VIDEO COMPRESSION USING RATE-DISTORTION OPTIMIZATION-BASED CHANNEL ATTENTION NETWORK

[1]B.Shravan Kumar, [2]Dr. V.Usha Shree, [3]Dr. Sumagna Patnaik

[1]Reasearch Scholar, JNTUH, Hyderabad, TS.

shravanbk6@gmail.com

[2]Principal & Professor of CSE department, Swami Vivekananda Institute of Technology, Secunderabad, TS.

valasani_usha1@yahoo.com

[3]Professor IT Department and DEAN IQAC, J.B. Institute of Engineering and Technology, Hyderabad, TS

ABSTRACT:

In this study, we provide a hybrid video compression methodology focused on perceptual quality optimization. Multi-scale optimal coding methods are used in the proposed framework, which is based on the newly released Versatile Video Coding (VVC). From the coding unit level to the video sequence level, specific attention has been paid to three key areas in order to significantly enhance compression efficiency. In order to remove block artefacts, we first suggest a block-level rate-distortion optimization (RDO) approach. Post-processing of each compressed image is then handled by convolutional neural networks with frame-level perceptual quality optimizations that use a channel attention method to capture and restore the key information in subjective assessment. To get the greatest balance between quality and bit rate, it is important to treat bit allocation as a dynamic programming problem. According to experimental data, the validation set in the video track of the proposed technique by CLIC-2021 achieved an MS-SSIM of 0.98658.

Keywords: Versatile Video Coding, rate-distortion optimization, and channel attention network are other related terms.

DOI Number: 10.14704/nq.2022.20.11.NQ66270 **NeuroQuantology 2022; 20(11): 2678-2683**

1. INTRODUCTION

Video compression has seen considerable technological advancements during the last three decades. In this regard, the most widely used and affected video coding standards are often indicated by the H.26x [1, 7, 10], AVS [4, 11], and AOM [2] series standards. On top of the hybrid video coding idea, several representative coding tools have been developed. Block-based prediction and transform-based coding, as well as scalar quantization of the prediction residual, are just some of the approaches used in this type of video coding. Refined predictive coding or transform coding has been shown to improve local rate-distortion (R-D) efficiency in numerous investigations. The above-mentioned video coding standards, on the other hand, tend to optimise the hybrid video coding architecture as a whole depending on the signal quality. In video coding, the distortion measure is frequently the pixel level-based objective quality mean square error (MSE), which generates a chasm between subjective and objective perceptions. Using the MS-SSIM index ([9]), which has a better link to subjective quality than MSE, is a typical distortion assessment to determine the level of visual pleasure in humans [10]. In recent video coding standards, MSSSIM-based optimization is often not possible with current coding tools. MS-SSIM video coding uses optimization techniques to outperform VVC's coding in terms of perceptual quality, as seen here. From the sequence level down to the coding unit (CU) level, this research primarily studies hierarchical subjective video

coding. The MSSSIM metric components can be adapted to these new tactics utilising both the current deep learning strategy and the traditional R-D optimization (RDO) approach. We've put together a detailed technical breakdown of our team's performance at CLIC-2021 Grand Challenge's video track in this article. An overview of this paper's most important findings follows.

At the CU level, we suggest using a cutting-edge RDO technique that takes into account the elimination of block artefacts. This method reduces boundary effects and increases structural similarity.

- For frame-level post processing, we suggest an unique convolutional neural network (CNN) in combination with a channel attention mechanism at the image level.
- Based on dynamic programming, we suggest and model the bit allocation for each video at the sequence level to achieve the best R-D trade-off for this issue.

2. METHODOLOGY

The suggested multi-scale optimization approaches from the frame level post processing network, the CU level RDO approach, and the sequence level bit allocation method are all presented in detail in this part.

2.1. Optimal Bit Allocation

The best bit allocation approach is provided to balance bit rates and perceptual distortion at the sequence level. A detailed examination of the CLIC-2021 challenge movies was conducted first, in which the films were divided into a wide range of aspects (e.g., animation, gaming and virtual reality-related videos). After that, we turned this trade-off into a dynamic programming issue, and we suggested a recursion-based approach to determine the best bit allocation technique given the supplied validation sequences. The average MS-SSIM weighted by pixels with a restriction on the total of data size and decoder size is the evaluation measure for the video track of CLIC-2021, which may be expressed as:

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

where D denotes the decoder's file size and B the set of the bitstream. The values ni, j and Mi, j denote the quantity of pixels and MS-SSIM for the jth sequence's ith frame, respectively. The bitstream and the decoder can fit in a maximum amount of space (Rt). In this challenge, R is computed as the weighted average of the size of the bitstream and the decoder:

R = Rb/0.019 + Rd. (2)

Rd is seen as a constant value for a certain decoder. such that the expression for Eqn:

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

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

Table 1. Notations for optimal bit allocation in this paper

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

A dynamic programming approach is suggested in order to find the best solution for the restricted optimization issue mentioned above. To produce bitstreams under varied conditions, we first encode each sequence using a variety of quantization settings (QP). Then, we individually calculate the MS-SSIM value of the compressed movies and the related bitstream cost size. Table 1 provides an overview of the notations used throughout the remaining section. Eqn. 5 is how the dynamic programming's initialization is expressed.

F1,j = N1 ∗ max{value1,k| cost1,k ≤ j, k ∈ P} (5)

eISSN 1303-5150

Eqn. 6 may be used to express the fundamental state transition. Fi may be generated recursively from Fi1 after the initialization of the first sequence in the manner shown below.

$$F_{i,j} = \max\{F_{i-1,j-costi,k} + N_i * valuei,k | k \in P\}. (6)$$

In parallel, we update variable Fi,j while using Li,j to record the selected coding parameters:

$$L_{i,j} = \arg\max k \{F_{i-1,j-costi,k} + N_i * valuei,k | k \in P\}. (7)$$

$$Bopt = \arg\max \{B\} \{FN,r | r\} (8)$$

The maximum weighted average MSSSIM, also known as the FN,r over the validation, may be attained after N 1 rounds. The best bit allocation approach is simultaneously used to allocate bits to each sequence.

2.2. Channel Attention Network

As a post processing method, we propose using a frame-level perceptual quality optimised CNN, which is based on the recent development of CNN-based picture restoration algorithms. In particular, the coding loop for MS-SSIM improvement is cut off from the proposed network. According to Fig. 1, following compression, the reconstruction frames are sent progressively into the network, which then produces improved frames. The channel attention modules and the residual units are the two different types of components we use to construct the network. Additionally, the network has a global connection that connects its origin and conclusion.

Given the past experience with CNN-based restoration, the channel attention module has a great deal of promise to enhance deep networks' capacity for feature aggregation and representation. The significance of each feature map may be ascertained via channel-wise interactive learning. An additional multiplicative step returns the newly-acquired map's meaning to the original feature maps. Maps with key features are highlighted, while those with less important data are buried. Efficient Channel Attention (ECA) is incorporated into our post-processing network architecture, considering the trade-off between performance and additional parameters. Residual units and the ECA module will be in this hybrid network. In addition to the channel attention module, the residual learning approach is considered when designing the post-processing network. The remaining units form the backbone of the post-processing network, comparable to the early studies on in-loop filtering [3].

In the residual unit, a Rectified Linear Unit (ReLU) separates a local shortcut and two subsequent convolutional layers [5]. The ECA module is added to the residual unit before the local shortcut, along with the other components already described. Figure 1 depicts the specifics of the network design. The post processing is carried out for each compressed picture at the frame level, as was already described. A frame-level switch flag is created to indicate whether or not if is on or off in order to maximise network benefits. It is only when the post-processing module improves the perceived quality of an image that it is set to 1. The binaryized frame-level choices must also be notified in order to guarantee consistency between encoder and decoder. They are broadcast independently along with the video bitstreams, as you should know.

