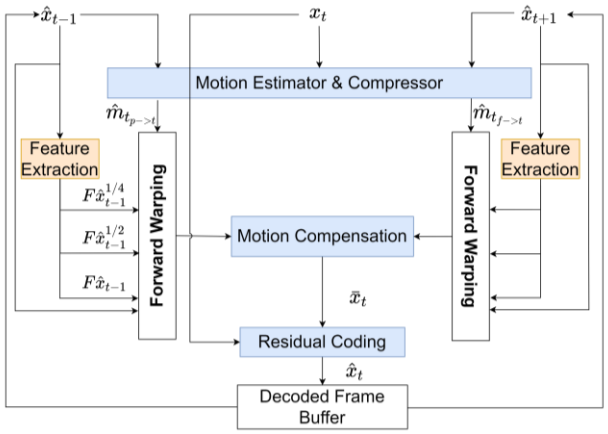

Fig. 1. Our proposed system. The decoded frames from past $\hat{x}_{t-1}$, and future $\hat{x}_{t+1}$ are used in the motion estimator process to produce the motion vectors $\hat{m}_{t_{p\to t}}$ and $\hat{m}_{t_{f\to t}}$.

The motion compensation module consists of pairs of mask-net and merge-net. As shown in Fig. 2, the inputs of this module are decoded motion vectors of both the past and future frames. The past/future frames are first converted to features in three scales using a feature extractor as shown in Fig 2. Then, the past/future frames in feature and pixel domains are fed into the mask-merge-net. The design of mask-net and merge-net is adopted from [6].

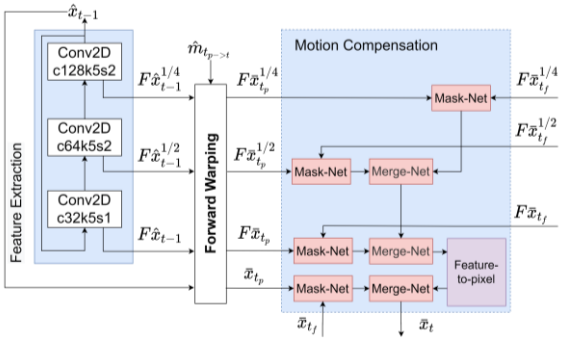

Fig. 2. We use forward warping to estimate the motion compensated feature/frame. Each of estimated frames is masked and merged using neural networks.

We adopt the conditional ANFIC [4] as our residual coder. Lin, et al. [8] introduced the Scaling-Net for video compression. It uses a conditional autoencoder modulated by the lambda parameter to achieve multi-rate coding. We include this Scaling-Net tool into our residual coder to adjust the bit rate to match the specified rates.

3. Training Setup

The batch size in training is 8 and each epoch includes three scaling parameters. The training process has three phases: (a) the motion compressor, (b) the motion compressor and compensator together, and (c) the entire system targeting at the MS-SSIM image quality metric. We trained the model using Vimeo-90k dataset [7] and using Adam optimizer. The training is performed using NVIDIA V100 for 3 days with various scaling parameters.

4. Evaluation

The evaluation was performed on the CLIC Validation dataset. The coding parameters are determined according to the sequences and set specifically for each video sequence empirically. Our system achieves 24.57dB PSNR at 0.1Mbps and 28.31dB at 1Mbps on the validation dataset. For test dataset, it achieves 25.577dB PSNR at 0.1 Mbps and 28.441dB at 1Mbps.

Acknowledgement

This work is partially supported by the Ministry of Science and Technology, Taiwan under Grant MOST 109-2634-F-009-020 through Pervasive AI Research (PAIR) Labs, National Yang Ming Chiao Tung University, Taiwan. Also, we would like to thank National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

[1] A. Wieckowski, et al. Towards A Live Software Decoder Implementation for The Upcoming Versatile Video Coding (VVC) Codec. International Conference on Image Processing (ICIP), 2020.

[2] A. Ranjan and M.J. Black. Optical Flow Estimation using a Spatial Pyramid Network. arXiv preprint arXiv:1611.00850, 2016.

[3] M.A. Yılmaz and A.M. Tekalp. End-to-End Rate-Distortion Optimized Learned Hierarchical Bi-Directional Video Compression. IEEE Transactions on Image Processing, 2021.

[4] Y.-H. Ho, et al. ANFIC: Image Compression Using Augmented Normalizing Flows. arXiv preprint arXiv: 2107.08470, 2021.

[5] S. Niklaus and F. Liu. Softmax Splatting for Video Frame Interpolation. IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[6] M. Akın Yılmaz, A.Murat Tekalp. End-to-End Rate-Distortion Optimized Learned Hierarchical Bi-Directional Video Compression. IEEE Transactions on Image Processing, 2021.

[7] T. Xue, et al. Video Enhancement with Task-Oriented Flow. International Journal of Computer Vision (IJCV), 2019.

[8] J. Lin, et al. A Deeply Modulated Scheme for Variable-Rate Video Compression. In 2021 IEEE International Conference on Image Processing (ICIP), 2021.