Regular Article

Decoder-ROI based Versatile Video Coding for Multi-Object Tracking Vision Task

Huong Bui Thanh[1,2], **Minh Do Ngoc**[1], **Xiem HoangVan**[1]

[1] VNU-University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam
[2] Hanoi University of Civil Engineering, Hanoi, Vietnam

Correspondence: Xiem HoangVan, xiemhoang@vnu.edu.vn

Communication: received 27 March 2024, revised 07 June 2024, accepted 20 June 2024
Online publication: 28 June 2024, Digital Object Identifier: 10.21553/rev-jec.371

Abstract– **Video encoding standards like High Efficiency Video Coding (HEVC) and more recently, Versatile Video Coding (VVC) have introduced significant advancements in multimedia communication applications, such as video conferencing, broadcasting, and notably, E-learning. However, recent developments in artificial intelligence (AI) and big data have given rise to an urgent need for a specialized video encoding model designed for image and video analysis applications, namely video coding for machines (VCM). In this context, we propose a novel video encoding approach that effectively combines the Region of Interest (ROI) coding algorithm with the VVC encoding model. The proposed coding solution identifies ROI within video frames through deep learning models. Consequently, we propose an adaptive compression method for each frame block, ensuring both the execution performance of machine learning applications and the minimal data encoding requirements. In addition, to achieve new coding scheme without adding bitrate, new feature extraction approach is utilized using only decoded information (Decoder-ROI). The results demonstrate that the Decoder-ROI based VVC achieved significant compression improvement when compared to the standard and relevant VCM schemes. Furthermore, ROI exploitation contributes to around 3.25% reduction in encoding time when compared to the baseline VVC encoding standard.**

Keywords– **Versatile video coding, ROI coding, machine vision.**

1 Introduction

Nowadays, video has become a ubiquitous medium for communication, entertainment, education, and marketing. With the proliferation of online platforms, the advent of high-speed internet and machine vision applications, the demand for video encoding has never been higher. Accordingly, researchers and organizations worldwide have made significant contributions to video coding technology. Video coding standards have been developed over the years, starting with the first generation, AVC (Advanced Video Coding) [1], followed by HEVC (High Efficiency Video Coding) [2], and the latest standard, VVC (Versatile Video Coding) [3]. Over the past three decades, many advanced coding technologies have been developed to improve compression performance for video. Particularly with the latest VVC video coding standard, research efforts are focused on enhancing encoding performance in various aspects, as demonstrated in studies such as [4–7].

Besides, A growing awareness highlights that the majority of video traffic is destined for machine vision consumption. In the contemporary landscape, societies are increasingly becoming multimedia-centric, data-driven, and highly automated. Automation, analysis, and intelligence are expanding beyond human interfaces to cater to the unique requirements of Machine-To-Machine (M2M) and Machine-To-Human (M2H) communications as shown in Figure 1. The ascent of AI-driven video intelligence solutions, exemplified by Video Coding for Machine (VCM) standards [8], tailored for M2M or M2H visual processes, will play a pivotal role in addressing the most profound challenges in multimedia computing, transmission, and storage. VCM is poised to revolutionize everyday operations in this ever-evolving landscape. Subsequently, in 2019, MPEG [9] established a group of experts dedicated to video coding research with the goal of investigating and developing the Video Coding for Machines (VCM) standard [10]. As a result, research related to VCM in the context of video coding has gained increasing attention and has extended to encompass all the current video coding standards.

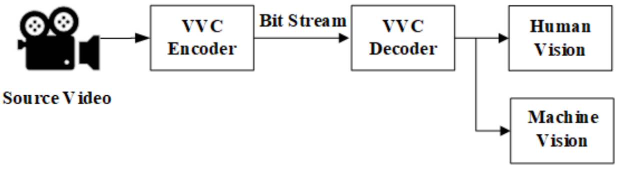


Figure 1. Standard Video Coding System for Machine Vision Applications framework.

To address the VCM problem, approaches are centered around encoding information related to significant features crucial for computer vision tasks. One such approach is Region of Interest (ROI) Coding. The term ROI Coding emerged early in the context of encoding with the JPEG2000 standard [11], focusing on encoding images so that the ROI is represented with higher quality compared to the rest of the image.

