A Novel Visual Analysis Oriented Rate Control Scheme for HEVC

Qi Zhang
Institute of Digital Media
Peking University
Beijing, China
ywwynm@pku.edu.cn

Shanshe Wang
Institute of Digital Media
Peking University
Beijing, China
sswang@pku.edu.cn

Siwei Ma
Institute of Digital Media
Peking University
Beijing, China
swma@pku.edu.cn

***Abstract*—Recent years have witnessed an explosion of machine visual intelligence. While impressive performance on visual analysis has been achieved by powerful Deep-Learning-based models, the texture and feature distortion caused by image and video coding is becoming a challenge in practical situations. In this paper, a new rate control scheme is proposed to improve visual analysis performance on coded video frames. Firstly, a new kind of visual analysis distortion is introduced to build a Rate-Joint-Distortion model. Secondly, the Rate-Joint-Distortion Optimization problem is solved by using Lagrange multiplier method, and the relationship between rate and Lagrange multiplier λ is described by a hyperbolic model. Thirdly, a logarithmic $\lambda - QP$ model is established to achieve minimum Rate-Joint-Distortion cost for given λs. The experimental results show that the proposed scheme can improve visual analysis performance with stable bits used for coding.**

***Index Terms*—video coding, rate-distortion optimization, rate control, visual analysis**

I. Introduction

In recent years, videos are broadly used for machine visual analysis rather than only for humans to watch. Large-scale, high-quality datasets significantly empower learning-based computer vision models, especially Deep-Learning models. However, due to the limitations of storage and transmission, videos are usually compressed and coded before sent to models that are deployed in application environments, leading to a decrease of video quality as well as noticeable performance loss for visual analysis. Particularly in low bit-rate situations, models may even lose the ability to extract adequate features on distorted video frames at all, which lower the performance to an unacceptable level.

As an essential part of video coding technologies, rate control plays an important role in real-world video applications. The goal of rate control is to encode video within a target bit-rate range and achieve the best possible video quality. It can be regarded as an optimization problem considering both bit-rate and video quality, or rate and distortion (RD) caused by video coding. Obviously, the definition and measurement of distortion influence the solution to Rate-Distortion Optimization (RDO) significantly. In common video coding standards and software, the distortion is often defined as signal-level and measured by pixel-wise calculations like mean squared error (MSE), which is fair for human watching. However, in the context of video coding for visual analysis, the signal-level distortion is not equivalent to the distortion of video frame features or vision task performance, resulting in non-optimal RD cost for visual analysis.

Some recent researches focus on rate control for visual analysis under the analyze-then-compress (ATC) scheme, which compresses and encodes only features from pristine video frames. Zhang et.al proposed a Rate-Accuracy Optimization model for SIFT feature coding, in which the feature distortion is estimated by ranking differences between the original and compressed feature descriptors when performing pair-wise feature matching [1]. Similarly, Ding et.al proposed a Rate-Performance-Loss Optimization model for deep feature coding, in which the retrieval performance loss is estimated from actual feature distortion by probabilistic models [2]. Li et.al proposed Texture-Feature-Quality-Index as the distortion caused by joint video and feature coding for face recognition, which is used as the optimization target during joint coding [3]. Nevertheless, the rate control scheme for pure video coding on the purpose of visual analysis is not studied sufficiently yet.

Our contributions in this work are as follows[1]:

- A new kind of visual analysis distortion is introduced to build a Rate-Joint-Distortion (RJD) model for visual analysis oriented rate control.
- A hyperbolic model is built to describe the relationship between rate and Lagrange multiplier of λ that is used to solve the Rate-Joint-Distortion Optimization (RJDO) problem.
- A logarithmic model is built to describe the relationship between λ and the best quantization parameter (QP) that minimizes RJD cost.

The rest of the paper is organized as follows. In section II, the RJD model is described and the solution to RJDO is given. In section III, the relationship between λ and QP is investigated and fitted. Section IV shows the experimental setups and results. And finally, section V concludes the paper.

