Hybrid video coding scheme based on VVC and spatio-temporal attention convolution neural network

Gang He[12†] Kepeng Xu[1†✉] Chang Wu[1] Zijia Ma[1] Xing Wen[2] Ming Sun[2]
[1] Xidian University
[2] Kuaishou Technology
kepengxu11@gmail.com

Abstract

In this paper, we propose a hybrid video coding framework. The framework is built on the basis of VVC (Versatile Video Coding) video coding standard and constructs an implicitly aligned multi-frame fusion model to accomplish subjective video quality enhancement. The proposed framework mainly optimizes video compression efficiency from two perspectives. First is the sequence-level dynamic rate control algorithm, which assigns the appropriate bitrate to each video to obtain the highest overall video quality. Second is the MAQE, a multi frame implicit alignment video quality enhancement model, which performs motion alignment through multiple convolutional kernels of different sizes, uses a residual aggregation layer to fuse features of different frames, and then uses an enhanced attention module to adaptively deflate features based on spatio-temporal contextual features, so as to more effectively fuse feature of multiple frames and obtain higher quality reconstructed frames. The proposed method is validated on two tracks of 0.1M code rate and 1M code rate on CLIC-2022 video compression task, Experimental results show that the proposed method achieves PSNR of 30.301 and 37.251 and obtains MS-SSIM of 0.9368 and 0.9875. This paper is a comprehensive presentation of the scheme used by the Night-Watch team of the CLIC-2022 video track.

1. Introduction

In the last decade, video content has accounted for an increasing share of traffic in the Internet, and as a result a large number of video compression techniques have been developed to compress video as much as possible, which include standards such as H263 [6],H264 [8],H265 [7],H266 [1],AVS [3] and AV1 [2]. Since H265, the code stream size of video content has been reduced as much as possible by building hybrid video coding methods that use a number of different tools for lossy encoding of video. These mentioned past standard methods compress video mainly by, extracting features, quantizing features, and entropy coding with PSNR (peak signal to noise ratio) as the optimization goal. In such an optimization model, MSE (Mean Square Error) is usually used as the fidelity measure to calculate R-D Loss, but since the subjective human perceptual quality is not absolutely proportional to PSNR, the optimal perceptual quality is not obtained by using PSNR as the optimization target exclusively. To more accurately evaluate the competent quality of video images, the multiscale structural similarity metric is widely accepted by researchers.

In order to effectively improve the subjective quality of reconstructed videos, this paper proposes a hybrid compression strategy that can obtain a subjective quality exceeding that of VVC. In this paper, first video sequences are dynamically bit-rate assigned to obtain the highest overall evaluation metric. And then an implicit motion alignment-based video subjective quality enhancement model is constructed. Such a hybrid video compression strategy can effectively improve the subjective video quality. The main contribution classes of this paper are summarized in the following points.

- We proposes the sequence-level dynamic rate control algorithm that can obtain the highest quality video with the overall bitrate of multiple videos less than the upper limit of the track and obtain the best R-D tradeoff.

- We propose a deep neural network model to enhance the subjective quality of the reconstructed video by implicit alignment, multi-frame fusion, and multiple residual feature attention modules.