Subsequently, this term found extensive use in both image and video coding. In video coding, ROI Coding involves encoding the region of interest with higher quality compared to other areas within a frame [12]. This approach has numerous applications, including online conferencing systems, video surveillance [13, 14], and supporting artificial intelligence tasks. Allocating more bits to the region of interest enhances the differentiation of human faces and behaviors from the surrounding environment, thereby improving video conferencing experiences and tracking performance.

Although there have been studies on VCM based on common video coding standards such as H.264/AVC, H.265/HEVC, and H.266/VVC, these studies have primarily focused on building encoders for VCM rather than providing a comprehensive architecture based on traditional standards and ROI Coding [12].

To address this issue, this paper focuses on presenting a model that combines ROI Coding and the VVC standard. This model addresses two main questions: (1) what information in the video should be encoded to enhance the efficiency of computer analysis and processing, and (2) how essential information for computers should be encoded. To elaborate this problem, next Section will detail the proposed decoder ROI based VVC framework in which the multi-object tracking task is considered and QP - mode map generation is specified. Afterwards, we discuss the coding performance with the proposed framework and finally, we give some conclusions and outline the future works.

2 Proposed Decoder-ROI based VVC Framework

2.1 Overall Framework

The conventional video encoding process, as typically employed in Standard-VCM, as shown in Figure 1, is not efficient for machine vision applications due to its inability to leverage crucial features within the images. In this process, information within each frame is uniformly compressed without distinguishing between different parts of the image. This leads to the loss of significant information, particularly those related to objects directly influencing computer analysis and processing. Furthermore, the traditional video encoding process does not prioritize the preservation of essential features for detection, recognition, and machine vision data processing. Instead, it focuses on reducing the video size and optimizing the Rate-Distortion (RD), which may result in the loss of important information for machine vision applications. Hence, the video encoding process requires enhancement and development to meet the demands of these applications.

For the sake of compressing extracted features from videos to serve VCM tasks, the proposed ROI coding framework, as illustrated in Figure 2, is introduced. Firstly, the reconstruction of previously coded frame, $\hat{F}_{t-1}$ is stored in the Decoded Picture Buffer (DPB). Then, to compress the current frame F_t, its Quantization Parameter (QP) map is created based on the ROI information extracted from $\hat{F}_{t-1}$, which is the frame most relevant to frame F_t in the video used for object detection tasks.

At the decoder side, the reference frame F_{t-1} is also decoded and stored in the DPB. Subsequently, the decoding of frame F_t relies on the QP adjustment information derived from the ROI of the previously reconstructed frame $\hat{F}_{t-1}$ stored in the DPB, mirroring the encoding process. In this context, the ROI determination process for frame F_t is synchronized in both transmitter and receiver sides without encoding any overhead information of QP adjustments.

2.2 Multi-Object Tracking Vision Task

In the field of artificial intelligence, there is a range of critical tasks related to real-time monitoring and localization of multiple objects. One of the most crucial tasks in this domain is Multiple Object Tracking (MOT) [15]. MOT plays a pivotal role in separating multiple objects, maintaining their identities, and generating individual trajectories for each object based on video input. The objects to be tracked can be pedestrians on the street, vehicles on the road, athletes on the field, or even groups of animals like birds, bats, ants, fish, cells, or bees. Tracking multiple objects can be seen as tracking different components of a single object.

Embarking on the journey to enhance the ability to track multiple objects, the JDE (Joint Detection and Embedding) network [16] has emerged as a critical tool. JDE combines the tasks of object detection and embedding them into feature space, enabling the tracking and maintaining of object identities in real-time scenarios. This integration allows JDE to achieve high accuracy and stability, making it a crucial tool for various real-world applications such as urban traffic management, security surveillance, and many other important tasks. Therefore, our focus will be on evaluating the impact of VCM on the pedestrian tracking task using the JDE model as depicted in Figure 3.

2.3 Decoder - ROI Extraction

ROI-based video coding is a technique where video compression is tailored to the importance of regions within the frame. This approach consists of two main components: defining the ROI and integrating the ROI into a codec framework. However, the challenge in locating ROIs lies in the consideration of time, precision, and the efficiency of computer vision tasks at the decoder side. In surveillance [17–19] or object tracking tasks, objects are often constantly in motion. Additionally, the proposed Decoder-ROI framework can operate with any algorithm that provides ROI or proposes bounding boxes for potential objects. To address this, two approaches are presented for ROI retrieval: a traditional machine learning method and a deep learning approach utilizing YOLO.

