Artificial Intelligence Based Adaptive GOP Size Selection for Effective Wyner-Ziv Video Coding

Thao Nguyen Thi Huong, Huy Phi Cong, Tien Vu Huu
Posts and Telecommunications Institute of Technology
(thaonth,huypc,tienvh)@ptit.edu.vn

Xiem HoangVan
VNU – University of Engineering and Technology
xiemhoang@vnu.edu.vn

***Abstract*—Wyner-Ziv video coding (WZVC) has been gaining many attentions in recent decades due to its low computational complexity and error resiliency benefits, notably when compared to traditional video coding standards such as H.264/AVC or High Efficiency Video Coding (HEVC) standards. In a Wyner-Ziv video coding scheme, the compression efficiency can be controlled by the length of the group of pictures (GOP) which typically consists of the two key and several WZ frames. However, the current Wyner-Ziv video coding solutions usually employ a fixed GOP size or simple adaptive GOP size mechanisms, which depend on some heuristic features extracted from video content. To address the limitation of the current GOP size adaptation solutions, we propose in this paper a novel Artificial Intelligence based GOP size adaptation mechanism and integrate it into the most advanced transform domain Wyner-Ziv video coding (TDWZ) architecture. In the proposed GOP size adaptation mechanism, the proper GOP size is learnt from the correlation between video features and the optimal compression performance. The power of machine learning techniques is used to select the most suitable video features and the model of GOP size and compression performance correlation. Experimental results shown that, using the obtained GOP size adaptation mechanism, the TDWZ achieved a compression performance when compared to relevant benchmarks.**

***Index Terms*—Artificial Intelligence, DVC**

I. Introduction

Nowadays, there are not only traditional applications such as broadcasting and video-on-demand but also emerging applications such as wireless video networks, mobile video cameras and multi-camera surveillance systems. These emerging applications have different requirements than those related to traditional video delivery systems. However, current popular video coding solutions such as H/264/AVC, HEVC [1], [2] rely on the powerful hybrid block-based transform and inter-frame predictive video paradigm. This architecture makes high complexity encoders and light decoders. This is well-suited for traditional applications where video is encoded once and decoded several times but becomes challenge when applied for emerging applications because there is a high number of encoders but only one decoder.

In order to fulfill these new requirements, it is essential to have a different video coding paradigm with a low-power and a low complextiy encoder with expense of a high complexity decoder. The most promising solution for this case is called Distributed Video Coding (DVC). To decrease the complexity of encoder, temporal correlations are exploited at the decoder rather than encoder. Therefore, the encoder complexity is much lighter than the decoder. Information theory results [3], [4] show that despite of independent encoding and jointly decoding, DVC systems can still achieve coding efficiency similar to current hybrid video coding standards.
In DVC codec, frames are split into keyframe and Wyner-Ziv (WZ) frame. Key frames are intracoded while WZ frames are intercoded. WZ frames is usually coded by channel codes such as turbo code or low density parity check (LDPC) code [5]. However, in order to decrease the number of transmitted bits, only the parity bits and intracoded key frames are sent to the decoder. At the decoder, a prediction of the WZ frame is created and named the Side Information (SI) [6]. SI is generated by performing motion estimation and compensation using decoded key frames. This SI, together with the received parity bits, will be used to obtain the original WZ frame. For this reason, the Rate-Distortion (RD) performance of DVC codec depends on the quality of SI, consequently, depends on the distance between the key frames or the Group Of Pictures (GOP). However, a fixed GOP size along the whole sequence may be inefficient because the temporal correlation is not fully exploited when the video content changes. For frame with high motion, the temporal correlation is low and the small GOP size should be selected. Conversely, for frame with low or medium motion, the temporal correlation is high and in this case, the longer GOP size could be used.