[1]This work was supported in part by the National Key R&D Program of China (2017YFC0821005), and High-performance Computing Platform of Peking University, which are gratefully acknowledged.

978-1-7281-8068-7/20/$31.00 ©2020 IEEE

II. Rate-Joint-Distortion Optimization

The RJDO problem for rate control can be defined as:

$$\min D^*, \quad \text{s.t. } R \leq R_C \tag{1}$$

where R is the actual bits used for coding and R_C is the target bits. D^* stands for the joint distortion caused by coding, which consists of signal-level distortion D_t and visual analysis distortion D_p with different weights of ω_t and ω_p:

$$D^* = \omega_t D_t + \omega_p D_p, \quad \omega_t + \omega_p = 1 \tag{2}$$

Equation (1) is a constrained optimization problem that can be converted to unconstrained as (3) by using Lagrange multiplier methods:

$$\min J^* = D^* + \lambda \cdot R \tag{3}$$

where J^* is the joint cost, and λ is the Lagrange multiplier that controls the relative importance between D^* and R.

In order to solve (3), let the partial derivative function of J^* to R equals 0:

$$\frac{\partial J^*}{\partial R} = \frac{\partial D^*}{\partial R} + \lambda = 0 \tag{4}$$

Then we have:

$$\lambda = -\frac{\partial D^*}{\partial R} \tag{5}$$

According to (5), λ is determined by R and D^*. Therefore, a $R-D^*$ model should be established. Ignoring D_p, the relationship between R and D_t can be described by a hyperbolic function [4]. We assume the $R-D_p$ model is in a similar formulation as (6):

$$D_p = C_p \cdot R^{-K_p}, \quad C_p, K_p > 0 \tag{6}$$

where C_p and K_p are model parameters.

Without loss of generality, taking object detection as an example of visual analysis task, the proposed definition of D_p is expressed as (7):

$$D_p = \frac{P(0) - P(R)}{P(0)} \tag{7}$$

where $P(0)$ is the object detection performance on pristine video frames that can be measured by mean average precision (mAP), and $P(R)$ is the performance on compressed and distorted frames. After normalization, the value of D_p is between 0 and 1.

To verify (6), experiments are performed on a subset of HEVC common test sequences. Firstly, we encode selected sequences using different QPs with rate control disabled and record the bits that are used. Secondly, we run the widely-used Faster RCNN [5] model with ResNet-101 [6] as feature extractor on encoded frames to get the object detection performance. It should be mentioned that the ground truth object boxes of pristine video frames are annotated by a much powerful state-of-the-art object detection model [7] from framework [8], instead of annotating them manually. Only boxes with a prediction score larger than 0.8 are preserved as ground truth to generate more reasonable and robust annotations. Finally, the $R-D_p$ curve is fitted by (6), with R expressed in terms of *bpp* (bits per pixel) and D_p calculated by (7) using prior records.

Currently, we focus on building the $R-D_p$ model for B-frames. Some fitting results are shown in Fig 1, which verifies the hyperbolic relationship between R and D_p. Therefore, a similar assumption is made and experimental results fitting is processed for $R-D^*$ model according to (8):

$$D^* = \omega_t D_t + \omega_p D_p = C^* \cdot R^{-K^*}, \quad C^*, K^* > 0 \tag{8}$$

where C^* and K^* are model parameters.

Because the original value of D_p is in $[0, 1]$, it is not directly comparable with D_t that is usually measured by MSE. We enlarge D_p by a scale factor $\gamma_p = 255$ due to some observations on the experiments. After balancing, some fitting results of (8) for different ω_t and ω_p are shown in Fig. 2.