2.3.1 Traditional ROI detectors: In this approach, the proposed Decoder-ROI architecture can operate with any ROI-search algorithm. To that end, three traditional approaches have been studied, suitable for locating

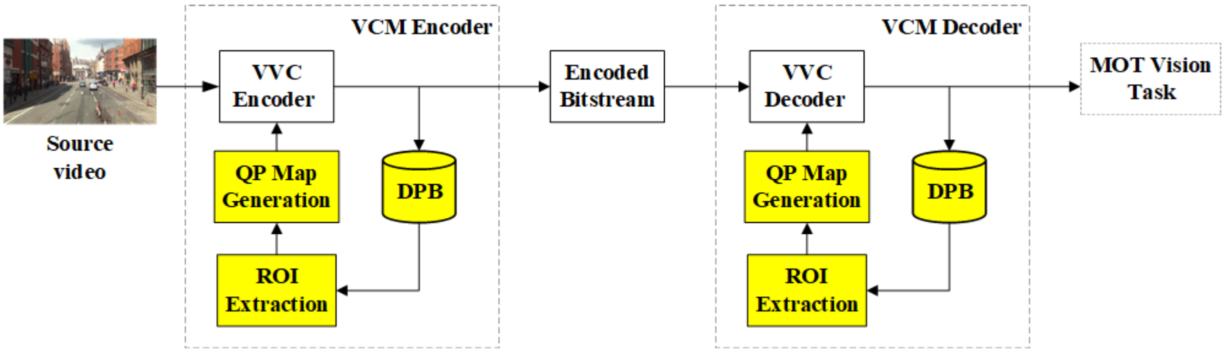


Figure 2. Decoder-ROI Multi-Object Tracking Vision Task.

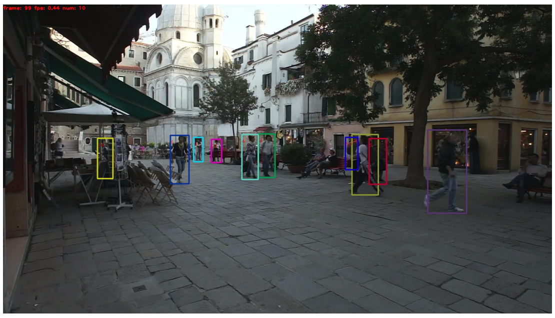
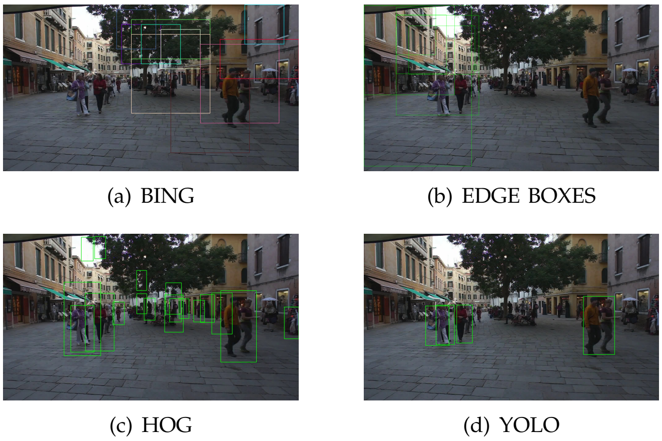


Figure 3. Object Tracking Application.



(a) BING

(b) EDGE BOXES

(c) HOG

(d) YOLO

Figure 4. The comparison of various ROI detectors aiming to achieve exemplary CTUs' ROI for the 10th frame of the MOT16-02 video.

ROIs in a VCM scenario and fast enough to be applied in an encoding process within practical scenarios. The first method, known as Edge Boxes [20], generates bounding box proposals without explicit classification, providing a user-defined number of boxes. Additionally, Binarized Normed Gradients (BING) [21] utilizes a feature representation of normalized gradients to yield bounding boxes linked to potential objects. Moreover, the fusion of Histogram of Oriented Gradients (HOG) with Support Vector Machine (SVM) contributes to the classification of significant regions in VCM. HOG extracts feature descriptors from images, subsequently integrated with SVM to efficiently classify regions. These methods, recognized for their efficiency and practical applicability, constitute the foundation of research into effective ROI encoding frameworks.