In the literature [7]–[10], efforts are made in order to control the GOP size according to the changes in the motion activity. The more accurate the motion type of frame is identified, the better the selection of GOP size and this could significantly reduce bitrate of the system. In [7], authors used features related to histogram and block variance to evaluate the activity along the video sequence. These features can detect changes in both global and local motion. This improved the performance up to 0.4 dB for the transform domain when compared to the fixed GOP size approach. Another idea from [8] used past system behavior in order to select the GOP size. Initially, a small set of size N of different GOP sizes is created. The coding performance of the each GOP size is calculated based on the ratio of the average estimated PSNR and average coding rate. The GOP size with the highest ratio will be selected as future GOP size. Krishna R.V et al. in [9] proposed a simple GOP size control algorithm in which the blocks in a frame are classified in to key, skip, and WZ blocks. The current

frame was considered as WZ or key frames depending on the number of the skip block. Results showed that the proposed algorithm achieved quite good results with negligible encoder complexity increase.

These GOP size adaptation algorithms, however, are relatively and mainly rely in some deterministic assumptions. Consequently, RD performance of DVC codec is insignificantly improved. The objective of this paper is to precisely classify GOP size based on video content. Therefore, this paper employs a powerful artificial intelligence algorithm to efficiently select GOP size for each video segment. Since the content of video data is typically diverse, several features extracted from every five frames are adopted for artificial intelligence algorithm. The results shows that the proposed algorithm brings a major quality improvement with negligible additional complexity when compared to relevant previous solutions and can be easily integrated in the prior DVC architectures.

The rest of the paper is organized as follows. Section 2 briefly introduces the architecture of transform domain Wyner–Ziv video codec. Section 3 describes the proposed machine learning based GOP size adaptation mechanism while experimental results are discussed in Section 4. Finally, some conclusions and future works are presented in Section 5.

II. Transform domain Wyner-Ziv video codec

The proposed architecture of the transform domain Wyner-Ziv video codec is illustrated in Fig.1 in which the novelty GOP adaptation module is highlighted.

A. Encoding process

In the proposed TDWZ encoder, the input video sequence is split into subsequences of 5 frames in order to process and GOP size selection is performed for each subsequence. GOP size is chose depending on the motion content for each subsequence. If the subsequence has high motion and/or complex texture, GOP 2 is selected. On the contrary, GOP 4 is considered. After GOP size is selected, each subsequence is split into key frames and WZ frames. Key frames, corresponding to the first frame of each GOP, are conventionally encoded using HEVC intra encoder. WZ frames are encoded using DVC principle. Firstly, WZ frame is block based transformed with an integer discrete cosine transform (DCT). The obtained transformed coefficients are uniform quantized. These coefficients are organized in bands where every band contains the coefficients associated to the same frequency in different blocks. The bit representing these coefficients are split into bitplanes which go through Low-Density-Parity-Check (LDPC) encoder. The LDPC encoder computes parity bits corresponding to the encoded bitplane. While the systematic bit are eliminated, the parity bits are stored in a buffer and progressively transmitted to the decoder depending on requests sent from the decoder during the decoding process, via feedback channel.

B. Decoding process

At the decoder side, encoded key frames are decoded using HEVC intra decoder. These decoded key frames are fed into the buffer in order to create the side information, which is an error version of the original WZ frames. The difference between the original WZ frame and the corresponding SI can be considered as correlation noise in a virtual channel. This correlation noise is modeled by Laplacian distribution. An integer DCT is carried out over the generated SI in order to obtain the integer DCT coefficients, a noisy version of the WZ frame DCT coefficients. Then, the LDPC decoder corrects the error bits in the transformed SI, using the parity bits of WZ frames sent from the encoder via the feedback channel, taking into acount the correlation noise. To decide whether more parity bits are needed for the successful decoding, a convergence criteria is used. The decoded WZ DCT coefficients are then reconstructed by doing the inverse of the quantization. Finally, the inverse integer DCT transform is carried out in order to obtain entire WZ frame in the pixel domain. The decoded video sequence is created by multiplexing the decoded key frames and WZ frames.