✉ Corresponding author
† Equal contribution

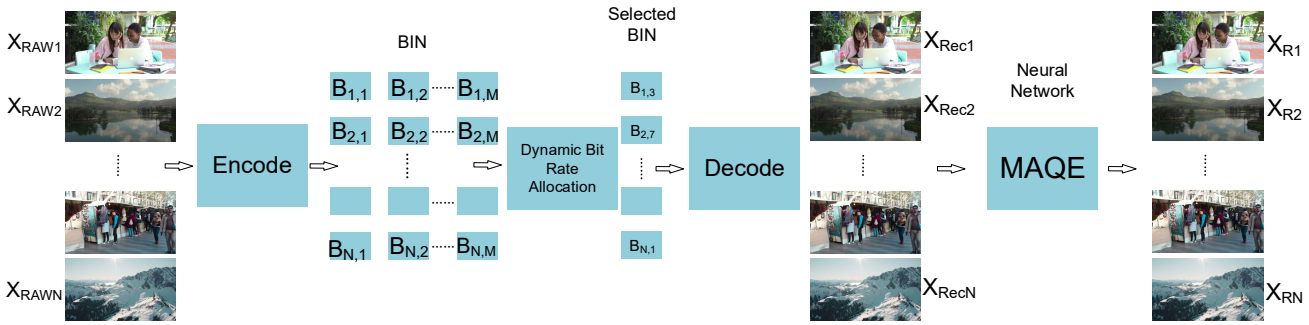

Figure 1. Framework of our solution. For the validated videos released by CLIC2022, we first encode each video using different QPs (using VTM) and next select the appropriate QP for each video by the proposed dynamic code rate allocation strategy, after which each video has a selected BIN stream. We transfer the selected stream to the decoder side and decode it using the decoder. Then the video quality is enhanced using our proposed MAQE model to obtain the final video.

2. Methodology

2.1. Framework

In this section, we introduce the proposed sequence-level dynamic rate control algorithm and subjective quality enhancement model. First we introduce the overall framework, the whole framework is shown in Figure 1. For the source video X_{RAWi} (total N videos), we first obtain the bitstream B_{ij} (i-th video, j-th QP) by encoding X_{RAW} in multiple steps (using several different QPs, total M QPs) using the VVC encoder. Next, B_{ij} is decoded using the VVC decoder to obtain X_{Recij}, and the bitstream size and PSNR are calculated for each reconstructed video. Given the overall upper limit of the bitrate, the optimal bitstream combination (OB_{ij}) is selected using the proposed sequence-level dynamic rate control algorithm, where one QP is selected for each video, so that we obtain the best bitrate combination. Next, the selected bitstreams OB_{ij} are transmitted to the decoder side for decoding and post-processing enhancement. At the decoder side, we first use the VVC decoder to decode the bitstream OB_{ij} to obtain the reconstructed video X_{Recij}, and next use the proposed MAQE to perform quality enhancement on X_{Recij} to output the final video sequence X_{Ri}.

The entire video encoding process can be formalized as a formula. At the encoder side, the source video is first encoded with multiple different QPs using VVC, and then the code rate is assigned using the sequence-level dynamic rate control algorithm (SDRC).

$$\begin{aligned} B_{ij} &= VVCEncoder(X_{RAW}) \\ OB_{ij} &= SDRC(B_{ij}, X_{RAW}) \end{aligned} \tag{1}$$

Decode the video at the decoder side using VVC and quality enhance the video using convolutional neural network model.

$$\begin{aligned} X_{Recij} &= VVCDecoder(BS_{ij}) \\ X_{Ri} &= MAQE(X_{Recij}) \end{aligned} \tag{2}$$

2.2. Sequence-level dynamic rate control

In the video track of CLIC-2022 challenge, for a given set of validation sequences, the evaluation metric is subjective coding quality under average bitrate constraints, including 0.1mbps and 1mbps. To facilitate the implementation of the algorithm, we introduce quantifiable objective metric, i.e., PSNR. This task can be formulated as :

$$\begin{aligned} \{B\} &= \arg\max_{\{B\}} \sum_{i=1}^{N} V_i \\ s.t. S &\leq N \cdot T \cdot R_{avg} \end{aligned} \tag{3}$$

where B is the set of the encoded bitstream. N is the number of sequences in the validation set. V_i represents the average PSNR of the ith sequence. S is the total size of all encoded bitstream. T is the duration of each sequence and it is fixed as 10. R_{avg} is the limited average bitrate.

Therefore, the task can be regarded as solving a dynamic programming problem. We propose a sequence-level dynamic rate control algorithm to achieve the trade-off between coding quality and bitrate (see Alg.1). First of all, each sequence is encoded with various quantization parameters (QPs) and calculate the PSNR and corresponding size of the bitstream under each QP. Generally, a larger bitstream size indicates a higher PSNR and the rising slope is gradually decreasing. And then, by setting a reasonable threshold λ and judging whether the slope is below it, the most suitable encoding settings are selected for each sequence through iterative optimization and ensure that the total size of all bitstream does not exceed the limit.

2.3. Multi-frame fusion quality enhancement model

Inspired from deep learning-based video recovery and hyper-segmentation methods, we propose a multi-frame fusion convolutional neural network model MAQE applied to the video compression post-processing stage. We will present the architectural composition of MAQE here, and the overall framework is shown in Figure 2.