2.3.2 Deep learning ROI detectors: The second approach involves applying a well-known deep learning model, YOLO to search for ROI area in the coded picture. Since the traditional method mentioned earlier defines ROIs based on features related to human perception, such as motion, it may not be suitable for tasks involving multi-object tracking as presented in Section 2.2. In this scenario, object tracking is accomplished using the JDE model, which is a neural network structure. Therefore, YOLO, a convolutional neural network, is proposed for the purpose of locating ROIs. An example of applying the traditional machine learning methods and a deep learning approach to an image is provided in Figure 4.

In the realm of object detection algorithms, YOLOv5 [22] stands out among several approaches that have achieved remarkable advancements. In the landscape of architectural object detection, two fundamental concepts have emerged: the One-stage detector and the Two-stage detector (refer to Figure 5).

A common thread among various object detection architectures is the processing of input image features. These features undergo compression via a feature extractor, typically known as the Backbone, before being directed towards the object detector. This object detector comprises the Detection Neck and Detection Head, as depicted in Figure 5. The Neck serves as a feature aggregator responsible for amalgamating and blending the features extracted within the Backbone. Its primary role is to prepare these features for the subsequent detection step executed in the Head.

The distinctive aspect here is that the Head manages the detection process, encompassing both localization and classification tasks for each bounding box. The Two-stage detector performs these tasks separately, merging their outcomes afterward (Sparse Detection). Conversely, the One-stage detector executes these tasks simultaneously (Dense Detection), as illustrated in Figure 5. YOLO, as a one-stage detector, embodies the principle of "You Only Look Once."

2.4 QP-Mode Map Generation

To focus the encoding on the information relevant to the object tracking task, it is necessary to incorporate this information into the encoder. Specifically, as detailed in section 2.3, we can identify the ROI area in

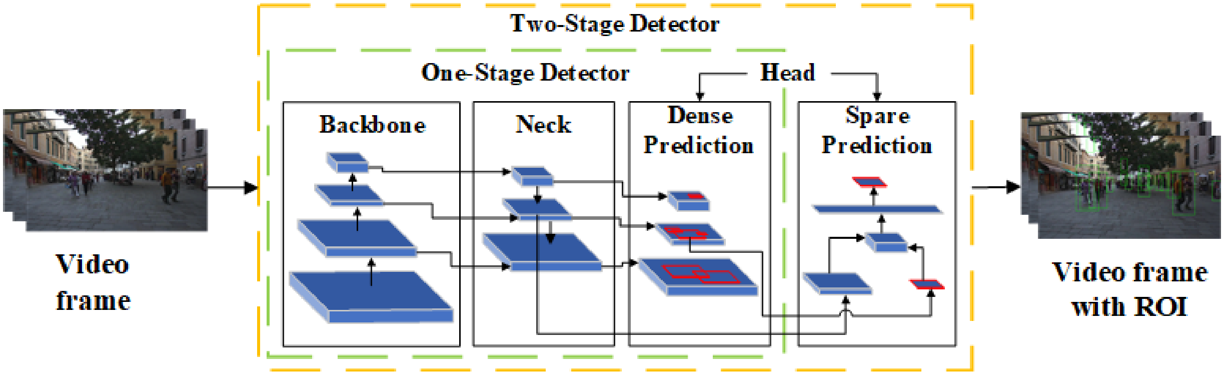


Figure 5. YOLOv5-based Region of Interest (ROI) detection architecture.

the coding picture and stored it as pixel coordinates. Subsequently, the frames, when encoded using the VVC standard, will be divided into Coding Tree Units (CTUs) of size 128 × 128 in sequence. Therefore, we can determine the CTUs containing the ROI based on their corresponding pixel coordinates.

However, to ensure a consistent bit rate when adjusting the QP values for each Coding Tree Unit (CTU), it is necessary to determine the ratio (α) between CTUs that contain the ROI and those that do not, as described in Equation (1).

$$\alpha = \frac{\text{Number of CTUs with objects}}{\text{Number of CTUs in total}}. \tag{1}$$

Subsequently, the QP values of each CTU will be adjusted according to Equation (2) to facilitate the encoding process. Specifically, for CTUs within the ROI, the QP value decreases by an amount $\alpha \times \delta_{QP}$, whereas for CTUs outside the ROI, the QP value increases correspondingly. This adjustment ensures consistent image quality throughout the encoding process

$$QP = QP \pm \alpha \times \Delta\text{QP}, \tag{2}$$

where QP represents the quality of each CTU, and ΔQP is the difference in QP values between regions with objects and regions without objects.

