ISSN No:-2456-2165

A Survey on Video Coding Optimizations using Machine Learning

Mahesh Pawaskar
School of Engineering,
Career Point University, Kota
Rajasthan, India

Dr. Gaurav Vijay
School of Engineering
Career Point University, Kota
Rajasthan, India

Abstract:- The most common type of data used globally is presently video data. The volume of video data has been rising explosively around the globe as a result of the quick development of video applications and the rising demand for higher-quality video services, giving the biggest challenge to multimedia processing, transmission, and storage. Video coding by compression has become somewhat saturated while the compression ratio has grown in the last three decades. Deep Learning algorithms offer new possibilities for improving video coding technologies since they can make data-driven predictions and learn from vast amounts of unstructured data. We explore machine learning-based video encoding optimization in this research, which lays a solid groundwork for further advancements in video coding. The video service's designer must choose a suitable video coding scheme to satisfy criteria like efficiency, complexity, rate distortion, flexibility, etc. This article also presents challenges associated with machine learning video coding optimization. The survey is mainly presented from two key aspects, first is low complexity optimization with the help of advanced learning tools, such as feed-forward CNN, deep RL, and deep NN, and second is learning-based visual quality assessment (VQA).

Keywords:- *Video Coding, Deep Learning, Machine Learning, High-Efficiency Video Coding Standard (HEVC), Versatile Video Coding (VVC), Visual Quality Assessment. (VQA).*

I. INTRODUCTION

Numerous video applications, including TV broadcasting, movies, video-on-demand, video conferences, mobile video, video surveillance, remote control, robotics, 3D videos, and free viewpoint TV, have emerged with the development of multimedia computing, communication, and display technologies. Numerous aspects of daily life, including industry, communication, national security, the military, education, medicine, and entertainment, have made extensive use of these video applications. The majority of data transmission over the internet today is video data, and its volume is increasing dramatically yearly. YouTube is extensively used to share information through video.

Digital video has many advantages over traditional analog video, which has led to its replacement. Text and audio data are compatible with digital videos. To properly store and transmit visual information, an effective video coding system is required. Digital videos consume large amounts of data, and if they are not compressed properly, it would be highly difficult to store and transmit video data. Although today, data storage capacity, network bandwidth, and computer power have increased tremendously, demands for better-quality video have never stopped.

YouTube is a video-sharing website, that enables to watch online videos. YouTube received 112.9B visits in the month of October 2023, with an average session duration of 35:09 which increased the traffic by 19.00% within two months. [1]

One of the key technologies in video applications is video coding, which makes it possible to compress and organize video data more efficiently for computing, transmission, and storage. In order to improve video compression efficiency, machine learning is the most advanced research topic to be explored. The study of machine learning allows for the analysis of data to find hidden patterns and drive decisions. Owing to its exceptional ability to learn from data, a number of recent studies have significantly enhanced video coding results by including machine learning algorithms in the process.

The idea of perceptual redundancy is explained by the fact that not all video distortions are perceptually visible by the Human Visual System (HVS), which is ultimately responsible for perceiving most videos. The eyes and the brain are the two functioning components of the HVS. Numerous visual characteristics and redundancies have been identified and inspired by HVS research that is based on physiological (eye) and psychological (brain) studies [2]. The notion of Just Noticeable Difference (JND) arises when multiple pixel values in an image exhibit a very fine-scale variation. In most cases, the distortion is unnoticeable. The eyes are responsible for these physiological perceptual redundancies. Additionally, the perceptual sensitivity differs depending on the video's subject matter, the viewer's consciousness, and their region of interest, or ROI, which corresponds to how the brain processes psychology. The goal of video coding is to preserve visual quality while utilizing signal and perceptual redundancy as much as possible.

ISSN No:-2456-2165

II. EXISTING STANDARDS

In this paper, we overview key challenging issues in video coding and recent advances in machine learning-based video coding optimization. The Motion Picture Expert Group (MPEG) of ISO/IEC and Video Coding Expert Group (VCEG) from ITU-T, play a vital role in standardizing video coding and advances in coding technologies. There are five leading standards from different generations, that are popular in video coding standards. These standards are H.261, MPEG-4, H.264/AVC, H.265/HEVC, and VVC. [3] [4]

H.261 was an early video coding standard developed for video conferencing applications. H.261 used basic compression techniques such as motion compensation and discrete cosine transform (DCT) for video compression. While these techniques are less advanced compared to modern video codecs, they were innovative at the time. It laid the foundation for subsequent video coding standards and contributed to the evolution of video compression technology. [5]

