Masking Effects Based Rate Control Scheme for High Efficiency Video Coding

Hao Wang[*†], Li Song[*†], Rong Xie[*†], Zhengyi Luo[‡], Xiangwen Wang[‡]
[*]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[†]Cooperative Medianet Innovation Center, Shanghai, China
[‡]School of Electronics and Information Engineering, Shanghai University of Electric Power
E-mail: {wanghao123, song_li, xierong}@sjtu.edu.cn, lzy@shiep.edu.cn, wxw21st@gmail.com

Abstract—**This paper presents a masking effects based rate control scheme for high efficiency video coding (HEVC). Rate control is regarded as a very effective tool to improve the performance of video coding under the limited bandwidth. However, the state-of-the-art rate control algorithm based on R-λ model ignores the characteristics of human visual system (HVS), which leads to poor performance in subjective quality. Moreover, some structural similarity (SSIM) or saliency based perceptual rate control algorithms only consider spatial characteristics. Since spatial and temporal visual masking effects can better reflect the characteristics of HVS, in this paper masking effects based perceptual factor for coding tree unit (CTU) is proposed, which takes both texture complexity and motion information into account. Then the proposed perceptual factor is utilized to guide bit allocation in CTU-level rate control. Experimental results show that the proposed scheme can effectively improve the coding performance compared with the R-λ algorithm.**

I. Introduction

High Efficiency Video Coding (HEVC) is the latest video coding standard developed by the Joint Collaborative Team on Video Coding (JCT-VC) [1], [2]. Compared with the preceding video coding standard H.264/AVC [3], multiple new coding tools including Advanced Motion Vector Prediction (AMVP) and Sample Adaptive Offset (SAO) have been integrated into HEVC, which have shown significant improvements. Almost 50% bitrate can be saved while keeping the same coding quality. As an essential tool in video coding, rate control contributes to maximizing the coding quality under the limited bandwidth, especially for real-time applications such as video conference. In previous video coding standards, different algorithms are adopted for rate control. For example, TM5 [4] algorithm is applied in MPEG-2, and TMN8 [5] is used in H.263. In MPEG-4 VM8 [6] is utilized. The rate control algorithm specified in [7] is adopted by H.264/AVC.

R-λ model based rate control algorithm [8] is adopted in HEVC, which achieves better coding performance than the previous algorithm based on unified RQ model [9]. However, the characteristics of human visual system (HVS) are not fully excavated and explored among these existing rate control algorithms. In addition, commonly used quality metrics such as peak signal-to-noise ratio (PSNR) are not always consistent with HVS. Recently, some visual quality assessment (VQA) metrics were proposed, which match well with HVS. As one of classical perceptual metrics, structural similarity (SSIM) [10] considers the degradation of structural information and comes closer to subjective perception. Zhao *et al.* [11] proposed an SSIM-Motivated rate control scheme in which SSIM was applied to bit allocation in CTU-level rate control. In addition, another VQA metric gradient magnitude similarity deviation (GMSD) [12], based on gradient information, was employed in perceptual-based CTU-level rate control [13]. Furthermore, saliency based rate control scheme was proposed to improve subjective visual quality, especially for the region of interest (ROI) [14]. However, temporal characteristics or motion information are not fully exploited in these methods.

Different from these common rate control schemes, a novel masking effects [15] based rate control scheme is proposed in this paper, which pays more attention to spatial and temporal characteristics of HVS as well as video contents. More specifically, if a local region is relatively smooth and stationary, more bits tend to be allocated there in order to reduce visual quality loss under the same distortion. On the contrary, less bits will be allocated to the regions with high texture complexity and motion intensity due to more evident visual masking effects occurring in these regions.

In this paper, we first propose a perceptual factor based on spatial and temporal visual masking effects, which is derived from a perceptual video quality metric, mean opinion score (MOSp) [16]. As a novel VQA metric, MOSp correlates much more closely with subjective visual quality compared with common quality metrics such as PSNR and SSIM. Then the perceptual factor is used to guide bit allocation in CTU-level rate control. The target bits of each CTU will be dynamically adjusted according to the texture complexity and the motion information of local regions. Experimental results show that the proposed scheme can achieve better coding performance compared with the state-of-the-art R-λ algorithm with regard to both subjective and objective quality metrics.