Furthermore, the complexity of algorithms for multi-object tracking and ROI determination, alongside the H266/VVC encoding standard, is high. In fact, the VVC encoding complexity is reported to be 30 times higher than HEVC [23]. Therefore, this research proposes a method to reduce the complexity of VVC encoding by adjusting the number of intra prediction modes for each region. Accordingly, each coding unit (CU) determines the number of intra prediction modes based on whether it contains an object or not, as illustrated in the Figure 6.

3 Performance Evaluation

3.1 Experiment Settings

To assess the efficiency of proposed Decoder ROI - VCM solution, the MOT16 Benchmark dataset [24] is ultilized due to its popular use in multi-object tracking task. This dataset comprises six sequences, including both forward-facing footage captured by mobile cameras and top-down surveillance videos. The characteristics of these sequences are particularly suitable for our method due to the minimal abrupt changes between frames, as described in Table I and Figure 7. Additionally, the MOT16 dataset is a standard benchmark widely used for evaluating models in multi-object tracking tasks, providing a reliable basis for comparison and validation. Afterwards, the videos are encoded using the VVenc software [25] and QP_{base} values ues of 22, 27, 30 and 37 in all intra configuration. Subsequently, the decoded videos were fed into JDE-1088x608 [16] obtain the multiple object tracking and measure its performance with the MOTA (Multiple Object Tracking Accuracy) metric [26].

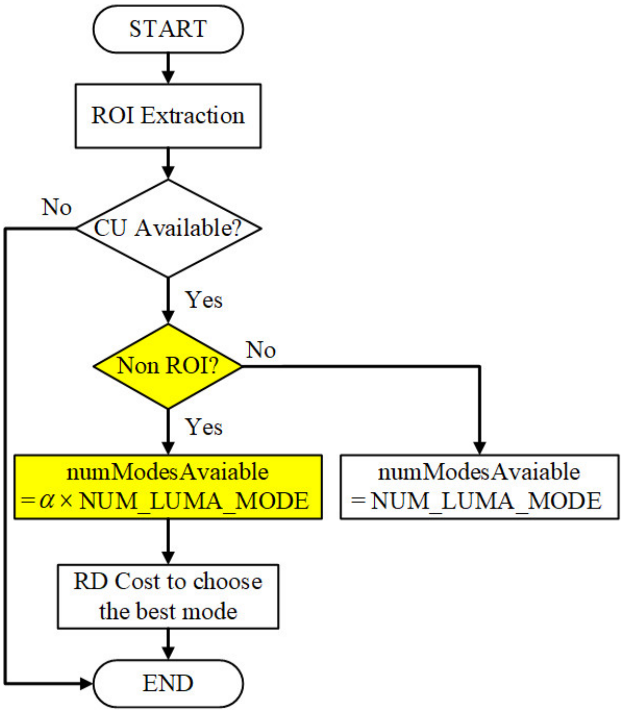

START
ROI Extraction
No
CU Available?
Yes
Non ROI?
No
Yes
numModesAvaiable =α × NUM_LUMA_MODE
numModesAvaiable =NUM_LUMA_MODE
RD Cost to choose the best mode
END

Figure 6. The flowchart of proposed method.

In order to assess the efficacy of the proposed approach, the encoding duration inherent to the proposed methodology is compared against that of the VVC standard [3]. The temporal efficiency, denoted as Time

(a) (b) (c)

(d) (e) (f)

Figure 7. First frame of testing video sequences.

Table I
The Time (Seconds) to Detect ROI by Various Algorithms on the MOT16-02 Sequence

Method	BING	EDGE	HOG	YOLO
Time	187.79	310.83	256.01	5.40

Saving (TS), of the proposed method is mathematically expressed by Equation 3:

$$TS = \frac{T_{Proposed} - T_{VVenC}}{T_{VVenC}} \times 100\%, \qquad (3)$$

where T_{VVenC} represents the total encoding time of the VVenC [25], $T_{Proposed}$ represents the total encoding time of the proposed method. In addition, BDBR and BD-MOTA [27] are computed to evaluate the performance of the proposed method compared to VVC. BDBR shows the difference in bitrate at the equivalent quality, while BD-MOTA shows the difference in MOTA at the equivalent bitrate.