According to Fig. 2, larger ω_p leads to more serious joint distortion. In this paper, we fix ω_p and ω_t to be 0.9 and 0.1, respectively. After taking more sequences into consideration, a more general $R-D^*$ model as (9) and the corresponding $R-\lambda$ model as (10) are given:

$$D^* = 6.1822 \times R^{-0.5299} \tag{9}$$

$$\lambda = -\frac{\partial D^*}{\partial R} = \alpha \cdot R^{\beta} = 3.276 \times R^{-1.5299} \tag{10}$$

The comparison between the fitted $R-D^*$ model and the existing $R-D_t$ model in HEVC reference software (HM 16.16) is shown in Fig 3. It can be observed that the joint-distortion is larger than signal-level distortion when bit-rate is lower, which indicates the serious visual analysis performance loss in such conditions.

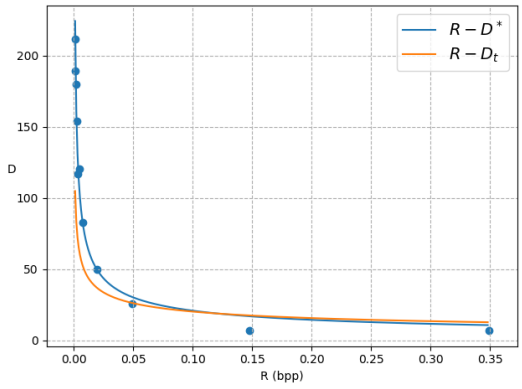
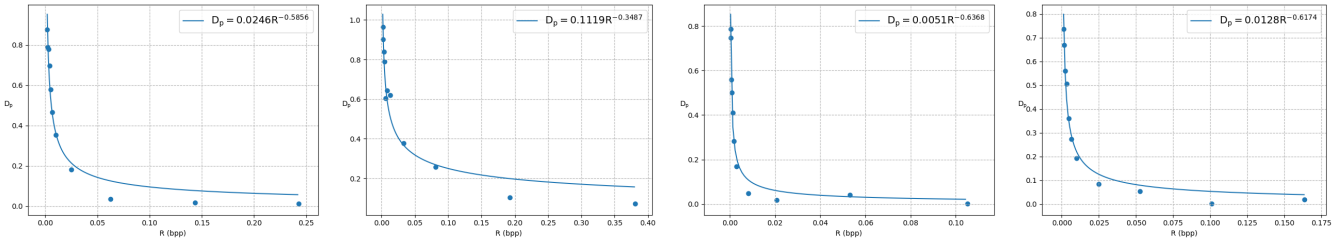


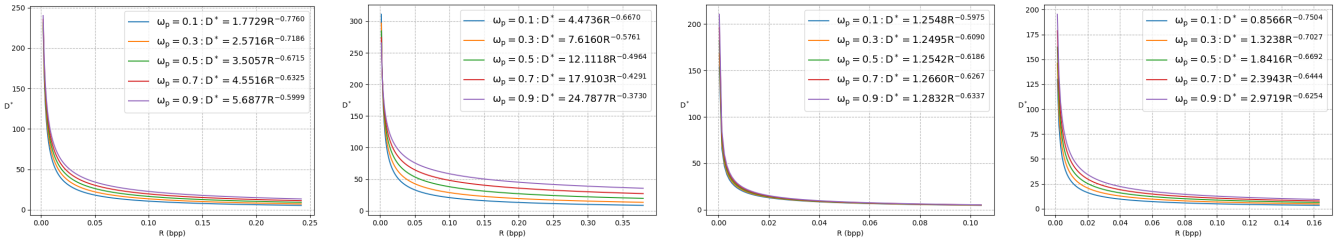
Fig. 3: Fitted $R-D^*$ model and HEVC's $R-D_t$ model

Since different frames are likely to have different $R-D^*$ characteristics, $R-\lambda$ model parameters, i.e. α and β, should be updated during coding. In this paper, we adopt the same update strategy and hyper-parameters like strides as HEVC.



