

LETTER

A CNN-Based Optimal CTU λ Decision for HEVC Intra Rate Control

Lili WEI[†], Nonmember, Zhenglong YANG^{†a)}, Member, Zhenming WANG^{††},
and Guozhong WANG^{†††}, Nonmembers

SUMMARY Since HEVC intra rate control has no prior information to rely on for coding, it is a difficult work to obtain the optimal λ for every coding tree unit (CTU). In this paper, a convolutional neural network (CNN) based intra rate control is proposed. Firstly, a CNN with two last output channels is used to predict the key parameters of the CTU R- λ curve. For well training the CNN, a combining loss function is built and the balance factor γ is explored to achieve the minimum loss result. Secondly, the initial CTU λ can be calculated by the predicted results of the CNN and the allocated bit per pixel (bpp). According to the rate distortion optimization (RDO) of a frame, a spatial equation is derived between the CTU λ and the frame λ . Lastly, The CTU clipping function is used to obtain the optimal CTU λ for the intra rate control. The experimental results show that the proposed algorithm improves the intra rate control performance significantly with a good rate control accuracy.

key words: HEVC, intra rate control, CNN, rate distortion optimization

1. Introduction

Rate control is mainly used to achieve optimal quality of the coding video under the given target bit rate. For high-efficiency video coding (HEVC), the λ domain rate control [1] builds a typical relationship between the Lagrange multiplier, the bit per pixel (bpp) and the rate control parameters. For λ domain rate control, some researches studied on the CTU level [2] and the frame level [3] bit allocation. Additionally, a separated λ decision algorithm [4] was proposed for the surveillance videos.

However, as the intra rate control has a significant effect on the following frames, it is a important work to improve the intra coding performance. Wang *et al.* [5] allocated bpp of intra CTU according to the gradient per pixel (gpp), but the bpp assigned by the traditional method was difficult to satisfy the CTU spatial coding feature. Therefore, the convolutional neural network (CNN)-based intra rate control has been proposed that achieves an outstanding performance. Santamaria *et al.* [6] gave the bit and distortion prediction for intra rate control by two CNNs. Lu

et al. [7] predicted the coefficients of the R-Quantization model by a designed CNN. As Li *et al.* [8] suggested that QP was not suitable for the flexible coding structure of HEVC which indicated that using QP as the key factor for rate control would limit coding performance. In [9], two CNNs were used to predict two parameters of the R- λ curve. However, it ignored the whole coding performance of a frame and the predicted λ might be singular. Therefore, it is necessary to explore the spatial relationship of every CTU λ for the reasonable coding in a whole frame.

As intra rate control has little prior coding information, it is difficult to improve intra coding quality under a high control accuracy. In this paper, a designed CNN is used to predict the key parameter of the R- λ relationship. To minimize the training loss of the CNN, a combining loss function is built and the balance factor γ is explored. Then, with the rate distortion optimization (RDO) process, a spatial equation of the CTU λ and the frame λ can be derived. Lastly, according to the RDO derivation, a special clipping method is used to obtain the optimal CTU λ for coding the whole frame reasonably.

2. Proposed CNN-Based Optimal CTU λ Decision

2.1 R- λ Relationship of CTU

We ran the HEVC encoder numerous times to observe the R- λ characteristics of different CTUs with QP = 22, 27, 32 and 37 respectively at all intra configuration. The experimental results are shown in Fig. 1.

It is observed from Fig. 1 that the R- λ curve of a CTU obeys the hyperbolic function that can be modeled as

$$\lambda = \alpha \cdot R^{-\beta} \quad (1)$$

where α and β are the key model parameters. R is mod-

Manuscript received May 2, 2021.

Manuscript revised June 20, 2021.

Manuscript publicized July 19, 2021.

[†]The authors are with the School of Urban Rail Transportation, Shanghai University of Engineering Science, Shanghai 201620, China.

^{††}The author is with the Shanghai Branch of Communication Information Group Co., Ltd, Shanghai 201600, China.

