Beyond VVC: Towards Perceptual Quality Optimized Video Compression Using Multi-Scale Hybrid Approaches

Zhimeng Huang[1*], Kai Lin[1*], Chuanmin Jia[1*], Shanshe Wang[1], Siwei Ma[1,2]
[1]Department of Computer Science, School of EE&CS, Peking University, Beijing 100871, China
[2]Information Technology R&D Innovation Center of Peking University, Shaoxing 312000, China
{zmhuang, kailin, cmjia, sswang, swma}@pku.edu.cn

Abstract

In this paper, we propose a perceptual quality optimization oriented video compression framework using hybrid approaches. The proposed framework, which is built on the top of recently-published Versatile Video Coding (VVC), contains multi-scale optimized coding techniques. Specifically, three major aspects of efforts from coding unit level to video sequence level have been dedicated to obtain substantial compression efficiency improvement. We first propose a block-level rate-distortion optimization (RDO) method with the consideration of block artifacts removal. Subsequently, we propose frame-level perceptual quality optimized convolutional neural networks for the post-processing of each compressed image, within which the channel attention mechanism has been employed to capture and restore the crucial detail in subjective evaluation. We additionally model the bit allocation as sequence-level dynamic programming problem such that optimal perception and bitrate tradeoff could be obtained. Experimental results show that the proposed method achieves 0.98658 MS-SSIM on the validation set in video track of CLIC-2021.

1. Introduction

Over the past three decades, video compression has embraced significant technology innovations and advancement. H.26x [1, 7, 10], AVS [4, 11] and AOM [2] series standards basically indicate the most widespread influenced video coding standards in this context. Plenty of representative coding tools were built on the top of hybrid video coding principle. More concretely, such hybrid video coding refers to the combination of multiple means to reduce different types of redundancy within the video signals, that is, block-based prediction plus transform coding with scalar quantization of the prediction residual. Evidence has been extensively shown that locally optimizing the predictive coding or transform coding would obviously benefit the rate-distortion (R-D) efficiency.

However, the aforementioned video coding standards tend to optimize the entire hybrid video coding framework based on the signal fidelity. The pixel level based objective quality mean square error (MSE) is usually adopted as the distortion metric in video coding, resulting in mismatch between the subjective and objective perceptions. Compared to MSE, multi-scale structure similarity (MS-SSIM) index [9] has higher relationship with the subjective quality, and it is also deployed as a widely accepted distortion measurement to reflect the degree of human visual satisfaction. Existing coding tools often does not support MS-SSIM based optimization in recent published video coding standards.

As such, we propose a MS-SSIM optimization oriented video coding framework in this paper to surpass the coding performance of VVC in terms of perceptual quality. In particular, this paper considers the subjective quality inspired video coding in a hierarchical fashion, from the sequence level to the coding unit (CU) level. These advanced techniques are proposed to adapt to the essential factors of MS-SSIM metric by leveraging both the recent deep learning based approach and conventional R-D optimization (RDO) approach. This paper is a comprehensive technical description of the team *DWH* in the video track of CLIC-2021 grand challenge. The major contribution of this paper can be summarized as follows.

- At CU level, we propose to utilize a novel RDO method with the consideration of block artifacts removal, which improves the structural similarity by suppressing the boundary effects.
- At picture level, we propose a novel convolutional neu-

This work was supported in part by the China Postdoctoral Science Foundation (2020M680238), the National Natural Science Foundation of China (62072008, 61931014) and High-performance Computing Platform of Peking University, which are gratefully acknowledged. * equal contribution.

ral network (CNN) combined with channel attention mechanism for frame-level post processing.

- At sequence level, we propose and model the bits allocation for each video based on dynamic programming such that the optimal R-D trade-off could be obtained for this challenge.

2. Methodology

In this section, we give a comprehensive presentation of the proposed multi-scale optimization methods from the sequence level bit allocation method, the frame level post processing network and the CU level RDO approach.

2.1. Optimal Bit Allocation

We propose an optimal bit allocation trade-off between bit-rates and perception distortion in sequence level. We first extensively analyzed the characteristics of the content in the video track of CLIC-2021 challenge, in which the videos are divided into several different categories, (e.g., animation, gaming and virtual reality related videos). Subsequently, we formulated such trade-off as a dynamic programming problem and a recursion based algorithm is proposed to calculate the optimal bit allocation strategy given the provided validation sequences.