(a) BasketballPass (b) BQMall (c) Market3 (d) RitualDance

Fig. 1: Fitted $R - D_p$ models on HEVC common test sequences



(a) BasketballPass (b) BQMall (c) Market3 (d) RitualDance

Fig. 2: Fitted $R - D^*$ models with different ω_t and ω_p on HEVC common test sequences

III. QP DETERMINATION BY λ

In order to actually perform rate control, the encoder should choose certain encoding parameters to encode the video according to λ. In HEVC, QP is an important parameter that influences both compressed video quality and bit-rate. Therefore, after calculating λ for a target bit-rate by $R-\lambda$ function, the QP value that minimizes RD cost is then determined by a statistical logarithmic model as (11):

$$QP = 4.2005 \times \ln \lambda + 13.7122 \tag{11}$$

Equation (11) is only valid for signal-level distortion D_t measured by MSE. After introducing visual analysis distortion D_p, the following optimization problem extended from (3) should be solved:

$$\min J^*(QP) = D^*(QP) + \lambda \cdot R(QP) \tag{12}$$

In (12), different λ or QP values lead to different R, D^* and J^*. However, λ can be determined by (10) before encoding with rate control enabled, then the only variable of (12) becomes just QP. Because QP values are discrete, a traversal for each QP values can be performed to solve (12).

In this paper, we retrain the $\lambda-QP$ relationship on a subset of HEVC common test sequences like in [9]. Firstly, multiple λ values in a range from 3.0 to 10000.0 are selected to find the best relevant QP values. Secondly, sequences are encoded under different QP values, after which the corresponding bits as well as visual analysis distortions are recorded. Thirdly, for each fixed λ as λ_s, an anchor QP of searching interval is computed through the exponential $QP-\lambda$ model from HEVC, which is written as QP_s. Then J^* can be calculated by (12) for each QP in $[QP_s - 4, QP_s + 8]$ using λ_s, and the best QP that minimizes J^* can be found. As such, the relationship of λ and QP is determined, for which we use a logarithmic function to fit like in HEVC. According to the fitted result, we have:

$$QP = 3.6 \times \ln (\lambda + 16.0129) + 16.1840 \tag{13}$$

Fig 4 compares the fitted model as (13) with the existing $\lambda - QP$ model in HEVC (HM 16.16). It can be observed that ours curve gives smaller QPs than HEVC's when λ enlarges.

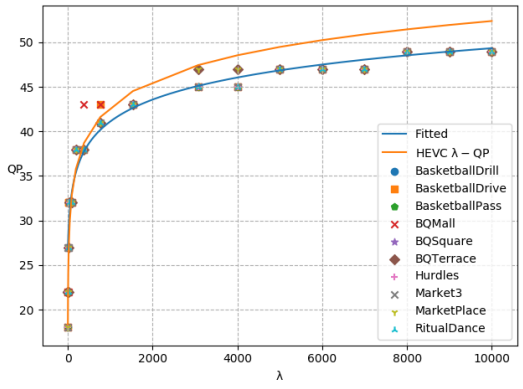


Fig. 4: Fitted $\lambda - QP$ model compared with HEVC

IV. EXPERIMENTS

To test the coding and visual analysis performance of the proposed rate control scheme, experiments are performed on HEVC common test sequences. Because some sequences don't

have valid objects or are too easy to analyze, only a subset of them is selected as shown in Table I to experiment on. The selected sequences are encoded using random access (RA) configuration for 5 seconds under 5 groups of bit-rate settings. The bit-rate group setups are shown in Table II.

TABLE I: Selected sequences for experiments

Resolution	Name	Frame rate	Object categories
416x240	BasketballPass	50	Person, Ball
	BQSquare	60	Person, Umbrella
832x480	BasketballDrill	50	Person, Ball
	BQMall	60	Person
1920x1080	BasketballDrive	50	Person, Ball
	BQTerrace	60	Person, Vehicle
	Hurdles	50	Person
	Market3		Person, Car
	MarketPlace	60	Person
	RitualDance		Person

TABLE II: Bit-rate groups for experiments

Resolution	Group (Bit-rate in kbps)				
	A	B	C	D	E
416x240	100	150	200	250	300
832x480	240	310	380	450	520
1920x1080	470	560	650	700	740