3.2 Results and Discussions

In this research, we conducted experiments using the proposed Decoder-ROI model across various ROI detection algorithms as presented in Section 2.3. Figure 8 illustrates the relationship between MOTA and rate on the MOT16-02 sequence. Notably, the YOLO-based approach achieved the best performance compared to other methods. Additionally, the illustration of applying ROI detection algorithms on a provided frame is depicted in Figure 4. In this context, the BING and EDGE BOXES algorithms identify bounding boxes containing most objects. However, these algorithms' bounding boxes lack precision in object adherence, thereby offering limited improvement in bitrate savings. Conversely, approaches utilizing HOG and YOLO features produce more accurate bounding boxes closely aligned with the objects, optimizing the bitrate allocation for encoding. While the HOG-based approaches consume more time, the ROI detection using YOLO exhibits the fastest processing time, as detailed in Table IV. Therefore, employing YOLO facilitates an optimization in both bitrate savings and encoding time.

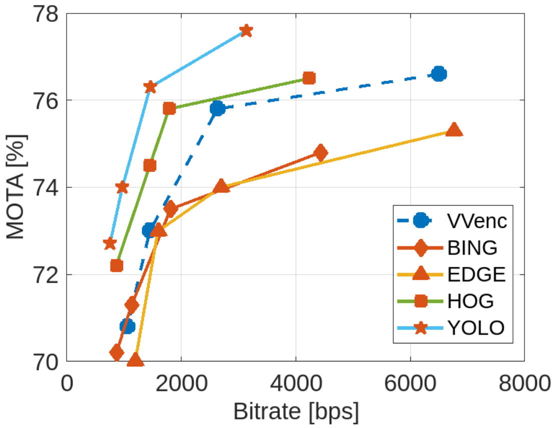


Figure 8. MOTA performance according to rate on MOT16-02 sequence.

Table II
BD-Rate Values in % with the MOTA as Quality Metric for ROI Detection Algorithms

Seq	BING	EDGE	HOG	YOLO
MOT16-02	17.43	34.8	-9.55	-43.95
MOT16-04	-65.66	-71.68	-55.12	-79.12
MOT16-09	-45.05	17.54	-17.99	-57.59
MOT16-10	-68.94	-70.75	-70.62	-85.80
MOT16-11	-22.77	-2.72	-27.73	-40.58
MOT16-13	-18.83	-17.23	-16.92	-66.31
Average	**-29.12**	**-15.82**	**-29.89**	**-53.34**

Table II presents the bit savings ratio computed for different ROI detection algorithms. On average across the MOT16 dataset, the YOLO-based approach achieved the highest bit savings ratio, reaching 53.34%. This was attributed to YOLO's accurate object detection and precise bounding box generation, enabling max-

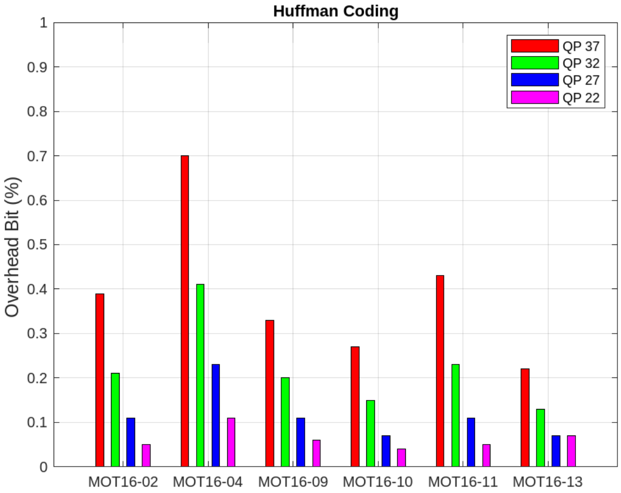


Figure 9. The percentage of overhead bits at different QPs in each video test sequence.

Table III
The Similarity of the QP Maps Generated on the Decoding Side and the QP Map Sent from the Encoding Side

SEQ	**MOT16**					
	02	**04**	**09**	**10**	**11**	**13**
Similarty (%)	99.04	94.42	97.25	97.85	94.43	98.70
Average	**96.95**					

imal bit savings while preserving crucial information for the object tracking task at the decoder end.

Additionally, the relevant approach (Encoder-ROI) relying on information from the transmitted QP map from the encoding side requires an extra bit for storing information about the non-object region, termed as overhead bit. Figure 9 illustrates the percentage of overhead bits for different QP values in each video sequence when encoded using Huffman encoding. However, in the case of the Decoder-ROI approach, the system doesn't necessitate transmitting overhead bits yet achieves effectiveness in generating the QP map. With Decoder-ROI, the generated QP map exhibits high similarity when compared to the QP map generated based on the received overhead bits from the encoding side, as presented in Table III. In Table III, the similarity is assessed based on the percentage of similarity between two QP maps of Decoder-ROI and Encoder-ROI.