^{†††}The author is with the Artificial Intelligence Industry Research Institute, Shanghai University of Engineering Science, Shanghai 201620, China.

a) E-mail: 949713872@qq.com

DOI: 10.1587/transinf.2021EDL8047

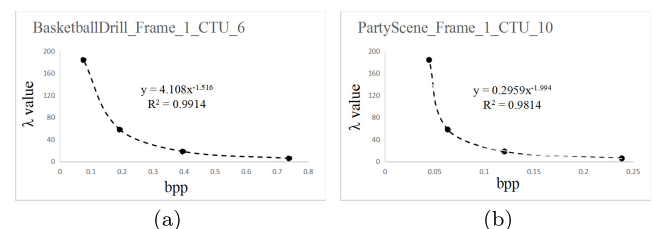


Fig. 1 R- λ curves for (a) the 6th CTU in BasketballDrill at the 1st frame, (b) the 10th CTU in PartyScene at the 1st frame

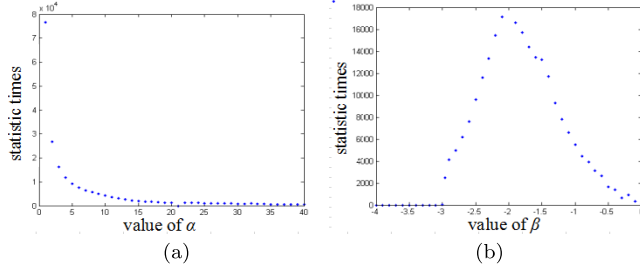


Fig. 2 Statistical results for parameter (a) α and (b) β

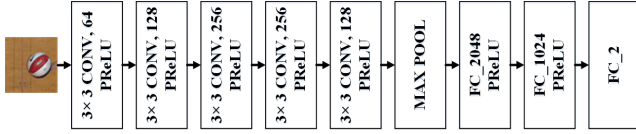


Fig. 3 CNN structure for predicting α and β

eled by bpp. We also observe that different CTUs have quite different R- λ characteristics from each other. On the other hand, if the accurate R- λ curve of a CTU can be obtained prior to coding, it will achieve the optimal coding performance with any allocated bpp as long as the R- λ point is located on the curve.

2.2 CNN-Based CTU R- λ Prediction

As described by Eq. (1), we use the CNN to predict α and β . We pick out approximately 5000 frames from different kinds of sequences. To reduce the correlation of the adjacent frames, ten frames that consist of the first frame and nine frames of the integer multiple of the frame rate after the first frame are selected from every sequence. Every frame is compressed by HEVC for all intra configuration with QP = 22, 27, 32 and 37. Every CTU in a single image will have four points, and then the R- λ curve of the CTU will be fitted by the four points to obtain α and β in Eq. (1). We obtain statistics for α and β from 300000 R- λ curves, with the results shown in Figs. 2 (a) and (b), respectively.

In Figs. 2 (a) and (b), the horizontal axis represents the values of α and β , respectively, and the vertical axis represents for the number of times that this value is obtained. It is observed that the values of α are mostly located in [0, 15]. The values of β are mostly located in [0, -3]. Therefore, α and β that are found in [0, 15] and [0, -3], respectively, will be used as the ground truth for training. Last, for the CNN, 200000 CTUs are used for training and 20000 CTUs are used for testing. The CNN structure is shown in Fig. 3.

It is observed from Fig. 3 that the CNN consist of five convolutional (CONV) layers, seven parametric rectified linear unit (PReLU) layers, one max-pooling layer and three full connection (FC) layers where the last FC layer will be used for prediction. As the last FC layer has two outputs, the combining loss function is defined as

$$\text{Loss} = \frac{1}{2N} \sum_{n=1}^N \{ \| (x_n - X_n) \|^2 + \gamma \cdot \| (y_n - Y_n) \|^2 \} \quad (2)$$

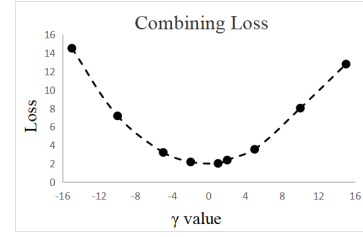


Fig. 4 Combining loss results with different γ

where N is the number of the CTUs used for training. x_n and y_n are the prediction of α and β from the CNN, respectively. X_n and Y_n are the ground truth of α and β , respectively. γ is the balance factor. The training is performed by stochastic gradient descent with standard back-propagation.