In the video track of CLIC-2021, the evaluation metric is defined as the average MS-SSIM weighted by pixels with a limitation of the sum of data size and decoder size, which could be formulated as:

$$\{B, D\}_{opt} = \arg\max_{\{B,D\}} \frac{\sum n_{i,j} * M_{i,j}}{\sum n_{i,j}} \quad s.t. R \leq R_t \qquad (1)$$

where B is the set of the bitstream and D indicates the file size of decoder. $n_{i,j}$ and $M_{i,j}$ represent the number of pixels and the MS-SSIM for the i_{th} frame in the j_{th} sequence respectively. R_t denotes the maximum space limit containing both the bitstream and the decoder. Specifically, R is calculated as the weighted mean of the size of the bitstream and the decoder in this challenge:

$$R = R_b/0.019 + R_d. \qquad (2)$$

For a certain decoder, R_d is regarded a constant value. Such that Eqn. 1 could be expressed as:

$$\{B\}_{opt} = \arg\max_{\{B\}} \sum n_{i,j} * M_{i,j} \quad s.t. R_b < R_c, \qquad (3)$$

$$R_c = (R_t - R_d) * 0.019. \qquad (4)$$

To solve the above constrained optimization problem, a dynamic programming algorithm is proposed to obtain the optimal solution. First, we encode each sequence with various quantization parameters (QP) to obtain bitstreams under different situation. Then we derive the MS-SSIM value $value$ of the compressed videos and corresponding size of the bitstream $cost$, separately.

Notation	Explanation
$\mathcal{F}_{i,j}$	The optimal weighted MS-SSIM for the first ith sequences with space cost j.
$cost_{i,j}$	The space cost of sequence i with coding parameter j.
$value_{i,j}$	The MS-SSIM of sequence i with coding parameter j.
$\mathcal{N}_i$	The number of pixels in sequence i
$L_{i,j}$	The number of the chosen coding parameters to gain $\mathcal{F}_{i,j}$.
P	The set of coding parameters
N	The number of the sequences in the validation dataset

Table 1. Notations for optimal bit allocation in this paper

The notations used in the rest of the subsection are summarized in Table 1. The initialization of the dynamic programming is formulated as Eqn. 5.

$$\mathcal{F}_{1,j} = \mathcal{N}_1 * \max\{value_{1,k} | \; cost_{1,k} \leq j, k \in P\} \qquad (5)$$

The core state transformation can be stated as Eqn. 6. After the initialization of the first sequence, $\mathcal{F}_i$ can be derived from $\mathcal{F}_{i-1}$ recursively as follows.

$$\mathcal{F}_{i,j} = \max\{\mathcal{F}_{i-1,j-cost_{i,k}} + \mathcal{N}_i * value_{i,k} | \; k \in P\}. \qquad (6)$$

At the same time, we utilize $L_{i,j}$ to record the chosen coding parameters with the update of variable $\mathcal{F}_{i,j}$:

$$L_{i,j} = \arg\max_{k}\{\mathcal{F}_{i-1,j-cost_{i,k}} + \mathcal{N}_i * value_{i,k} | \; k \in P\}. \qquad (7)$$

$$B_{opt} = \arg\max_{\{B\}}\{\mathcal{F}_{N,r} | \; r < R_c\}. \qquad (8)$$

After $N-1$ iterations, the maximum weighted average MS-SSIM denoted as the $\mathcal{F}_{N,r}$ over the validation can be obtained. At the same time, the bits are assigned to each sequence according to the optimal bit allocation strategy.

2.2. Channel Attention Network

Motivated by the recent development of CNN based image restoration methods, we propose a frame-level perceptual quality optimized CNN as the post processing technique. In particular, the proposed network is separated from the coding loop for MS-SSIM enhancement. As shown in Fig. 1, the reconstruction frames after compression are fed into the network sequentially, and enhanced frames are subsequently generated by the proposed network. We build the network using two kinds of components, which are channel attention modules and residual units. There is also a global link from the beginning to output of the network.