Finally, the research explores the exploitation of ROI characteristics within the framework to reduce complexity by adjusting the intra prediction mode. The results are compared between the proposed method using approaches based on YOLO and the H.266/VVC video encoding standard with the reference software VVenc as described in Table IV. Accordingly, the proposed method yields an average bitrate saving of 62.25% and a 4.65 increase in BD-MOTA. Meanwhile, these YOLO-based approaches contribute to a 3.25% reduction in encoding time compared to the original H.266/VVC video encoding standard.

Table IV
Time Saving and BDBR Loss Comparison

SEQ	**TS**	**BD-Rate**	**BD-MOTA**
MOT16-02	-0.7%	-43.95	2.86
MOT16-04	-1.5%	-79.12	1.17
MOT16-09	-1.6%	-57.59	8.35
MOT16-10	-1.1%	-85.80	5.01
MOT16-11	-1.0%	-40.58	3.27
MOT16-13	-13.6%	-66.31	7.24
Average	**-3.25%**	**-62.23**	**4.65**

4 CONCLUSION

In this paper, to achieve highly efficient video compression for machines, we propose a Decoder-ROI based VVC architecture, which involves modifications to the latest VVC coding standard by integrating a ROI extraction and an adaptive rate allocation structure. The proposed Decoder-ROI VVC leverages solely on the decoded information to predict the ROI, thereby obviating the necessity for the encoder to transmit additional bitrate to achieve a commensurate ROI representation. The performance evaluation demonstrates that the proposed Decoder - ROI VVC framework achieves significant bitrate savings when employing YOLO for ROI detection. Specifically, leveraging YOLO results in a 53% bitrate reduction, with the potential for maximum savings of approximately 86% on the MOT16 test dataset. Future work could focus on extending the Decoder-ROI concept to alternative configurations such as random access (RA) and low-delay P (LDP). Moreover, extracting ROI information from frames of different perspectives could further adapt VVC to VCM tasks.

Acknowledgment

This research was funded by the University of Engineering and Technology, Vietnam National University, Hanoi, under grant number CN23.18.

References

[1] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] 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*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the Versatile Video Coding (VVC) Standard and its Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[4] X. HoangVan, L. Dao Thi Hue, and T. Nguyen Canh, "A trellis based temporal rate allocation and virtual reference frames for high efficiency video coding," *Electronics*, vol. 10, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/12/1384

[5] X. HoangVan, S. NguyenQuang, M. DinhBao, M. DoNgoc, and D. Trieu Duong, "Fast QTMT for H.266/VVC Intra Prediction using Early-Terminated Hierarchical CNN model," in *Proceedings of the 2021 International*

Conference on Advanced Technologies for Communications (ATC), 2021, pp. 195–200.

[6] X. HoangVan, S. NguyenQuang, and F. Pereira, “Versatile video coding based quality scalability with joint layer reference,” *IEEE Signal Processing Letters*, vol. 27, pp. 2079–2083, 2020.

[7] X. HoangVan, “Adaptive Quantization Parameter Estimation for HEVC Based Surveillance Scalable Video Coding,” *Electronics*, vol. 9, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/6/915

[8] Y. Zhang, M. Rafie, and S. Liu, “Use cases and requirements for video coding for machines,” *ISO/IEC JTC*, vol. 1, 2021.

[9] Y. Zhang and P. Dong, “MPEG-M49944: Report of the AhG on VCM,” *Moving Picture Experts Group (MPEG) of ISO/IEC JTC1/SC29/WG11*, Oct. 2019.

[10] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, “Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics,” *IEEE Transactions on Image Processing*, vol. 29, p. 8680–8695, jan 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.3016485

[11] D. Taubman and M. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards and Practice*. Springer Science and Business Media, 2012, vol. 642.