For training the CNN, γ is changed to explore the combining loss result. We set the maximum number of training iterations to 1500000 and observe that after approximately 1100000 iterations, the training and testing loss results tend to be stable. The combining loss results with different γ are shown in Fig. 4.

From Fig. 4, it can be seen that the combining loss has the smallest value, which is nearly to 1.87, when γ equals to 1. Therefore, we obtain α and β for Eq. (1) from the CNN which is trained with the combining loss function as γ equals to 1.

2.3 Optimal CTU λ Decision

To minimize the distortion for the intra rate control, the RDO problem of a frame can be modeled as

$$\begin{aligned} \min_{\{r_{CTU}(i)\}_{i=1}^{N_{CTU}}} D_F &= \sum_{i=1}^{N_{CTU}} d_{CTU}(i) \\ \text{s.t. } R_F &= \sum_{i=1}^{N_{CTU}} r_{CTU}(i) \leq T_F \end{aligned} \quad (3)$$

where D_F and R_F are the distortion and the bit rate of the current frame, respectively. N_{CTU} is the number of CTUs in the current frame. T_F is the target bit rate. Then, the Lagrange multiplier method can be converted to an equivalent unconstrained problem as

$$\min_{\{r_{CTU}(i)\}_{i=1}^{N_{CTU}}} \sum_{i=1}^{N_{CTU}} d_{CTU}(i) + \lambda_F \cdot \sum_{i=1}^{N_{CTU}} r_{CTU}(i) \quad (4)$$

where $z \in [1, N_{CTU}]$ and λ_F is the Lagrange multiplier of the current frame. Then, Eq. (4) can be solved by setting its derivative to zero,

$$\frac{\partial \sum_{i=1}^{N_{CTU}} d_{CTU}(i)}{\partial r_{CTU}(z)} + \lambda_F \frac{\partial \sum_{i=1}^{N_{CTU}} r_{CTU}(i)}{\partial r_{CTU}(z)} = 0 \quad (5)$$

As every CTU is encoded separately for intra rate control. Taking $\lambda_{CTU}(i) = -\partial d_{CTU}(i) / \partial r_{CTU}(i)$ into Eq. (5), we can obtain,

$$\sum_{i=1}^{N_{CTU}} \lambda_{CTU}(i)/N_{CTU} = \lambda_F \quad (6)$$

where Eq. (6) indicates that the mean λ of the CTUs in a frame should be equal to the current frame λ . As all the CTU λ s can not be obtained once time for calculating the final mean λ in Eq. (6), a clipping method is used to make the mean λ of the encoded CTUs tend to the frame λ , which is defined as

$$\begin{aligned} & \text{clip}\{\max[\sum_{u=1}^{i-1} \lambda_{CTU}(u)/(i-1) \cdot a, \lambda_F \cdot a], \\ & \min[\sum_{u=1}^{i-1} \lambda_{CTU}(u)/(i-1) \cdot b, \lambda_F \cdot b], \lambda_{CTU}(i)\} \end{aligned} \quad (7)$$

where a and b are the empirical values which are $2^{-1/3}$ and $2^{1/3}$, respectively. In Eq. (7), $\lambda_{CTU}(i)$ of i th encoding CTU is effected not only by the frame λ but also by the mean CTU λ of the encoded CTUs. Thus, Eq. (6) can be approximately satisfied.

3. Experimental Results

The proposed algorithm is implemented into HM16.9. The default intra rate control method, which are the CTU level (HMModel1) and frame level rate control (HMModel2), of HM16.9 and the intra rate control of Li *et al.* [9] are used for comparison. It should note that we use our data set for training the CNNs which are proposed in [9], and turn off rate control to compress the first frame with QP = 22, 27, 32 and 37, then use the resulting bit rates as the target bit rates. We follow the HEVC common test conditions with 15 test sequences. The bit rate accuracy is defined as

$$M = |T_{\text{actual}} - T_{\text{target}}|/T_{\text{target}} \quad (8)$$

where T_{actual} is the actual bit rate. T_{target} is the target bit rate.