MPEG-4 is part of the MPEG suite of standards and was introduced as a successor to earlier video coding standards like MPEG-2. It was first published in 1998 and has gone through several revisions and extensions over the years. It uses advanced video compression techniques, including motion compensation, transform coding such as DCT, and quantization to reduce data size while maintaining video quality. [6]

H.264 has had a profound impact on the multimedia industry, and it is one of the most widely adopted video compression standards. Digital Multimedia Broadcasting (DMB), Digital Video Broadcasting-Handheld (DVB-H), and iPod are just a few of the video coding apps that have embraced and grown to love this standard. It offers several feature sets for coding algorithms that have been found to satisfy particular application requirements. The market share of the leading online video codecs and containers has been studied recently, and it shows that in 2018, 82% of online video streams were encoded with AVC/H.264. [7]

High Efficiency Video Coding (HEVC), which is also referred to as H.265 is the latest video coding standard. The HEVC standard is advanced and standardized collaboratively by the International Telecommunication Union-T Video Quality Experts Group (VQEG) and ISO/IEC MPEG organizations. With certain modifications from previous standards, block-based motion-compensated hybrid video coding techniques serve as the foundation for the architecture of the video coding layer [8]. A few market players incorporated HEVC into their product lines. However, this coder's market share is not particularly high and already seems to be saturated. Business uncertainty resulted from the unexpected complexity and delay in realizing the entire cost of licensing for implementation. According to recent research on the market share of top online video codecs and containers, in the year 2018, only 12 % of video streams on the internet were coded with HEVC [7].

In July 2020, Joint Video Exploration Team (JVET) finalized the Versatile Video Coding (VVC) standard. It increases compression capabilities by adding new tools, reaching up to approximately 50%-bit rate reduction for equivalent video quality when compared with HEVC. It is useful for emerging applications such as HDR/WCG video, 360^0 immersive video, screen content coding etc. [9]

The main objective of video coding standards is to minimize the bit rate without significantly damaging visual quality. While achieving bit rate and visual quality, there is a challenge of maintaining low complexity. Now, researchers are focusing on learning-based approaches to upgrade video coding performance. There are a number of different areas to use learning-based approaches. In this article, two areas are considered for learning-based approaches, which are low-complexity coding optimization, and high-quality coding optimization. Subsequent sections discuss existing learning-based approaches.

III. LEARNING BASED LOW COMPLEXITY CODING OPTIMIZATION

In predictive coding, refined variable block size partitioning can improve prediction accuracy, which in turn can lower coding residue and improve coding efficiency. Variable block size partition and number of prediction modes may increase the complexity of the coding. There are different numbers of intra-prediction modes. For example, there are 1, 4/9, 35, and 67 prediction modes for MPEG-2, H.264/AVC, H.265, and VVC respectively. Refined modes, Transform Unit (TU), Motion Estimation (ME), and loop filtering techniques are incorporated in inter-prediction mode to increase coding efficiency. The selection of a more appropriate mode in minimum time is desirable to speed up video coding. Additionally, there are multiple decision layers that are to be explored in a recursive way. A number of approaches are adopted for mode decision. These approaches have very limited complexity and are fast. But there are some drawbacks which are 1) very few features are exploited which restricts the discriminability for distinguishing each mode. 2) Due to limited statistical analyses, thresholding of these algorithms may not be optimized.

Learning of mode decisions can be made using classification problems. In this paper, research on machine learning-based mode decision approaches is explored. In this section, four different machine learning-based approaches are discussed.

Decision tree, binary classifier, including support vector machine (SVM), Back Propagation Nural Network (BPNN) unsupervised learning are different machine learning models that are applied to skip some modes or to select the best mode among seven different modes of H.264/AVC.

Eduardo Martinez-Enriquez et al. proposed two-level classification-based approach for inter-mode decisions in AVC. The first classifier determines whether to skip the mode or DIRECT the mode. Whereas the second determines whether to use small modes such as 8×8, 8×4, 4×8, and 4×4

or large modes such as 16×16, 16×8, 8×16, and 8×8. The experimental result showed that, compared to others, this method saves 60% of the total encoding time. Obviously, there is a small amount of compromise with the rate-distortion parameter. [10].