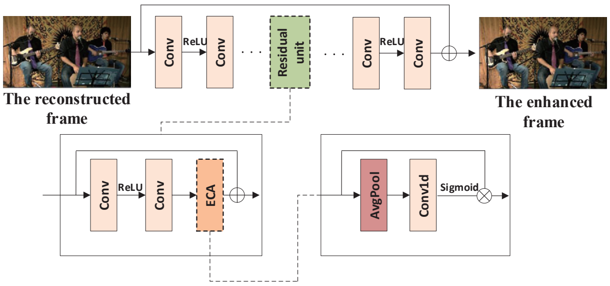

Figure 1. The network architecture of the post processing neural network.

Given the prior knowledge in CNN based restoration, the channel attention module is of great potential to further improve the feature aggregation and representation ability of deep networks. Through the channel-wise interactive learning, the importance of each feature map can be derived. Subsequently, the learnt importance map is multiplied back to the initial feature maps. As such, the meaningful feature maps are retained and highlighted while irrelevant information is suppressed. Considering the trade-off between performance and extra parameters increment, we apply the lightweight Efficient Channel Attention (ECA) module [6] into our post processing network design, yielding a hybrid network containing both residual units and ECA module.

Besides the channel attention module, the residual learning strategy is another principle when designing the post processing network. Similar with the early explorations of in-loop filtering [3], the backbone of the post processing network is made up of several residual units. More specially, the residual unit consists of a local shortcut and two consecutive convolutional layers separated by a Rectified Linear Units (ReLU) [5]. Besides described factors, the ECA module is placed in the residual unit before the local shortcut. The details of the network architecture is shown as Fig. 1.

As aforementioned, the post processing is performed at frame level for each compressed image. In order to maximize the effects of the network, we design a frame level switch flag to indicate the on/off of if. Only when the post processing module improves the perceptual quality of current frame, the frame-level decision is set to be 1. To ensure the consistency between encoder and decoder, the binarized frame-level decisions are also required to signalled. Note that they are separately transmitted with the video bitstreams.

2.3. Perceptual Based RDO for Each CU

In the current encoder side design of VVC, the deblocking filter is not considered within the coding loop. However, the deblocking filter plays an essential part in improving perceptual quality for video compression. The boundary effects suppression would definitely benefit the the MS-SSIM based evaluation because less blocking artifacts will smooth the textures and signal variance could be reduced. In this context, it is desirable to adopt the block artifacts removal when encoding each CU. In our proposed framework, we add the deblocking filter to the full RDO process of each CU, which means that the boundary of each CU is mandatory to be filtered after CU reconstruction in the coding loop. With the approach, the block artifacts could be suppressed and MS-SSIM score could be improved.

3. Experiments

The experimental results of the proposed perceptual quality oriented video coding framework are provided in this section. Specifically, we follow the test conditions defined by the CLIC-2021 challenge and report our results.

3.1. Training Details

To evaluate the performance of the post processing module, we build the network by 10 residual units, and the number of feature maps is set as 64. The network construction and training process are implemented based on Pytorch.

We utilize the provided 462 video sequences as originals and compressed them with VTM-12.0 [1] under RA configuration. The quantization parameter (QP) is set as 37, and the other necessary parameters are set as default during the compression process. As such, pair wise training samples containing compressed frame and ground truth are generated. In this paper, we use MS-SSIM as the loss function to update the network parameters. Adam is adopted as the optimization method during the training process. The learning rate is set to be 1e-4 initially and decreases to 1e-5.

3.2. R-D Performance on Validation Set

3.2.1 Validation Dataset Description

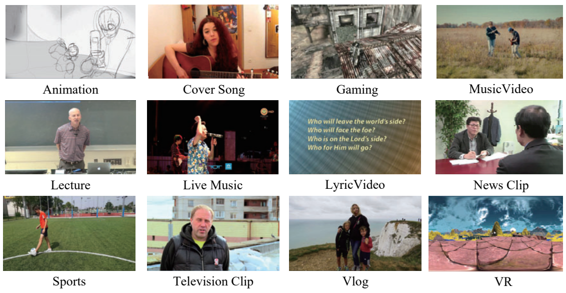

Figure 2. The thumbnail images of typical sequences in validation set.

