Joint Learned and Traditional Video Compression for P Frame

Zhao Wang, Ru-Ling Liao, Yan Ye
XG Lab of Damo Academy, Alibaba Group
Team:DAMO_XG

Abstract

In this paper, we propose a joint learned and traditional video compression framework for the P frame track on learned image compression hosted at CVPR2020. The main difference between video compression and image compression is that the former has high degree of similarity between the successive frames which can be utilized to reduce the temporal redundancy. Therefore, we first introduce a decoder-side template-based inter prediction method as an efficient way to obtain reference blocks without the need to signal the motion vectors. Secondly, a CNN post filter is proposed to suppress visual artifacts and improve the decoded image quality. Specifically, the spatial and temporal information is jointly exploited by taking both the current block and similar block in reference frame into consideration. Furthermore, an advanced SSIM based rate-distortion optimization model is proposed to achieve best balance between the coding bits and the decoded image quality. Experimental results show that the proposed P frame compression scheme achieves higher reconstruction quality in terms of both PSNR and MS-SSIM.

1. Introduction

Efficient video compression (also named video coding) has been a critical factor for enabling many popular consumer applications, e.g., TV broadcasting, video conference, social networking, e-commerce, remote education, and so on. For example, without video compression, the High Definition (HD) video at 1080p resolution and 60 frames per second demands about 1.44Gbps bandwidth to transmit, which can't be adopted by any video consumer. Video compression systems exploit the internal redundancy of the video signal to significantly reduce the storage size and transmission bandwidth.

Over the past decades, a large number of companies and research institutes around the world have been working on video compression and released several video coding standards, such as the H.264/MPEG4 part 10 AVC standard [1] and the H.265/HEVC standard [2]. In recent years, a new Versatile Video Coding (VVC) standard [3] is under development to further improve video coding efficiency. In all these standards, a block-based hybrid video coding framework is used to exploit the spatial redundancy, temporal redundancy and information entropy redundancy in video.

In VVC, the to-be-coded frame is first divided into non-overlapping equal-sized image regions, such as 128×128, and then further divided into smaller blocks called Coding Units (CU), following a hierarchical quad-ternary-binary partitioning tree to adapt to the local content properties. A CU can be coded by intra- or inter- prediction. If intra prediction is used, spatial neighboring samples are used to predict the current block. If inter prediction is used, one or more similar blocks will be searched from the already coded pictures and used to predict the current block. The relative position shift between the current block and its similar blocks (also called reference blocks) is called motion vector (MV) and also need to be signalled to decoder. The residual, namely the difference between the current block and the prediction block, is sent to the transform and quantization modules to generate the quantized residual coefficients, which are then sent to entropy coding module to be coded. At the decoder side, the quantized residual coefficients will be inverse quantized and inverse transformed to obtain the reconstructed residual. The intra or inter prediction block and the reconstructed residual are added together to form the reconstructed block.

In recent years, convolutional neutral networks (CNN) based image/video compression has become an active research area. Many works have revealed great potentials in learned image compression [4, 5, 6], such as high-efficiency transforms, soft-to-hard quantization, and learned entropy model (e.g., hyperpriors for probability estimation and joint priors from autoregressive neighbors and hyperpriors). An end-to-end learned video compression framework was proposed in CVPR2019 [7]. Specifically, learning based optical flow estimation is utilized to obtain the motion information and reconstruct the current frames. Then two auto-encoder style neural networks are deployed to compress the corresponding motion and residual information.

By investigating the traditional and learned methods on

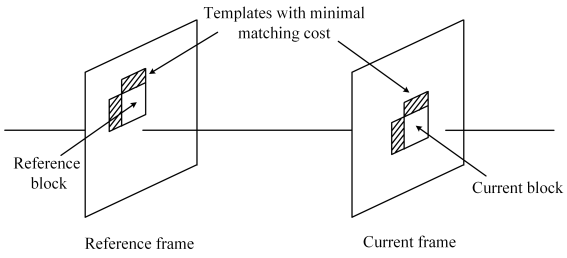


Figure 1. Illustration of DTIP.

video compression, it is found that the learned method cannot outperform the traditional inter prediction in view of reducing temporal redundancy, especially in the case when few reference frames are available. However, the traditional video compression framework consumes much rate cost to signal the motion information which limits the overall compression efficiency. This observation motivates us to propose a decoder-side template-based inter prediction method to efficiently obtain similar reference blocks without explicitly signalling the motion information. Considering visually annoying artifacts are often observed in the reconstructed frames, a CNN filter is adopted to address this issue and improve the reconstructed video quality. Specifically, the spatial and temporal information is jointly exploited by taking both the current block and the reference block into consideration during the processing of CNN filter. To achieve the best performance in balancing the coding bits and the reconstructed distortion, a rate-distortion optimization function is trained offline and deployed to dynamically adjust the weight between rate and distortion.