Yu-Huan et al. proposed a multi-phase nearest mean classification based on RD cost clustering for fast mode decisions. It is an unsupervised clustering machine learning model. This method achieved a 68% reduction of time at the expense of a slight increase in bit rate. [11]

Jui-Chiu Chiang et al proposed a fast stereo video encoding algorithm. This algorithm is based on hierarchical two-stage neural classification, including fast prediction source determination and fast block partition selection. [12]

Paula Carrillo et al. suggested a machine learning based approach. In this technique, they have proposed three-level topology for inter-mode decision. The first level improves speed by SKIP early decision. In the second level, there is a direct division between inter 8×8 and sub modes against inter 16×16 and sub modes. If other leaf is selected, third classification between inter 16×16 sub modes and intra 4×4 is evaluated. [13]

H.265/HEVC has a large number of decision modes. That makes the task more challenging. H.265/HEVC has more complex decision computation as compared to H.264/AVC. H.265/HEVC includes recursive quad-tree CU mode decision, multi-class CU and TU mode decision. In the following sections, machine learning based HEVC INTRA and INTER coding optimization are discussed. In recent years Deep Nural Network (NN) has been widely used in visual signal processing. Researchers are putting their efforts into exploring end-to-end deep learning-based decision schemes.

Zhenyu Liu et al. used Convolution Neural Network (CNN) to analyse the texture of images and reduces the number of CU mode. Whatever CU modes are available are undergone through an exhaustive Rate-Distortion-Optimization process. In this encoding, CNN determines the texture of CU and then identifies the optimal CU/PU configuration. They have incorporated quantization parameters in CNN architecture. This method could save 63% intra-coding time with the cost of a 2.66 % BDBR increase [14].

Thorsten Laude et al. developed deep learning intra prediction mode decision process for H.265/HEVC. Input values of block samples to be coded are fed through a deep convolution neural network. Without RD optimization of all feasible modes, the choice of intra-prediction mode is expressed as a classification problem. [15]

Tianyi Li et al. proposed a complexity reduction approach for INTRA mode. This model learns Deep Convolution Neural Network to predict CTU partition instead of RDO. They established a large-scale database with diversiform patterns of CTU partition. Then, they created a model by partitioning as a three-level classification problem. In order to solve classification problem, they used various sizes of convolution kernels and trainable parameters. Experimental result showed that, their approach reduced intra encoding time with negligible Bjontegaard delta bit-rate [16].

As brute-force searches for rate-distortion optimization, the quad-tree partition of the Coding Unit (CU) is responsible for complexity in encoding. Mai Xu et al. collected large-scale database which includes CU partition data for intra and inter-modes of HEVC. This is used for deep learning CU partition. They used a hierarchical CU partition map (HCPM) to depict the CU partition of a whole coding tree unit. Next, they suggested, using an early terminated hierarchical CNN (ETH-CNN) to develop prediction skills for the HCPM. Consequently, by using ETH-CNN to determine the CU partition instead of a brute-force search, the encoding difficulty of intra-mode HEVC can be significantly decreased. Third, to discover the CU partition's temporal correlation, an ETH-LSTM is suggested. A combination of ETH-LSTM and the ETH-CNN is used to predict the CU partition, which reduces HEVC complexity in inter-mode. Experimental results showed that, their method outperformed other state-of-the-art approaches in terms of complexity [17].

In order to reduce complexity in H.264 to HEVC, Jingyao Xu et al. suggested deep learning based approach to replace brute-force searching for rate-distortion optimization. They built large-scale transcoding database. After that, determined correlation between HEVC CTU partition and H.264 features. These relation helps to find out temporal and spatial-temporal similarities of the CTU partition. Next, they proposed hierarchical long short-term memory (H-LSTM) architecture network. This deep learning-based architecture predict the CTU partition of HEVC. The performance of (H-LSTM) is compared with other methods. [18]

- *Discussion on learning based low complexity coding optimization:*

Low complexity optimization becomes more important when the coding complexity grows exponentially. In the meantime, the complexity of each mode decision problem in VVC increases. In order to solve complicated decision problems, advanced learning tools, such as feed-forward CNN, deep RL, and deep NN, are better options.

IV. LEARNING-BASED VISUAL QUALITY ASSESSMENT (VQA)

Minimizing distortion (D) or increasing quality (Q) is the aim of video coding. The quality Q is determined by PSNR, based on the pixel-by-pixel difference between the original and reconstructed pictures, and the distortion D is determined by MSE. But, there is no guarantee that PSNR and MSE reflect the real perceived quality of HVS. There are a number of Visual Quality Assessment (VQA) metrics that have been developed, such as SSIM, FSIM, Multi-Scale SSIM etc. Creating a useful visual quality metric that is in line with human perception is difficult. Through the extraction of visual elements from data and the development

of data-driven solutions, machine learning opens up new possibilities. Visual Quality Assessment (VQA) matrix can be categorized as No Reference (NR), Reduced Reference (RR), and Full Reference (FR). In the FR matrix full reference frame is available, RR needs partial side information and NR doesn't have any reference frame or information. Recently, there have been attempts to apply deep learning to VQA by researchers due to the progress of deep learning in image recognition.