The validation dataset consists of 562 videos (314,175 frames) taken from the UGC dataset [8]. The videos could be divided into 13 categories: Animation, Cover Song, Gaming, How To, Lecture, Live Music, Lyric Video, Music Video, News Clip, Sports, Television Clip, Vlog, and VR.

As shown in Fig. 2, videos in different categories show various visual characteristics. For example, Animation videos usually contain more smooth edges, while Vlog videos may have more complex textures; Sports and Game videos always perform lots of motion estimation but the backgrounds of Cover Song or Lecture tend to keep still. Considering these different characteristics, the sequence level bit allocation strategy performs huge improvement on the validation set.

3.2.2 Performance

The proposed video compression approach shows an upgrade on MS-SSIM over VTM-12.0. Since CLIC-2021 limits the data size of coded files, we utilize the sum of the consuming space to represent the bitrate. And the evaluation metric MS-SSIM is regarded as the distortion measurement. The R-D performance curve is shown in Fig. 3. The proposed video compression approach shows better performance than VTM-12.0 at every bitrate. Note that the proposed framework is able to generate different bit-rate points. Specifically, R_d is 9608.5 KBytes in our submission. According to Eqn. 4, R_c and the corresponding MS-SSIM are 24108.58 KBytes and 0.9865, separately. Although a channel attention network is embedded into our decoder, it only cost 1688 seconds to decode the validation dataset according to the leaderboard[1].

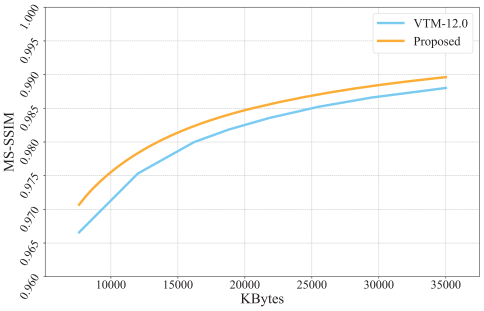

Figure 3. The comparison between the proposed algorithm and VTM-12.0

Table 2 shows the performance for each tool. Tool1 indicates the proposed bit allocation mentioned in Section 2.1. Tool2 represents the channel attention network illustrated in Section 2.2. And Tool3 is the perceptual based CU RDO proposed in Section 2.3.

Table 2. Coding performance for each proposed tool

Method	Performance
VTM 12.0	0.98452
VTM+Tool1	0.98630
VTM+Tool1+Tool2	0.98648
VTM+Tool1+Tool2+Tool3	0.98658

4. Conclusion

This paper presents a novel MS-SSIM oriented video compression framework by leveraging both the state-of-the-art coding standard VVC and multi-scale optimization coding tools. The proposed codec consistently optimizes the coding efficiency under the MS-SSIM quality. In this regard, three major advanced techniques from CU level to sequence level are proposed. The block artifact removal was particularly considered within the RDO process of each CU. Subsequently, we proposed a channel attention based post processing network to enhance the subjective quality. The optimal bit allocation was finally modeled as dynamic programming for each sequence.

References

[1] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang. Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc). *Proceedings of the IEEE*, 2021. 1, 3

[2] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, et al. An overview of core coding tools in the av1 video codec. In *Picture Coding Symposium (PCS)*, pages 41–45. IEEE, 2018. 1

[3] K. Lin, C. Jia, Z. Zhao, L. Wang, S. Wang, S. Ma, and W. Gao. Residual in residual based convolutional neural network in-loop filter for avs3. In *Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2019. 3

[4] S. Ma, T. Huang, C. Reader, and W. Gao. Avs2? making video coding smarter [standards in a nutshell]. *IEEE Signal Processing Magazine*, 32(2):172–183, 2015. 1

[5] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 3

[6] W. Qilong, W. Banggu, Z. Pengfei, L. Peihua, Z. Wangmeng, and H. Qinghua. Eca-net: Efficient channel attention for deep convolutional neural networks. 2020. 3

[7] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. 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] Y. Wang, S. Inguva, and B. Adsumilli. Youtube ugc dataset for video compression research. In *International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 3

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402. IEEE, 2003. 1

[10] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. 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

[11] J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma, and W. Gao. Recent development of avs video coding standard: Avs3. In *Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2019. 1

[1] http://compression.cc/leaderboard/video/valid/