The rest of this paper is organized as follows. In Section II, masking effects based perceptual factor for each CTU is proposed, followed by a new bit allocation scheme for CTU-level rate control. Section III shows the experimental results. Finally, Section IV concludes this paper.

II. Proposed Rate Control Scheme

In general, visual masking effect is a common phenomenon in HVS, which occurs in both the spatial and the temporal domain. More specifically, the presence of a stimulus can influence the visibility of another one. In other words, one

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

stimulus is usually masked by another, which makes itself invisible or not detectable.

Spatial and temporal masking effects contribute much more to perceptual quality metrics than other visual characteristics [17]. As for spatial masking effect, the distortion in textured regions is usually hard to be perceived by human eyes than smooth regions. That is, HVS is less sensitive to the distortion from textured regions. Therefore, regions with much complex texture can tolerate more distortion than smooth ones. Similarly, for temporal masking effect, the details and the distortion in moving regions cannot be easily perceived by human eyes than stationary regions. As the motion becomes faster, the masking effect tends to be more obvious. So the distortion in moving regions is more tolerable by human eyes. As a result, better perceptual performance can be achieved in textured or moving regions than smooth or stationary ones if the same distortion is introduced.

In this section, we first propose a novel perceptual factor based on visual masking effects, which is derived from MOSp and calculated according to texture complexity and motion information. Then the proposed perceptual factor is utilized as new weight to guide bit allocation in CTU-level rate control.

A. Masking Effects Based Perceptual Factor

Spatial edge information is usually employed to estimate texture details [12]. In this paper, edge strength is used to measure the texture complexity of a local region. Sobel edge detection algorithm [18] is adopted to obtain edge strength in consideration of its simplicity and robustness. Specifically, horizontal and vertical gradient magnitude are first obtained respectively for each pixel. Then the average edge strength of each CTU is derived as follows.

$$G=\frac{1}{M}\sum\nolimits_{i,j}\sqrt{\left|G_h(i,j)\right|^2+\left|G_v(i,j)\right|^2} \qquad (1)$$

where G_h and G_v represent horizontal and vertical gradient at each pixel, respectively. M denotes the total number of pixels of a coding block. Then the final edge strength k_g of each CTU is obtained by the average edge strength of the whole frame.

$$k_g=\frac{G(i)}{\frac{1}{N}\sum_{j=1}^{N}G(j)} \qquad (2)$$

where G(i) denotes the edge strength of the i-th CTU computed by (1). N is the total number of CTUs in current frame.

In addition, the variance coefficient related to SSIM was proposed to adjust the Lagrange multiplier in rate-distortion optimization (RDO) process and contribute to improving subjective quality concerning SSIM [19], which also indicates the texture complexity of a region. The more textured a region, the larger the Lagrange multiplier. As a result, less bits will be allocated to this region in RDO process, which corresponds to the spatial masking effect. So the variance coefficient k_σ is taken as a supplement to k_g, which is expressed as:

$$k_\sigma=\frac{2\sigma_i^2+c_2}{exp\left(\frac{1}{N}\sum_{j=1}^{N}log\left(2\sigma_j^2+c_2\right)\right)} \qquad (3)$$

where σ_i^2 denotes the average variance of the i-th CTU, and N represents the total number of CTUs in current frame. c_2 is a constant used for numerical stability.

After combining (2) and (3), the spatial perceptual factor k_s is defined as follows.

$$k_s=(1-\tau)\times k_g+\tau\times k_\sigma \qquad (4)$$

where τ is a weight value ranging in [0,1]. It's obvious that larger edge strength and variance contribute to larger k_s, which indicates that more complex texture occurs in this region.

In order to measure temporal motion information, the motion intensity of local regions is characterized as the temporal perceptual factor utilizing motion vectors, which is similar to the motion activity term defined in [20].

$$k_t=log\sqrt{\frac{\sum\nolimits_{i,j}v_x^2(i,j)+v_y^2(i,j)}{d(i,j)}} \qquad (5)$$

where (v_x, v_y) is the motion vector of current block and $\mathrm{d(i,j)}$ denotes the distance from current frame to its reference frame. Considering the trade-off between accuracy and computation complexity, the motion vectors are obtained separately for every nonoverlapping 16×16 block in current CTU. Moreover, in this paper the previous frame of current frame is chosen to be its reference frame, so as to reduce complexity.