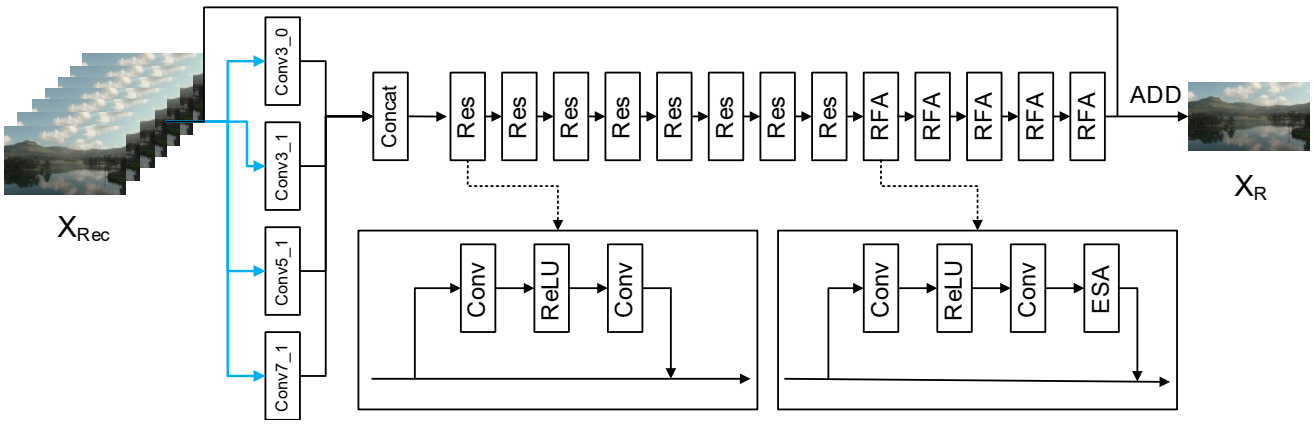

Figure 2. Architecture of proposed MAQE. The current frame and the six neighboring frames are input to the model simultaneously, and the spatio-temporal information is first extracted using convolutional kernels with kernel size of 3, 3, 5, 7 and dilation size of 0, 1, 1, 1 to complete the implicit feature alignment. Next, it is input to 8 residual blocks, followed by 4 RFA modules, which are able to perform spatio-temporal attention to better extract information useful for the current frame. Finally the reconstructed frames are output after vanilla convolution and global skip connection.ESA module is the same as [5].

Algorithm 1: Sequence-level dynamic rate control

Data: largest QP Q_l, smallest QP Q_s, the average PSNR under QP V_Q, the size of bitstream under QP S_Q, threshold λ
Result: Optimal QP Q_t

```
for Q ← Q_l to Q_{s+1} do
    k ← (V_{Q+1} − V_Q)/(S_{Q+1} − S_Q);
    if k>λ then
        continue;
    else
        Q_t ← Q;
        break;
    end
end
```

For the video enhancement task, due to the existence of spatio-temporal information correlation in videos, constructing a spatio-temporal information interaction mechanism can effectively enhance the enhancement capability of the model. Thus, in this paper, we design a simple multi-frame restoration model to capture motion information through convolutional kernels with different void rates, so as to implicitly accomplish motion alignment and effectively fuse video multi-frame information. After extracting multi-frame information, extracting effective information from multi-frame information can enhance video recovery, so the model needs to be able to perform differential feature extraction for different regions of different frames, and to achieve this function, we use a residual module based on the spatial residual attention mechanism. In order to achieve this function, we use a residual module based on the spatial residual attention mechanism.

Our specific implementation is described next. Firstly, the decoded video frame X_{Rec} is passed through Conv3 $\times$ 3 with dilation 0, Conv3 $\times$ 3 with dilation 1, Conv5 $\times$ 5 with dilation 1 and Conv7 $\times$ 7 with dilation 1 convolution kernels, and then the obtained features are stitched. Then, the features are input to 8 residual modules in series, through which the advanced features of video frames can be extracted and the spatio-temporal information of video frames can be fused, and then the output features are input to 4 RFA [5] residual modules, which can perform spatio-temporal attention to the input video frame information and weight the features in different time spaces, so that the fidelity features in different frames can be extracted more specifically. The RFA module can perform spatio-temporal attention on the input video frame information and weight the features in different time spaces, so that the image features with high fidelity in different frames can be extracted in a more targeted manner, which can reduce the artifacts brought by multi-frame fusion. The final output X_R is obtained by summing the input frames after a vanilla convolution with a kernel size of 3.