Le kang et al. developed a Convolution Neural Network for No-Reference image quality assessment. As a complete optimization process, they combined feature learning and regression. This enabled to employ of new training techniques to boost performance. The suggested approach delivers state-of-the-art performance on common Image Quality Assessment (IQA) datasets and produces predictions of image quality that are highly correlated with human perception. They demonstrated that the proposed method could estimate quality in local regions [19].

Sebastian et al presented a work, that developed a deep neural network-based approach for Image Quality Assessment (IQA). The proposed network comprised 10 convolution layers, 5 pooling layers for feature extraction, and two fully connected layers for regression. This model is useful for No Reference as well as Full Reference images. This model allows for joint learning of local quality and local weights. The proposed model was evaluated on LIVE, CISQ and TID2013 database as well as the wild image quality database. This model shows superior performance to state-of-art NR and FR IQA. [20]

Chunling Fan et al. developed a multi-expert Convolutional Neural Networks (CNNs) based NR IQA algorithm. This network consisted of distortion type classification, CNN based IQA algorithms, and fusion algorithm. To determine the type of distortion, present in the input image, they first introduce a distortion-type classifier. Then, they provide an IQA method based on multi-expert CNN for every kind of distortion. In the end, a fusion technique combines the multi-expert CNN-based image quality forecasts and the distortion kinds' classification results. Model was assessed with LIVE II database and cross-dataset on CSIQ database. The proposed algorithm shows improvement for NR IQA. [21].

Hak Gu Kim et al. proposed a deep learning-based virtual reality image quality assessment method. The proposed deep network consists of a virtual reality (VR) quality score predictor and human perception guider. By encoding the positional feature and visual feature of a patch on the omnidirectional image, the proposed VR quality score predictor learns the positional and visual properties of the image. Patch weight and patch quality score are estimated using the encoded positional feature and visual feature. The image quality score is then anticipated by adding together all of the patch scores and weights. Using adversarial learning, the suggested human perception guide assesses the projected quality score by comparing it to the human subjective score. The experimental results demonstrate that, for omnidirectional images, the proposed method outperforms both the state-of-the-art and the current two-dimensional image quality models. [22]

The fact that deep learning-based methods need a huge volume of labelled data for training is one of the difficult problems. If the training dataset is insufficient or does not accurately replicate real-world movies, the learning model can have trouble handling a variety of contents and distortions. Very limited labelled images are available in quality assessment.

- *Discussion on Learning-based Visual Quality Assessment:*

To use the VQA algorithm as the quality objective, the block-based VQA algorithm adaptation is desirable compared to the image- or video-based algorithms. Another point is that while rate-distortion theory was first developed using the MSE, it should also be reevaluated in terms of adaptation. Creating a mathematical relationship between VQA and MSE before implementing it in video coding is one way to solve this issue. Compute complexity is another difficult problem. The computational complexity rises dramatically when sophisticated feature extraction methods and trained classifiers are used in quality prediction. The coding algorithm will become exceedingly complex due to the high frequency of invoking of learning-based VQA algorithms, particularly the deep learning-based schemes, in the RDO. It is worthwhile to investigate how to include the VQA—particularly the learning-based VQA with superior performance—into video coding with manageable complexity.

V. CONCLUSION

We reviewed the development of several video coding standards in this research. VVC is an advanced video coding standard that has been shown to compress video more effectively. The great accuracy used by the HEVC and VVC algorithms is a result of the progress of more computer power. This article also presents challenges associated with machine learning video coding optimization. The survey is mainly presented from two key aspects, first is low complexity optimization with the help of advanced learning tools, such as feed-forward CNN, deep RL, and deep NN, and second is learning-based visual quality assessment (VQA). In each case, the problem formulation, advantages, and challenge issues are presented. To sum up, learning-based coding optimizations have a lot of benefits and promise, and the academic and industrial groups will find this to be a promising future.

There are other factors than coding efficiency that influence the industry's choice of video coding technology for goods and services. Appropriate licensing conditions are crucial when selecting video coding options.