As discussed in section I, MOSp performs much better than common perceptual quality metrics such as SSIM in subjective assessment [16], which is defined as:

$$MOSp=1-k\times MSE \qquad (6)$$

where k is the slope related to edge strength, and MSE is weighted to simulate HVS response. Similarly, a new VQA metric (VQM) based on MOSp was proposed to guide window-level rate control and accomplish consistent visual quality control in [20], which is defined as follows.

$$VQM=\omega\times(1-k\times MSE) \qquad (7)$$

where ω denotes the motion activity term. Inspired by this VQM presented in (7), we propose a new perceptual factor to guide bit allocation in CTU-level rate control. After obtaining k_s and k_t with (4) and (5) respectively, the perceptual factor for each CTU k_p is defined as:

$$k_p=\frac{c}{c+k_t}(1-k_s\times MSE) \qquad (8)$$

where c is a constant with consistent magnitude of k_t. The derived perceptual factor k_p is closely associated with both spatial texture complexity and temporal motion intensity, which fairly correspond to spatial and temporal masking effects. For textured and moving regions, k_s and k_t tend to have larger values. From (8), it clearly indicates that higher spatial and temporal complexity will lead to smaller perceptual factor k_p and less target bits that will be allocated in rate control.

B. CTU-Level Bit Allocation

As for group of pictures (GOP) level and picture level rate control, the bit allocation scheme is still in accordance with [8]. However, in CTU-level rate control, the perceptual factor k_p for each CTU is utilized as new weight to guide bit allocation, which is conducted in the initial bit allocation for every CTU before encoding the whole frame. The new initial allocated target bits T_{CTU}^{*} of current CTU is expressed as follows.

$$T_{CTU}^{*}=\frac{T_{pic}}{\sum\limits_{AllCTUs}k_{p}}\times k_{p} \tag{9}$$

where T_{pic} denotes the target bits of current frame. T_{CTU}^{*} is further clipped and limited to the range between 0.25 and 4 times the anchor initial target bits $T_{CTU}^{\sim}$ to avoid very few extreme fluctuation, which is conducted as follows.

$$T_{CTU}^{*}=clip\left(0.25\times T_{CTU}^{\sim},4\times T_{CTU}^{\sim},T_{CTU}^{*}\right) \tag{10}$$

To achieve a balance between the perceptual weight and the anchor weight for bit allocation, the final initial target bits T_{CTU} is weighted as follows. τ is still the same weight value as (4).

$$T_{CTU}=\left(1-\tau\right)\times T_{CTU}^{\sim}+\tau\times T_{CTU}^{*} \tag{11}$$

Finally, the actual target bits of current CTU $T_{currCTU}$ will be dynamically adjusted according to current buffer status and the initial allocated target bits T_{CTU}, which is expressed as:

$$T_{currCTU}=T_{CTU}+\frac{\sum\limits_{NotCoded}T_{CTU}-B_{left}}{SW} \tag{12}$$

where B_{left} denotes the actual left bits of current frame, and SW represents the smooth window size.

III. EXPERIMENTAL RESULTS

The proposed perceptual bit allocation scheme described in Section II is implemented in HM 16.15 [21]. Experiments are conducted to verify the performance of the proposed scheme with default low delay P configuration [22]. Encoding results including actual bitrate, PSNR, SSIM and MOSp are recorded, respectively. BD-Rate, bitrate mismatch and time complexity are calculated to fully compare the coding performance with the anchor scheme in HM 16.15, where R-λ rate control algorithm is incorporated. Moreover, the target bitrates for test sequences set in this experiment are shown in Table I.

A. R-D Performance

Popular video quality metrics including PSNR, SSIM and MOSp are adopted in this paper to measure and compare the coding performance. The last two metrics, proved to be much more consistent with HVS, provide a good estimation of perceptual visual quality. Table II shows the detailed BD-Rate results of the proposed scheme compared with the anchor.