2. Proposed method

The proposed method is implemented on VVC test model [3] and includes three additional components: 1) decoder-side template-based inter prediction, 2) joint spatial-temporal CNN filter, 3) rate-distortion optimization model, as detailed in the following.

2.1. Decoder-side template-based inter prediction

Based on the JEM codec [8], a decoder-side template-based inter prediction (DTIP) is proposed to fetch the reference block from the reference frame (namely the input frame in the P frame challenge) without cost of signalling motion information. The proposed DTIP is based on the truth of high correlations in spatial neighboring samples. Because the block-based hybrid coding scheme compresses frame block by block following the top-to-bottom and left-to-right order, the left and top image regions of the current to-be-coded block have been reconstructed, which contains some information that can be used to generate the current block. As illustrated in Figure 1, for the current block, template is specified as the left and top neighboring samples, such as the left two columns and the top two rows in our implementation. In the reference frame, motion estimation is applied to search the reference template which has the minimal matching cost with the current template. Then, the co-located block corresponding to the template in the reference frame is fetched to be the prediction of the current block.

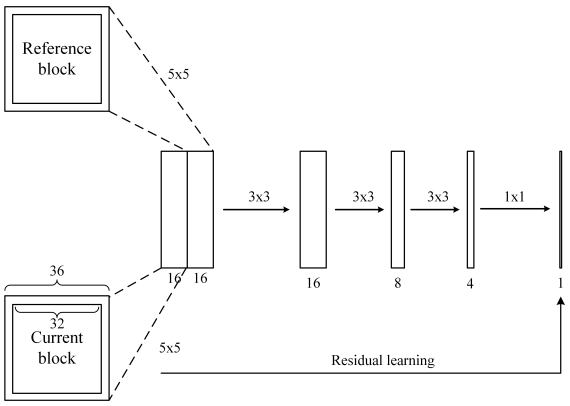


Figure 2. Architecture of the spatial-temporal CNN filter.

Though the motion estimation introduces computation cost at the decoder, it is tolerable because DTIP serves as an optional prediction mode. Whether to be used is decided for each block at the encoder and one flag is signalled to decoder. To further reduce the computational complexity, a fast motion estimation algorithm based on local greedy strategy is applied.

2.2. Spatial-temporal block based CNN filter

To reduce the visually annoying artifacts and further enhance the decoded video quality, a post-processing CNN filter is added. There are two novel designs in the proposed CNN filter. Firstly, besides the spatial samples, the temporal samples in the reference frame will also be taken into consideration and jointly trained. Secondly, the proposed network is deployed block by block and the input samples include the current/reference blocks and corresponding their margins. Specifically, 32×32 processing size is adopted in our implementation while 36×36 regions (adding the neighbouring four rows/columns) are used as input. The reference block is obtained by motion estimation in the reference frame.

The structure of the proposed network is shown in Figure 2 where the feature map numbers of each layer are also provided. The current image region and the reference image region are input to the first layer and convolutional operations with 5×5 kernel are conducted to extract the spatial features. The output features of the first layer are stacked for subsequent layers through fusion of spatial feature maps.

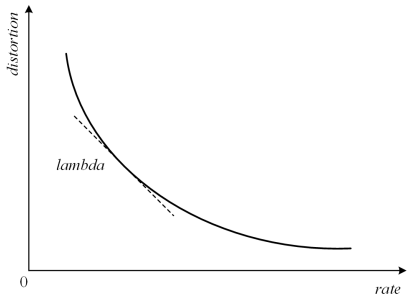


Figure 3. Illustration of rate-distortion optimization.

Three more convolution layers are followed and all of them utilize 3×3 filters. It is worth noting that rectified linear units (ReLU) are adopted as nonlinear mapping for all convolution layers [9]. To accelerate the speed of training, we design the convolution layers as residue learning [10] and the final output of the fifth layer is the element-wise sum of the current block.

To handle various quality levels of the decoded frames, the training data are generated in a wide bit-per-pixel (bpp) range from 0.002 to 0.03. Before the training and deployed of network, all frames will be extended to four more rows and four more columns by padding and motion estimation is conducted for each block. With respect to the loss function, only the distortion in terms of Structure Similarity Index (SSIM) is taken into consideration since no additional bits are introduced by this procedure.