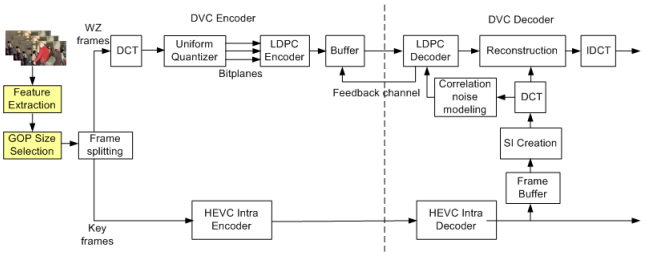

Fig. 1. Architecture of the transform domain WZ video codec

III. Artificial intelligence based GOP size adaptation mechanism

This section describes the proposed algorithm. First, features describing motion and texture of each subsequence are presented. Then, J48 decision tree based classification is detailed.

A. Features definition

As mentioned above, selected features must fully reflect the nature of video content, so some metrics are related to both global motion and local motion while others are related to the texture. The features include Sum of Absolute Difference (SAD), Difference of Histogram (DoH), Average of Motion Vectors (AMV), Number of Motion Vectors (NMV), Average Subsequence Variance (ASV), Average Subsequence Mean (ASM), DC value Variance (DCV), DC value Mean (DCM), AC value Variance (ACV) and AC value Mean (ACM). They are defined as follows.

$$SAD = \frac{1}{N-1}\sum_{k=1}^{N-1}\left(\sum_{x=1}^{H}\sum_{y=1}^{W}|f_{k+1}(x,y) - f_k(x,y)|\right) \quad (1)$$

where k, N represents the key frame index and number of key frames in a subsequence. In this paper, subsequence length equals to 5, thus $N = 3$. H, W describe the height and width

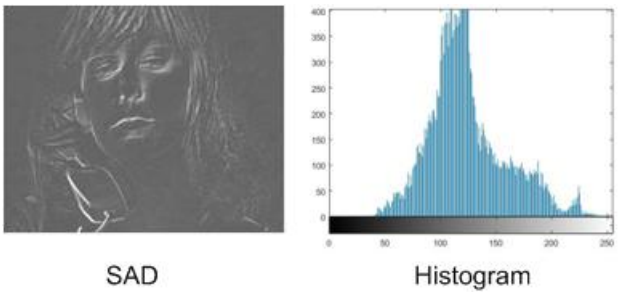

Fig. 2. SAD and Histogram feature of the first GOP in Suzie sequence

of frames. x,y and f is the coordinate and luminance value of pixel in the frame.

$$DoH = \frac{1}{N}\sum_{k=1}^{N-1}\left(\frac{1}{H.W}\sum_{i=0}^{L}|h_{k+1}(i) - h_k(i)|\right) \quad (2)$$

where h is the histogram operator with L levels.

$$AMV = \frac{1}{N-1}\sum_{k=1}^{N-1} MV(k+1,k) \quad (3)$$

where $MV(k+1,k)$ is total length of motion vector between key frames $k+1$ and k.

$$NMV = \frac{1}{N-1}\sum_{k=1}^{N-1} NMV(k+1,k) \quad (4)$$

where $NMV(k+1,k)$ is number of motion vector between key frames $k+1$ and k.

$$\text{ASV} = \frac{1}{N}\sum_{k=1}^{N}\sigma^2(k) \quad (5)$$

where $\sigma^2(k)$ is variance of pixel value in the key frame k.

$$\text{ASM} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{H.W}\sum_{x=1}^{H}\sum_{y=1}^{W} f_k(x,y)\right) \quad (6)$$

where $f_k(x,y)$ is pixel value of pixel (x,y) in the key frame k.

$$DCV = \sigma^2_{DC} \quad (7)$$

where σ^2_{DC} is variance of DC coefficient value of key frames in a subsequence.

$$DCM = \frac{1}{N}\sum_{k=1}^{N} DC(k) \quad (8)$$

where $DC(k)$ is DC coefficient value of key frames k in a subsequence.

$$ACV = \frac{1}{N}\sum_{k=1}^{N}\sigma^2_{AC}(k) \quad (9)$$

where $\sigma^2_{AC}(k)$ is variance of AC coefficient value in the key frame k.

$$ACM = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{H.W-1} AC_i(k) \quad (10)$$

where $AC_i(k)$ is AC coefficient i^{th} value in the key frame k.