As shown in Table II, the average BD-Rate reaches -1.29% for PSNR, -3.81% for SSIM and -4.67% for MOSp. It can be seen that perceptual quality has been significantly improved. Meanwhile, objective quality has not declined and also been refined. In the best situation, the proposed scheme can achieve up to -10.1% and -12.5% gains for SSIM and MOSp, respectively. These results have effectively demonstrated that better improvements can be achieved in perceptual coding performance with the proposed scheme, especially for the sequences with high texture complexity and motion intensity such as BasketballDrive and PartyScene. Besides this, Fig. 1 shows the R-D curves of sequences, BQSquare and BQTerrace, which also indicate that the proposed scheme can provide significant perceptual quality improvements.

TABLE I. TARGET BITRATE

	Sequence	Target bitrate (kbps)
Class B	ParkScene, Kimono	6000,4000,1600,1000
	Cactus,BasketballDrive,BQTerrace	10000,7000,3000,2000
Class C	BasketballDrill, BQMall PartyScene, RaceHorses	2000,1200,512,384
Class D	BasketballPass, BQSquare BlowingBubbles, RaceHorses	1500,850,384,256
Class E	FourPeople, KristenAndSara	4000,2000,1000,512

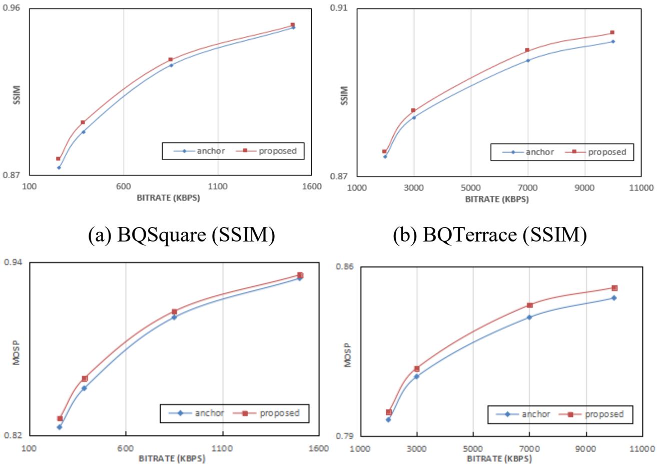
TABLE II. BD-RATE

	Sequence	BD-Rate (PSNR)	BD-Rate (SSIM)	BD-Rate (MOSp)
Class B	Kimono	-0.8%	-2.4%	-3.3%
	ParkScene	-2.7%	-4.7%	-5.2%
	Cactus	-1.8%	-3.6%	-5.5%
	BasketballDrive	-1.8%	-3.9%	-3.4%
	BQTerrace	-1.9%	-10.1%	-12.5%
Class C	BasketballDrill	-0.7%	-1.0%	-0.7%
	BQMall	-1.6%	-2.9%	-4.1%
	PartyScene	-2.4%	-3.7%	-4.3%
	RaceHorses	-0.8%	-2.6%	-2.5%
Class D	BasketballPass	-1.0%	-3.2%	-3.7%
	BQSquare	1.2%	-8.3%	-8.6%
	BlowingBubbles	-1.6%	-5.3%	-6.3%
	RaceHorses	0.0%	-3.3%	-3.7%
Class E	FourPeople	-1.3%	-0.8%	-3.8%
	KristenAndSara	-2.1%	-1.3%	-2.5%
	Average	-1.29%	-3.81%	-4.67%

In addition, Fig. 2 shows the overall fluctuation of SSIM and MOSp of the sequence BlowingBubbles encoded at 256 kbps. From Fig. 2(a) and 2(b), it can be seen that higher perceptual quality is achieved for most frames with the proposed scheme. Subsection C will present more intuitive details.

B. Rate Control Performance

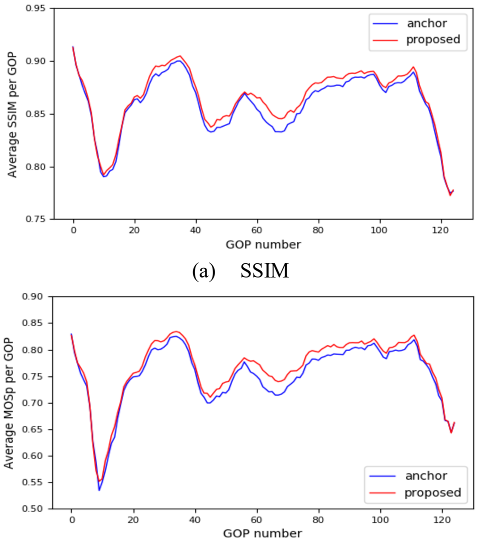
Bitrate mismatch ratio is defined to estimate and compare the accuracy of rate control algorithm, which is expressed as:



(a) BQSquare (SSIM) (b) BQTerrace (SSIM)

(c) BQSquare (MOSp) (d) BQTerrace (MOSp)

Fig. 1 R-D Curve



(a) SSIM

(b) MOSp

Fig. 2 Fluctuation of SSIM and MOSp

$$Mismatch\% = \frac{\left|R_{actual} - R_{target}\right|}{R_{target}} \times 100\% \qquad (13)$$

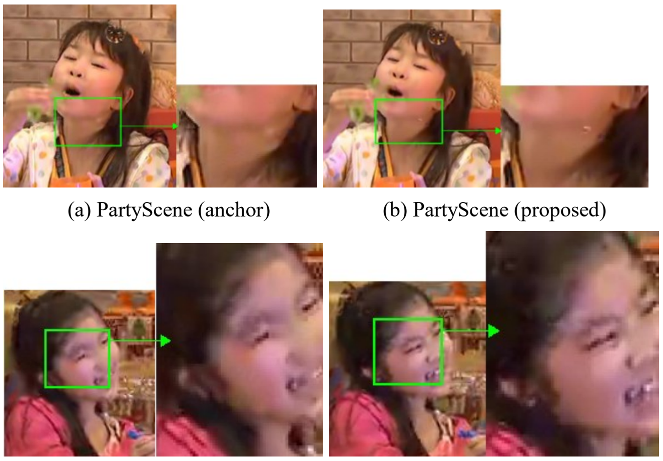
where R_{actual} and R_{target} represent actual bitrate and target bitrate, respectively. The average of mismatch ratios at four bitrate points presented in Table I is taken as the final mismatch ratio of a sequence. Table III shows the bitrate mismatch ratios resulting from the anchor and the proposed scheme. It can be observed that the proposed scheme can achieve almost the same average rate control accuracy as the anchor scheme.

C. *Subjective Improvement*

From Table II, it clearly shows that the proposed scheme can effectively improve the perceptual quality. In this subsection, several decoded frames are shown in Fig. 3 to visually verify the improvements in subjective quality. Fig. 3 shows the 360th decoded frame of PartyScene and the 498th decoded frame of BlowingBubbles, which are encoded at 1200 kbps and 256 kbps, respectively. It can be seen that the details of the bubble on the neck in Fig. 3(b) and the face including the eyes and the nose in Fig. 3(d) become much more clear after adopting the proposed scheme.

TABLE III. MISMATCH OF RATE CONTROL

	Sequence	Anchor	Proposed
Class B	Kimono	0.05%	0.05%
	ParkScene	0.04%	0.05%
	Cactus	0.04%	0.04%
	BasketballDrive	0.05%	0.04%
	BQTerrace	0.06%	0.05%
Class C	BasketballDrill	0.24%	0.24%
	BQMall	0.29%	0.29%
	PartyScene	0.24%	0.24%
	RaceHorses	0.14%	0.13%
Class D	BasketballPass	0.38%	0.34%
	BQSquare	0.41%	0.41%
	BlowingBubbles	0.35%	0.34%
	RaceHorses	0.21%	0.19%
Class E	FourPeople	0.18%	0.15%
	KristenAndSara	0.19%	0.22%
	Average	0.19%	0.19%



(a) PartyScene (anchor)

(b) PartyScene (proposed)

(c) BlowingBubbles (anchor)

(d) BlowingBubbles (proposed)

Fig. 3 Comparison of Subjective Quality

D. *Coding Complexity*

The encoding time of several video sequences is recorded for the anchor and the proposed scheme respectively to estimate and compare the coding complexity. The test platform is a PC equipped with 2.70 GHz Intel(R) Xeon(R) CPU and 64G memory. The encoding time ratio is obtained with the encoding time of the proposed scheme divided by that of the anchor for each sequence. The comparing results show that the total average encoding time of the proposed scheme has increased nearly 1.2% compared with the anchor, which is almost negligible. That means the proposed scheme has reached a compromise between the coding performance and the time complexity. Most of the increased time is consumed in the computation of the texture complexity and the motion intensity.