eISSN 1303-5150

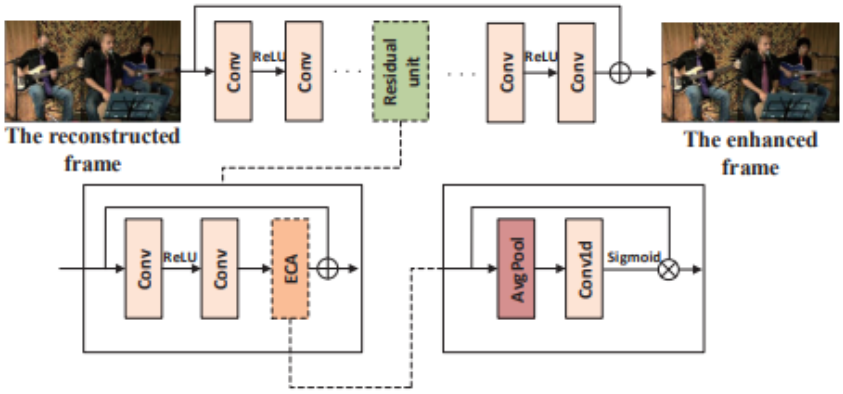

Figure 1. The network architecture of the post processing neural network

2.3. Perceptual Based RDO for Each CU

The deblocking filter is not taken into account inside the coding loop in the present encoder side architecture of VVC. The deblocking filter is crucial in enhancing perceived quality for video compression, however. The MSSSIM-based assessment will unquestionably benefit from the boundary effects reduction since less blocking artefacts would smooth the textures and the signal variance may be decreased. When encoding each CU in this situation, it is preferable to use the block artefacts elimination method. As a result of the deblocking filter being incorporated into the RDO process for each CU, our proposed design calls for the CU border to be filtered in the coding loop following CU reconstruction. Block artefacts were reduced, and the MS-SSIM score was improved as a result of this technique.

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

3. EXPERIMENTS

This section provides the experimental findings for the suggested perceptual quality-oriented video coding system. Particularly, we conduct the tests within the guidelines established by the CLIC-2021 challenge and submit our findings.

3.1. Training Details

We construct the network using 10 residual units, and the number of feature maps is set to 64 in order to assess the effectiveness of the post processing module. Pytorch is used to implement the network creation and training procedure. We use the 462 video sequences that have been given as is, and we use VTM-12.0 [1] with RA setup to compress them. Except for the quantization parameter (QP), which is set to 37 during compression, all other critical parameters are left at their default values. As a result, compressed frame and ground truth training samples are generated pairwise. MS-SSIM is the loss function we'll be using to tweak the network's settings in this post. Adam is employed as an optimization approach during the training phase. The learning rate begins at 1e-4 and gradually decreases to 1e-5.

3.2. R-D Performance on Validation Set

3.2.1 Validation Dataset Description

562 films (314,175 frames) from the UGC dataset were used to create the validation dataset [8]. 13 categories—including animation, cover songs, gaming, how-to, lectures, live music, lyric videos, music videos, news clips, sports, television clips, vlogs, and virtual reality—could be used to

eISSN 1303-5150

categorise the videos. As shown in Fig. 2, films in various categories display a variety of aesthetic traits. For instance, whereas Vlog videos may have more intricate textures than animation videos, the backdrops of cover songs and lectures tend to remain static while those in sports and games often use motion estimation. Taking into account these various traits, the sequence level bit allocation technique significantly improves on the validation set.

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

3.2.2 Performance

The suggested method for compressing videos outperforms VTM-12.0's MS-SSIM. As CLIC-2021 restricts the data capacity of coded files, we use the total amount of space that is used to describe the bitrate.

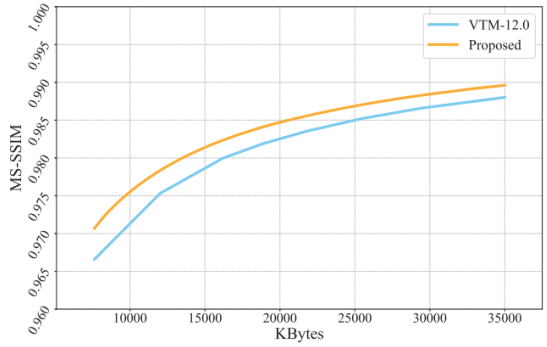

Figure 3. The comparison between the proposed algorithm and VTM-12.0

Additionally, the distortion is measured using the assessment metric MS-SSIM. In Fig. 3, the R-D performance curve is shown. At every bitrate, the suggested video compression method performs better than VTM-12.0. It should be noted that the suggested framework may provide various bit-rate points. In our proposal, Rd is specifically 9608.5 KBytes in size. Rc and the associated MS-SSIM are 24108.58 KBytes and 0.9865, respectively, according to Eqn. 4. Despite the fact that our decoder incorporates a channel attention network, the leaderboard1 indicates that it only took 1688 seconds to decode the validation dataset.

4. CONCLUSION

A unique MS-SSIM-focused video compression framework is presented in this research by combining the state-of-the-art VVC coding standard with multi-scale optimization coding tools. The suggested codec regularly improves coding effectiveness while maintaining MS-SSIM quality. In this context, three significant advanced strategies are presented, ranging from CU to sequence level. Each CU's RDO procedure carefully considered the elimination of block artefacts. We subsequently presented a channel attention-based post-processing network to

eISSN 1303-5150

improve the subjective quality. Finally, dynamic programming was used to predict the best bit allocation 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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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, 201