3. Experiments

3.1. Experimental setup

We choose the validation set provided by CLIC2022 to evaluate the proposed method, and choose MS-SSIM and PSNR as the evaluation metric. For the first half of the coding framework we choose VTM, a standard implementation of VVC, as the codec, and the coding configuration are set in table 1. In the training process of the deep neural network, we choose the dataset provided by MFQE2.0 [4] to

Option	Description
–InputFile	Selects the input file
–BitstreamFile	Path of bistream file
–SourceWidth	Video width
–SourceHeight	Video height
-c encoder randomaccess vtm.cfg	Coding configuration
–IntraPeriod=-1	Intra Period: A single Intra frame is selected
–QP qp	Value of the quantization parameter
–SliceChromaQPOffsetPeriodicity=1	Periodicity for inter slices that use the slice-level chroma QP offsets
–PerceptQPA=1	Applies perceptually optimized QP adaptation

Table 1. VTM configuration.

Method	0.1M PSNR	0.1M MS-SSIM	1M PSNR	1M MS-SSIM
VTM	29.536	0.9324	36.467	0.9832
Ours	30.301	0.9368	37.251	0.9875

Table 2. Quantitative Results.We tested the CLIC2022 at 0.1M and 1M bitrate track respectively.

Method	Params(M)	Runtime(ms)
MAQE	1.93	570

Table 3. Parameter size and inference time of MAQE.

train the model. We use the RAdam optimizer ,the initialized learning rate is set to 0.0001, and the learning rate decreases to 1/10 of the original one every 20,000 iteration, and MS-SSIM is used as the loss function.

3.2. Results

To validate the effectiveness of each component of the proposed framework, we evaluate the proposed dynamic bitrate control algorithm and deep learning post-processing model separately. As seen in Table 2, the proposed video compression method shows improvement over VTM on MS-SSIM. The proposed video compression method shows better performance than VTM at every bitrate.

In addition, we further calculate the inference speed of the proposed enhanced model as shown in Table 3, where the inference times are calculated using a Tesla P100 GPU at 1080P resolution.

4. Conclusion

In this paper, we propose a two-stage hybrid video coding framework that can significantly improve the competent quality of encoded videos by effectively enhancing the video coding efficiency through a sequence-level dynamic rate control algorithm with a post-processing subjective quality enhancement model. The code rate allocation in the proposed scheme can improve the objective quality of the video with limited overall code rate. An implicit multi-frame aligned quality enhancement model is proposed and incorporates a spatial residual adaptive attention module to enable the model to effectively identify multi-frame quality differences and selectively perform feature fusion to further improve video quality.

References

[1] Benjamin Bross, Jianle Chen, Jens-Rainer Ohm, Gary J Sullivan, and Ye-Kui Wang. Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc). *Proceedings of the IEEE*, 109(9):1463–1493, 2021. 1

[2] Yue Chen, Debargha Murherjee, Jingning Han, Adrian Grange, Yaowu Xu, Zoe Liu, Sarah Parker, Cheng Chen, Hui Su, Urvang Joshi, et al. An overview of core coding tools in the av1 video codec. In *2018 Picture Coding Symposium (PCS)*, pages 41–45. IEEE, 2018. 1

[3] Liang Fan, Siwei Ma, and Feng Wu. Overview of avs video standard. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 1, pages 423–426. IEEE, 2004. 1

[4] Zhenyu Guan, Qunliang Xing, Mai Xu, Ren Yang, Tie Liu, and Zulin Wang. Mfqe 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):949–963, 2019. 3

[5] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020. 3

[6] Karel Rijkse. H. 263: Video coding for low-bit-rate communication. *IEEE Communications magazine*, 34(12):42–45, 1996. 1

[7] 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. 1

[8] 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. 1