[12] H. Meuel, J. Schmidt, M. Munderloh, and J. Ostermann, “Region of Interest Coding for Aerial Video Sequences Using Landscape Models,” in *Advanced Video Coding for Next-Generation Multimedia Services*, Y.-S. Ho, Ed. Rijeka: IntechOpen, 2013, ch. 3. [Online]. Available: https://doi.org/10.5772/52904

[13] H. B. Thanh, S. N. Quang, T. V. Huu, and X. HoangVan, “Learning adaptive motion search for fast versatile video coding in visual surveillance systems,” *IET Image Processing*, vol. 18, no. 4, pp. 981–995, 2024.

[14] T. H. Le Dao, P. Van Giap, and H. Van Xiem, “Adaptive long-term reference selection for efficient scalable surveillance video coding,” in *Proceedings of the 2018 IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*. IEEE, 2018, pp. 69–73.

[15] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, “Multiple object tracking: A literature review,” *Artificial Intelligence*, vol. 293, p. 103448, 2021.

[16] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, “Towards real-time multi-object tracking,” in *Proceedings of the European conference on computer vision*. Springer, pp. 107–122.

[17] T. H. Le Dao, P. V. Giap, and H. V. Xiem, “Adaptive long-term reference selection for efficient scalable surveillance video coding,” in *Proceedings of the 2018 IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, 2018, pp. 69–73.

[18] T. Nguyen Thi Huong, H. Phi Cong, X. HoangVan, and T. V. Huu, “A practical high efficiency video coding solution for visual sensor network using raspberry pi platform,” in *Proceedings of the 2018 IEEE 12th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, 2018, pp. 64–68.

[19] L. T. Hue, L. Van, D. Trieu, and X. HoangVan, “Efficient and low complexity surveillance video compression using distributed scalable video coding,” *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 34, no. 1, 2018. [Online]. Available: jcsce.vnu.edu.vn/index.php/jcsce/article/view/198

[20] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *Proceedings of the Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 391–405.

[21] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, “BING: Binarized Normed Gradients for Objectness Estimation at 300fps,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3286–3293.

[22] G. Jocher, “YOLOv5 by Ultralytics,” May 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[23] F. Pakdaman, M. A. Adelimanesh, M. Gabbouj, and M. R. Hashemi, “Complexity Analysis Of Next-Generation VVC Encoding And Decoding,” in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3134–3138.

[24] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, “MOT16: A Benchmark for Multi-Object Tracking,” 2016.

[25] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, “VVenC: An Open And Optimized VVC Encoder Implementation,” in *Proceedings of the IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–2.

[26] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: the clear mot metrics,” *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[27] G. Bjontegaard, “Calculation of average psnr differences between rd-curves,” *ITU SG16 Doc. VCEG-M33*, 2001.

Bui Thanh Huong received the BSc degree in Electronics and Telecommunications, in 2000; and MSc degree in Information processing and Communication, in 2004, all from Hanoi University of Science and Technology (HUST). She is currently a main lecturer in the Department of Computer Engineering, Hanoi University of Civill Engineering, Vietnam. Her research interests are digital systems, digital signal processing and Multimedia Technology.

Minh Do Ngoc received a B.S. and M.S. degree in Computer Engineering from the VNU University of Engineering and Technology, in 2021, 2022. His research interests include AI Vision and Signal processing. Currently, he is currently member at the Faculty of Electronics and Telecommunications, VNU University of Engineering and Technology, Vietnam.

A/Prof. Hoang Van Xiem is the Head and founding member of the Department of Robotics Engineering, Vietnam National University – University of Engineering and Technology (VNU-UET). He was the former Director of the Center for Quality Assurance in VNU – UET (2021 – 2022 term). He received a Ph.D. degree from Lisbon University, Portugal, in 2015, a M.Sc. degree from Sungkyunkwan University, South Korea, in 2011, and BE degree from Hanoi University of Science and Technology, 2009, all in Electrical and Computer Engineering. He has published nearly 100 papers on image/video processing and robotics vision. He is an editor of the Frontiers in Signal Processing Journal and VNU-Journal of Science and reviewed for a number of top IEEE, Elsevier and Springer Journals. He has received a number of prestigious awards including 2023 IEEE RIVF, 2022 REV-ECIT, 2018 IWAIT and 2015 PCS best paper awards, his work on distributed video coding received the 2021 REV-AWARD, 2020 Innova Patent Silver Award (Croatia), and the Fraunhofer Portugal Challenge Award 2015. He is one of ten young research scientists (one of two in ICT) was awarded the Golden Globe in Science and Technology 2019 and in 2021 – 2022 term, he was one of the youngest associate professors appointed by the State council for professorship in Vietnam.