REFERENCES

[1]. https://www.semrush.com/website/youtube.com/overview/
[2]. Zhang, Yun, Sam Kwong, and Shiqi Wang. "Machine learning based video coding optimizations: A survey." Information Sciences 506 (2020): 395-423.
[3]. https://mpeg.chiariglione.org/who-we-are
[4]. Richardson, Iain E. The H. 264 advanced video compression standard. John Wiley & Sons, 2011.
[5]. Akramullah, Shahriar. Digital video concepts, methods, and metrics: quality, compression, performance, and power trade-off analysis. Springer Nature, 2014.
[6]. T. Sikora, "The MPEG-4 video standard verification model," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 7, no. 1, pp. 19-31, Feb. 1997, doi: 10.1109/76.554415.
[7]. "Market share of top online video codecs and containers worldwide from 2016 to 2018," Statista,New York, 2019. [Online]. Available: https://www.statista.com/statistics/710673/worldwide-
[8]. G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.
[9]. B. Bross et al., "Overview of the Versatile Video Coding (VVC) Standard and its Applications," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3736-3764, Oct. 2021, doi: 10.1109/TCSVT.2021.3101953.
[10]. E. Martinez-Enriquez, A. Jimenez-Moreno, M. Angel-Pellon and F. Diaz-de-Maria, "A Two-Level Classification-Based Approach to Inter Mode Decision in H.264/AVC," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 11, pp. 1719-1732, Nov. 2011, doi: 10.1109/TCSVT.2011.2134010.
[11]. Y. -H. Sung and J. -C. Wang, "Fast Mode Decision for H.264/AVC Based on Rate-Distortion Clustering," in IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 693-702, June 2012, doi: 10.1109/TMM.2012.2186793.
[12]. J. -C. Chiang, W. -C. Chen, L. -M. Liu, K. -F. Hsu and W. -N. Lie, "A Fast H.264/AVC-Based Stereo Video Encoding Algorithm Based on Hierarchical Two-Stage Neural Classification," in IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 2, pp. 309-320, April 2011, doi: 10.1109/JSTSP.2010.2066956.
[13]. P. Carrillo, Tao Pin and H. Kalva, "Low complexity H.264 video encoder design using machine learning techniques," 2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2010, pp. 461-462, doi: 10.1109/ICCE.2010.5418749.
[14]. Z. Liu, X. Yu, Y. Gao, S. Chen, X. Ji and D. Wang, "CU Partition Mode Decision for HEVC Hardwired Intra Encoder Using Convolution Neural Network," in IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5088-5103, Nov. 2016, doi: 10.1109/TIP.2016.2601264.
[15]. T. Laude and J. Ostermann, "Deep learning-based intra prediction mode decision for HEVC," 2016 Picture Coding Symposium (PCS), Nuremberg, Germany, 2016, pp. 1-5, doi: 10.1109/PCS.2016.7906399.
[16]. T. Li, M. Xu and X. Deng, "A deep convolutional neural network approach for complexity reduction on intra-mode HEVC," 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 2017, pp. 1255-1260, doi: 10.1109/ICME.2017.8019316.
[17]. M. Xu, T. Li, Z. Wang, X. Deng, R. Yang and Z. Guan, "Reducing Complexity of HEVC: A Deep Learning Approach," in IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 5044-5059, Oct. 2018, doi: 10.1109/TIP.2018.2847035.
[18]. J. Xu, M. Xu, Y. Wei, Z. Wang and Z. Guan, "Fast H.264 to HEVC Transcoding: A Deep Learning Method," in IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1633-1645, July 2019, doi: 10.1109/TMM.2018.2885921.
[19]. L. Kang, P. Ye, Y. Li and D. Doermann, "Convolutional Neural Networks for No-Reference Image Quality Assessment," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1733-1740, doi: 10.1109/CVPR.2014.224.
[20]. S. Bosse, D. Maniry, K. -R. Müller, T. Wiegand and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," in IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 206-219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.
[21]. C. Fan, Y. Zhang, L. Feng and Q. Jiang, "No Reference Image Quality Assessment based on Multi-Expert Convolutional Neural Networks," in IEEE Access, vol. 6, pp. 8934-8943, 2018, doi: 10.1109/ACCESS.2018.2802498.
[22]. H. G. Kim, H. -T. Lim and Y. M. Ro, "Deep Virtual Reality Image Quality Assessment With Human Perception Guider for Omnidirectional Image," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 4, pp. 917-928, April 2020, doi: 10.1109/TCSVT.2019.2898732.