2.3. Rate-distortion optimization

In video coding scheme, the compression efficiency is jointly evaluated by the bitrate and the coding distortion between the original and the reconstructed video. In general, for the same codec, compressed bitrate and coding distortion are two balancing factors. When more bits are consumed, more details can be reserved and hence lower distortion is achieved, as shown in Figure 3. Therefore, rate-distortion optimization (RDO) dedicated to achieving the optimal balance between the rate and the distortion plays a crucial role in video coding scheme [11]. The following RDO cost function is used,

$$J = R + \lambda \cdot D \tag{1}$$

where D, R and J denote the rate, the distortion and the joint cost, respectively. The factor λ is the Lagrangian multiplier, which is quite important in the RDO cost function.

In our scheme, the distortion is evaluated by SSIM and the RDO cost function is converted into,

$$J = R + \lambda \cdot n \cdot (1 - SSIM) \tag{2}$$

where n is the number of image samples. Assuming the rate R and the distortion D are differentiable everywhere, the

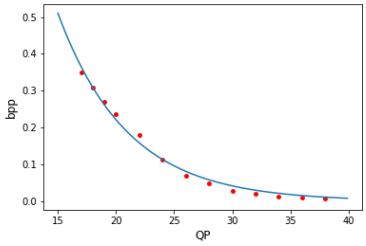


Figure 4. Illustration of the relationship between rate and QP.

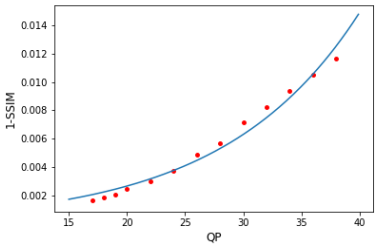


Figure 5. Illustration of the relationship between distortion and QP.

minimum of the RDO cost J is given by setting its derivative to zero,

$$\lambda = -\frac{\partial R}{\partial D} \tag{3}$$

Though the $R - D$ model is complex in real video coding scenarios, the most important factor lies in the quantization level which decides the relationship between rate and distortion directly. Therefore, the lambda derivation can be modelled as,

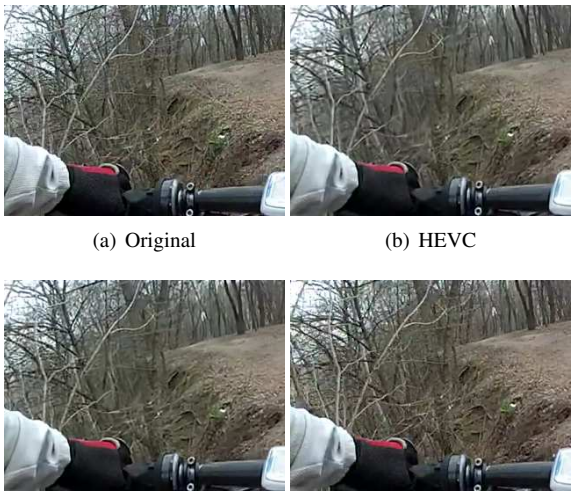
$$\lambda = -\frac{\partial R(QP)}{\partial D(QP)} = -\frac{\partial R/\partial QP}{\partial D/\partial QP} \tag{4}$$

where QP represents the quantization parameter used in the codec.

To exactly explore the expressions of $\partial R(QP)$ and $\partial D(QP)$, extensive experiments are conducted on videos with different contents and motion activities. The tested videos are compressed with the QP in the range between 15 and 40 and the corresponding bitrate and distortion per pixel are recorded. By averaging the recorded data at each QP, the relationship of $R - QP$ and $D - QP$ are illustrated in Figure 4& 5, respectively. From figure 4, it is observed that there exists a clear exponential relationship between bpp and QP. This relationship can be modelled as:

$$R/n = bpp = p \cdot e^{-q \cdot QP} \tag{5}$$

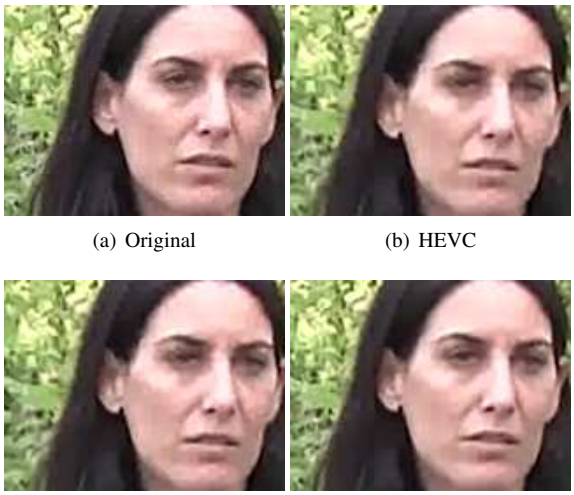
where p and q are the model parameters, and are set to be 6.28 and 0.167 respectively. With respect to the relationship between distortion and QP, it can be approximately