The bit rate accuracy results are shown in Table 1.

It is observed from Table 1 that the average bit rate accuracy results of the four algorithm are 1.61%, 2.21%, 5.75% and 13.06%, respectively. Since the average bit rate accuracy of HMModel2 is much worse than those of the other algorithms, it is inconsistent with the main purpose of the rate control. On the other hand, control of the bit rate is a highly challenging task for intra coding with only one frame. The accuracies of the proposed algorithm, the algorithm of Li *et al.* [9] and HMModel1 remain high.

We use HMModel1, HMModel2 and the algorithm of Li *et al.* [9] as the comparison respectively to obtain the BDRate and BDPSNR results, which are shown in Table 2. An examination of the data presented in Table 2 shows that the average BDRate and BDPSNR of the proposed algorithm and HMModel1 are -1.36% and 0.07 dB. This means that the proposed algorithm always uses less bit rate, but improves the coding quality. For the comparison of the proposed algorithm and the algorithm of Li *et al.* [9], the BDRate is decreased by 2.28% and the the BDPSNR is increased by 0.13 dB. The proposed algorithm has a better coding performance mainly about two reasons, the first one is that a deeper CNN structure is used which will lead to a more accurate prediction and the calculated λ is more suitable for every CTU. The other one is that every CTU λ is clipped together and the whole frame can be coded reason-

Table 1 Bit rate accuracy results of HMModel1, the algorithm of Li *et al.* [9], the proposed algorithm and HMModel2

Sequences	HMModel1 M%	Li <i>et al.</i> [9] M%	Proposed M%	HMModel2 M%
Class A	3.01	5.22	8.50	15.33
Class B	1.41	1.40	6.27	12.88
Class C	1.04	1.03	4.40	12.52
Class D	0.95	1.16	3.83	11.53
Average	1.61	2.21	5.75	13.06

Table 2 Experimental comparison results for the proposed algorithm with HMModel1, the algorithms of Li *et al.* [9] and HMModel2

Sequences		Proposed versus HMModel1		Proposed versus Li <i>et al.</i> [9]		Proposed versus HMModel2	
		BDRate	BDPSNR	BDRate	BDPSNR	BDRate	BDPSNR
Class A	PeopleOnStreet	-0.86	0.05	-1.82	0.10	2.36	-0.13
	Traffic	-1.10	0.06	-2.29	0.13	2.57	-0.14
	BasketballDrive	-2.37	0.06	-2.30	0.05	2.64	-0.07
	BQTerrace	-1.84	0.11	-3.06	0.19	4.04	-0.23
Class B	Cactus	-2.36	0.09	-3.58	0.14	5.13	-0.19
	Kimono	-1.42	0.05	-2.49	0.09	2.78	-0.10
	ParkScene	-1.79	0.08	-2.55	0.11	2.33	-0.10
	BasketballDrill	-2.08	0.10	-2.41	0.11	2.56	-0.12
Class C	BQMall	-1.07	0.07	-2.12	0.14	2.97	-0.18
	PartyScene	-0.56	0.05	-0.75	0.06	0.48	-0.03
	RaceHorses	-1.20	0.08	-1.85	0.12	1.87	-0.12
	BasketballPass	-0.68	0.04	-1.87	0.12	3.05	-0.19
Class D	BlowingBubbles	-0.60	0.03	-0.73	0.04	1.11	-0.06
	BQSquare	-0.63	0.05	-3.40	0.29	4.36	-0.36
	RaceHorses	-1.86	0.14	-2.99	0.21	3.55	-0.24
Average		-1.36	0.07	-2.28	0.13	2.79	-0.15

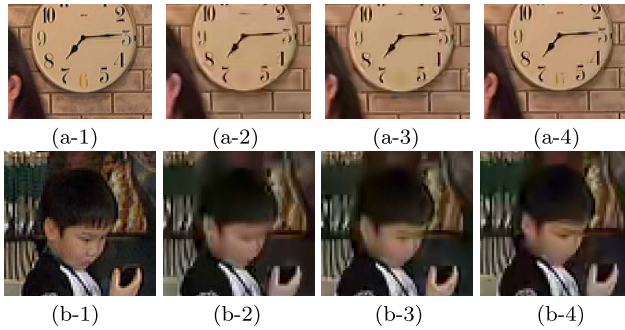


Fig. 5 Partial subjective performance of PartyScene for 1st frame @ 19605.2 kbps and BQMall for 1st frame @ 5971.2 kbps. (a-1) and (b-1) are the ground truth images, (a-2) and (b-2) are the images from HMMode1, (a-3) and (b-3) are the images from Li et al. [9], (a-4) and (b-4) are the images from the proposed algorithm.