B. Training and classification

Classification is the process of building a model of classes from a set of records that contain class labels. A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The performance comparison of Decision Tree Algorithms and Artificial Neural Network, and Nave Bayes Classifier on a set of attributes was performed. On the basis of results it has been examined that Decision Tree Algorithms performs better than the Artificial Neural Network and Nave Bayes Classifier. So, J48 decision tree method is chosen as the optimal for the problem as it has shown better results than the algorithms. The J48 decision tree method is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team [12].

1) J48 model training: The J48 model must be offline trained and only once before used for classification stage. First, features mentioned above are extracted from 352 subsequences of the five sequences *Foreman, Hall Monitor, News, Husky and Mobile*. Together with these features, the class, GOP 2 or GOP4, created by comparing Bjntegaard-Delta Peak Signal to Noise Ratio (BD-PSNR) in order to choose the size of GOP, are used to train J48 model.

2) Testing feature extraction: For each input video sequence, every five frames are considered as a subsequence and the features proposed above are extracted from each subsequence.

3) J48 classification: The classification is performed for a set of extracted features with J48 trained model. The output of the classification is the GOP size for each subsequence including five frames.

IV. Experimental results

A. Test conditions

In order to evaluate the proposed algorithm, BD-PSNR metric is used for comparision. BD-PSNR metric described in [11] to provide relative gain between two methods, by measuring average difference between the two RD-curves with a RD curve is chosen as base curve. If BD-PSNR is positive, it means that the second curve is better than the base curve. In this assessment, RD curves of GOP4 and the proposed method named Adaptive GOP are compared with the base curve GOP2. In this experiment, four video sequences are used for assessment including *Coastguard, Suzie,Pamphlet and Harbour* with the characteristics summarized in Table 1 while the first frames of four sequences are shown in Fig.3.

Fig. 3. The first frame of video test sequences

TABLE I
CHARACTERISTICS OF TEST SEQUENCES

Test sequences	Spatial resolution	Number of frames	Quantization parameters
Coastguard	176x144	300	{26,30,34,38}
Suzie		150	{25,29,34,40}
Pamphlet		150	{25,29,34,40}
Harbour		150	{25,29,34,40}

B. Performance evaluation

RD performance results for four test video sequences are presented in Table II and III.

As shown in Table II, PSNR values of the proposed method are better than the values of GOP 4 and approximated to the values of GOP2. Bitrate values of the proposed method are higher than the values of GOP4 and lower than the values of GOP2. Thus, the selection between GOP2 and GOP4 depends on the trade-off between PSNR and Bitrate. The results show that the reduction quality of video (in term of PSNR value) in the proposed method is negligible while the Bitrate saving is rather high. Table III shows that the Bitrate saving of proposed method is 3.37% and 9.62% compared to GOP2 and GOP4, respectively.