(a) Original (b) HEVC

(c) VVC (d) Proposed

Figure 6. Visual quality comparisons among HEVC, VVC and the proposed method.



(a) Original (b) HEVC

(c) VVC (d) Proposed

Figure 7. Visual quality comparisons among HEVC, VVC and the proposed method.

expressed as

$$D/n = 1 - SSIM = a \cdot e^{(x+b)/c} \tag{6}$$

where the parameters a, b and c are set to be 0.00052, 0.238 and 12.05, respectively. By incorporating 5 and 6 into 4, the multiplier λ can be determined as

$$\lambda = 24302 \times e^{-(0.25 \cdot QP + 0.02)} \tag{7}$$

3. Experimental results

In our proposed scheme, a rate control algorithm is designed to allocate bits for each frame dynamically. Each video is compressed with multiple quantization levels. The slopes between delta MS-SSIM and delta rate under different quantization levels are computed, which indicates the benefits to image quality when the rate is increased, and the data points with the largest slopes will be selected. According to the CLIC challenge requirement, the target bitrate is set to be about 0.075 bpp. To verify the performance of the proposed method, we have submitted one result named DAMO_XG for compressing the frames in the validation dataset. Table 1 demonstrates the coding performance of different methods in the validation phase. Among these methods, the proposed system achieves higher reconstruction quality in terms of both PSNR and MS-SSIM when the target bitrate is satisfied. With respect to our submission in the test phase, it achieves 0.9968 in MS-SSIM and 41.547dB in PSNR.

The visual quality comparisons are provided in Figure 6 and Figure 7, where the tested frames are compressed by HEVC, VVC and the proposed method respectively. From Figure 6, it is observed that the proposed method can achieve visually much better quality, especially for the regions rich with textures. From Figure 7, blocking artifacts are observed in the decoded frames of HEVC and VVC, while it is suppressed in the frames compressed by the proposed method, and such visual benefits mainly come from the proposed joint spatial-temporal CNN filter.

Table 1. Evaluation results on CLIC2020 validation dataset.

Team	MS-SSIM	PSNR	Data size
TUCODEC_SSIM	0.9969	37.309	37870015
IMCL_MSSSIM	0.9968	37.309	37960950
ZJUCSEFj	0.9967	36.585	38113147
DAMO_XG	0.9966	41.158	37312334
EDVC	0.9961	37.029	37941105
HUST_ZX	0.9960	42.344	38701571
Man	0.9959	41.811	38587535
Dolores_baseline	0.9953	41.563	31223104
....	....	....	....

4. Conclusions

In this paper, a novel joint learned and traditional video compression scheme is proposed for the P frame track in CLIC2020 challenge. We first propose a decoder-side template-based inter prediction method to predict the current block without signaling overhead of motion information. Secondly, a spatial-temporal CNN post filter is proposed to suppress visual artifacts and improve the decoded image quality. Furthermore, an advanced SSIM based rate-distortion optimization model is proposed to achieve best balance between the coding bits and the decoded image quality. Experimental results show that the proposed method can achieve higher reconstruction quality.

References

[1] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003.

[2] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

[3] Chen Jianle, Ye Yan, and Kim Seung Hwan. Algorithm description for Versatile Video Coding and Test Model (VTM6). *Doc. JVET-O2002, Joint Video Exploration Team (JVET)*, 2019.

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[5] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, pages 1141–1151, 2017.

[6] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. *arXiv preprint arXiv:1809.10452*, 2018.

[7] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019.

[8] Jianle Chen, Elina Alshina, and Gary J Sullivan. Algorithm description of Joint Exploration Test Model 7 (JEM7). *Doc. JVET-G1001, Joint Video Exploration Team (JVET)*, 2017.

[9] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Gary J Sullivan and Thomas Wiegand. Rate-distortion optimization for video compression. *IEEE signal processing magazine*, 15(6):74–90, 1998.