ably. With the comparison of HMMode2, the BDRate increases 2.79 and the BDPSNR decreases 0.15. However, HMMode2 sacrifices too much rate control accuracy which does not satisfy the purpose of intra rate control. Therefore, the proposed algorithm has the better rate control performance than the other algorithms. Some subjective comparisons are shown in Fig. 5.

It is observed from Fig. 5 (a-2) that the number 6 on the clock is missing as well as that in Fig. 5 (a-3). However, the proposed algorithm, which can be seen in Fig. 5 (a-4), has a clear result in the same region. Similarly, the subjective qualities of the face regions presented in Figs. 5 (b-2) and (b-3) are worse than that of the image in Fig. 5 (b-4). Since the intra rate control frame will be referred by the following frames, it is important to improve the intra coding quality as much as possible under the target bit rate.

4. Conclusions

In this paper, a CNN-based intra rate control is proposed. Firstly, a CNN is designed to predict CTU $R-\lambda$ relationship. Then, the combining loss function with the balance factor γ is built for training the CNN. Secondly, the CTU λ can be obtained by the training results and the allocated bpp.

However, the calculated λ is not suitable for coding. Then, with the RDO precess of a frame, the spatial equation relationship of the CTU λ and the frame λ can be derived. Lastly, the spacial limitation of every CTU in a frame is built for the reasonable coding of the whole frame. For the experimental results, the proposed algorithm has a better rate control performance than the comparison algorithms.

Acknowledgements

This work was supported by the major project of Zhangjiang (Grant No. ZJ2020-ZD-009).

References

- [1] B. Li, H. Li, L. Li, and J. Zhang, “ λ domain rate control algorithm for High Efficiency Video Coding,” *IEEE Trans. Image Process.*, vol.23, no.9, pp.3841–3854, 2014.
- [2] Z. Chen and X. Pan, “An Optimized Rate Control for Low-Delay H.265/HEVC,” *IEEE Trans. Image Process.*, vol.28, no.9, pp.4541–4552, 2019.
- [3] H. Guo, C. Zhu, S. Li, and Y. Gao, “Optimal Bit Allocation at Frame Level for Rate Control in HEVC,” *IEEE Trans. Broadcast.*, vol.65, no.2, pp.270–281, 2019.
- [4] Z. Yang, G. Wang, and G.-W. Teng, “A Foreground-Background-Based CTU λ Decision Algorithm for HEVC Rate Control of Surveillance Videos,” *IEICE Trans. Inf. & Syst.*, vol.E102-D, no.3, pp.670–674, 2019.
- [5] M. Wang, K.N. Ngan, and H. Li, “An Efficient Frame-Content Based Intra Frame Rate Control for High Efficiency Video Coding,” *IEEE Signal Process. Lett.*, vol.22, no.7, pp.896–900, 2015.
- [6] M. Santamaria, E. Izquierdo, S. Blasi, and M. Mrak, “Estimation of Rate Control Parameters for Video Coding Using CNN,” 2018 IEEE Visual Communications and Image Processing (VCIP) IEEE, 2018.
- [7] X. Lu, B. Zhou, X. Jin, and G. Martin, “A Rate Control Scheme for HEVC Intra Coding Using Convolution Neural Network (CNN),” 2020 Data Compression Conference (DCC), 2020.
- [8] B. Li et al., “Rate control by R-lambda model for HEVC,” presented at 11th Meeting on JCTVC-K0103, JCTVC of ISO/IEC and ITU-T, Shanghai, China, 2012.
- [9] Y. Li, B. Li, D. Liu, and Z. Chen, “A convolutional neural network-based approach to rate control in HEVC intra coding,” 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017.