TABLE II
RD PERFORMANCE FOR TEST SEQUENCES

Sequence	QP	GOP2		GOP4		Adaptive GOP	
		Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR
Coastguard	26	27760	38.18	28242	34.65	27735	38.14
	30	17131	34.87	16140	32.48	17058	34.84
	34	9838	31.88	8228	30.36	9760	31.85
	38	5256	29.14	3781	28.23	5199	29.12
Average		**14996.25**	**33.52**	**14097.75**	**31.43**	**14938**	**33.49**
Suzie	26	18424	41.58	19719	41.26	18565	41.34
	30	10869	38.56	11172	38.23	10530	38.26
	34	5725	35.41	5588	35.15	5283	35.29
	38	2667	32.24	2353	32.04	2270	32.19
Average		**9421.25**	**36.95**	**9708.00**	**36.67**	**9162.00**	**36.77**
Pamphlet	26	23893.93	41.15	23128.28	41.35	22453.65	41.37
	30	15669.90	37.42	14900.70	37.51	14504.50	37.56
	34	9013.55	33.18	8567.73	33.24	8349.78	33.29
	38	3897.73	28.86	3667.88	28.91	3587.02	28.95
Average		**13118.78**	**35.15**	**12566.15**	**35.25**	**12223.74**	**35.29**
Harbour	26	45656.58	38.04	45680.28	37.62	45337.92	37.81
	30	29713.93	34.18	28617.86	33.73	28830.11	33.96
	34	16805.14	30.36	15471.99	30.03	15889.86	30.23
	38	7646.22	26.24	6768.94	26.09	7082.92	26.22
Average		**24955.47**	**32.20**	**24134.77**	**31.86**	**24285.20**	**32.06**

TABLE III
BD-RATE SAVING

Sequences	Adaptive GOP vs. GOP2	Adaptive GOP vs. GOP4
Coastguard	-0.04	-26.24
Suzie	-2.28	-7.52
Pamphlet	-9.04	-3.26
Harbour	-2.12	-1.48
Average	**-3.37**	**-9.62**

V. CONCLUSION

In this paper, machine learning based GOP size selection is proposed for DVC codec. J48 decision tree algorithm is used for training and classification a set of video segments in order to choose the suitable GOP size for each segment including five frames. The results show that performance of the proposed method is better than using fixed GOP sizes or at least, it could choose the best size between GOP2 and GOP4. Future works will focus on finding more effective features and more powerful machine learning algorithm in order to improve the performance of DVC codec.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01 - 2016.15.

REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjontegaard, A. Luthra, "Overview of the H.264/AVC Video Coding Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, Issue 7, pp.560 - 576, Jul. 2003.

[2] G.J. Sullivan et al., "Overview of the High Efficiency Video Coding (HEVC) standard," IEEE TCSVT, vol.22, no. 12, pp. 1649-1668, Dec. 2012.

[3] J. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources," IEEE Trans. on Information Theory, vol. 19, no. 4, pp. 471-480, July 1973.

[4] A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," IEEE Trans. on Information Theory, vol. 22, no. 1, pp. 1-10, January 1976.

[5] X. HoangVan and B. Jeon, "Flexible Complexity Control Solution for Transform Domain Wyner-Ziv Video Coding," IEEE Transaction on Broadcasting, pp. 209-220, Vol. 58, Issue 2, June 2012

[6] X. HoangVan, et al., "A Flexible Side Information Generation Scheme using Adaptive Search Range and Overlapped Block Motion Compensation," Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, No. 46, Korea, February 2011.

[7] J. Ascenso, C. Brites and F. Pereira, "Content adaptive WZ video coding driven by motion activity," Proceedings of the International Conference on Image Processing, Atlanta, Georgia, USA, 2006.

[8] I. Ahmad, Z. Ahmad and I. Abou-Faycal, "Delay efficient GOP size control algorithm in WZ video coding," IEEE International Symposium on Signal Processing and Information Technology, 2009.
[9] K. R. Vijayanagar and J. Kim, "Dynamic GOP size control for low-delay distributed video coding," 2011 18th IEEE International Conference on Image Processing
[10] K. DinhQuoc, X. HoangVan, and B. Jeon, "An Iterative Algorithm for Efficient Adaptive GOP Size in Transform Domain Wyner-Ziv Video Coding," Lecture Note in Computer Science, vol.7088, pp.348-358, 2011.
[11] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," Proceedings of the ITU-T Video Coding Experts Group (VCEG) Thirteenth Meeting, Austin, April 2001.
[12] G. Holmes, A. Donkin, I.H. Witten, "Weka: A machine learning workbench," Proceedings of Australian New Zealand Intelligent Information Systems Conference, Australia, 1994.