We compare the proposed rate control scheme with the existing one in HM 16.16. The RJD performance of rate control schemes is evaluated from three aspects, the rate performance, the signal-level video quality, and the visual analysis performance. The rate performance is measured by bit-rate error (BE) that is calculated from the difference of target bit-rate R_{target} and actual bit-rate R_{actual} as shown in (14). The signal-level video quality is measured by peak signal-to-noise ratio (PSNR) of luma component. As for testing the visual analysis performance, two important and hot computer vision tasks are selected, which are object detection and human pose estimation. For object detection, the models for annotation as well as prediction are introduced in section II. For human pose estimation, the ground truth annotations are made by Pose-HRNet-w48 [10] and predictions are from Pose-ResNet-101 [11]. We use mAP to evaluate the prediction results on coded video frames as visual analysis performance for both tasks.

$$BE = \frac{|R_{actual} - R_{target}|}{R_{target}} \times 100\% \tag{14}$$

The experimental results are shown in Table III, where AP_1 and AP_2 are object detection mAPs for all categories and just person category, respectively, and AP_3 is pose estimation mAP. Both BE and APs are expressed in terms of percentage. According to the results, the proposed scheme provides better visual analysis performance, which has averaged gains of 0.85%, 1.18% and 1.54% for three kinds of mAPs in absolute values, considering all of the test sequences and bit-rate settings. The rate performance as well as the signal-level video quality are close to HEVC.

TABLE III: Coding and visual analysis performance comparison between HEVC and proposed rate control scheme

Bit-rate Group	HEVC					Proposed				
	BE	PSNR	AP_1	AP_2	AP_3	BE	PSNR	AP_1	AP_2	AP_3
A	4.29	29.06	15.03	49.95	52.10	4.46	28.93	**15.68**	**51.23**	**53.97**
B	2.97	30.23	16.67	54.47	56.73	3.66	30.16	**17.52**	**55.81**	**58.21**
C	2.92	31.12	18.20	58.64	60.14	3.04	31.05	**19.11**	**59.70**	**61.98**
D	2.61	31.77	18.99	60.12	61.79	2.62	31.71	**20.06**	**61.52**	**63.17**
E	2.23	32.30	19.61	61.56	63.18	2.36	32.24	**20.39**	**62.40**	**64.30**

V. Conclusions

In this paper, we propose a new rate control scheme to improve visual analysis performance on coded video frames. Firstly, a new kind of visual analysis distortion measured by normalized task performance loss is introduced to build a Rate-Joint-Distortion model. Secondly, the Rate-Joint-Distortion Optimization problem is solved by introducing a Lagrange multiplier λ, and the relationship between rate and λ is then fitted by a hyperbolic model. Thirdly, a logarithmic $\lambda - QP$ model is established to achieve minimum Rate-Joint-Distortion cost for given λ by traversing through all possible QP values in the searching range. The experiments prove the effectiveness of the proposed scheme.

References

[1] X. Zhang, S. Ma, S. Wang, X. Zhang, H. Sun, and W. Gao, "A joint compression scheme of video feature descriptors and visual content," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 633–647, 2016.

[2] L. Ding, Y. Tian, H. Fan, Y. Wang, and T. Huang, "Rate-performance-loss optimization for inter-frame deep feature coding from videos," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5743–5757, 2017.

[3] Y. Li, C. Jia, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Joint rate-distortion optimization for simultaneous texture and deep feature compression of facial images," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pp. 1–5, IEEE, 2018.

[4] B. Li, H. Li, L. Li, and J. Zhang, "λ domain rate control algorithm for high efficiency video coding," *IEEE transactions on Image Processing*, vol. 23, no. 9, pp. 3841–3854, 2014.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[7] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, *et al.*, "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4974–4983, 2019.

[8] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[9] B. Li, J. Xu, D. Zhang, and H. Li, "Qp refinement according to lagrange multiplier for high efficiency video coding," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 477–480, IEEE, 2013.

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

[11] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.