IV. CONCLUSIONS

In this paper, masking effects based rate control scheme is proposed to achieve better coding performance. Since the spatial and temporal masking effects can better reflect the characteristics of HVS, masking effects based perceptual factor for each CTU is derived from MOSp and employed as new weight to guide bit allocation in CTU-level rate control. Since the regions with low spatial and temporal complexity are more sensitive to distortion, more bits will be allocated to these regions to reduce visual quality loss. Experimental results show that the proposed scheme achieves better perceptual coding performance compared with the R-λ rate control algorithm.

ACKNOWLEDGMENT

This work was supported by NSFC (61671296, U1611461, and 61521062), the 111 Project (B07022 and Sheitc No.150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

References

[1] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[2] B. Li and J. Xu, “An introduction to high efficiency video coding range extensions,” http://wwwen.zte.com.cn/endata/magazine/ztecommunications/2016/1/articles/201603/t20160311_448969.html, accessed March 11, 2016.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[4] *Test Model 5* [Online]. Available: http://www.mpeg.org/MPEG/MSSG/tm5/

[5] J. Ribas-Corbera and S. M. Lei, “Rate control in DCT video coding for low-delay communications”, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp.172 – 185, 1999.

[6] H.-J. Lee, T. Chiang, and Y.-Q. Zhang, “Scalable rate control for MPEG-4 video,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 6, pp. 878–894, Sep. 2000.

[7] K.-P. Lim, G. Sullivan, and T. Wiegand, *Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods*, document Rec. JVT-N046, Hong Kong, China, Jan. 2005.

[8] B. Li, H. Li, L. Li, and J. Zhang, “λ domain based rate control for high efficiency video coding,” *IEEE TIP*, vol. 23, no. 9, Sep. 2014.

[9] H. Choi, J. Nam, J. Yoo, D. Sim and I.V. Bajić, “Rate control based on unified RQ model for HEVC,” ITU-T SG16 Contribution, JCTVC-H0213, San José, Feb. 2012.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

[11] H. Zhao, W. Xie, Y. Zhang, L. Yu and A. Men. An SSIM-motivated LCU-level rate control algorithm for HEVC. *Picture Coding Symposium (PCS)*, 85–88, 2013.

[12] W. Xue, L. Zhang, X. Mou, A. Bovik, "Gradient magnitude similarity deviation: A highly efficiency perceptual image quality index", *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684-695, 2014.

[13] A. Yang, H. Zeng, L. Ma, J. Chen, C. Cai, and K. K. Ma. A perceptual-based rate control for HEVC. In *2016 Sixth International Conference on Image Processing Theory*, Tools and Applications (IPTA), pages 1–5, Dec 2016.

[14] L. Bai, L. Song, R. Xie, J. Xie, and M. Chen. Saliency based rate control scheme for high efficiency video coding. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6, Dec 2016.

[15] J. T. Enns, V. D. Lollo, "What's new in visual masking", *Trends in Cognitive Sci.*, vol. 4, no. 9, pp. 345-352, Sep. 2000.

[16] A. Bhat, I. Richardson, and S. Kannangara, “A new perceptual quality metric for compressed video,” in *Proc. IEEE ICASSP*, Apr. 2009, pp. 993–936.

[17] X. Zhu, W. Hong, H. Xu, L. Yu, and Y. Zhao. Spatial quality index based rate perceptual-distortion optimization for video coding. *Journal of Visual Communication and Image Representation*, 38:423 – 432, 2016.

[18] X. Ran and N. Farvardin, “A perceptually motivated three-component image model, part 1: Description of the model,” *IEEE Trans. Image Process.*, vol. 4, no. 4, pp. 401–415, Apr. 1995.

[19] C. Yeo, H. L. Tan and Y. H. Tan. On rate distortion optimization using SSIM. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7): 1170–1181, 2013.

[20] L. Xu, S. Li, K. N. Ngan, and L. Ma. Consistent visual quality control in video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(6):975–989, June 2013.

[21] *HEVC Test Model* (HM) [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[22] F. Bossen. Document JCTVC-L1100: Common test conditions and software reference configurations. *JCT-VC Meeting*, *Geneva, Switzerland*, *Tech. Rep*, 2013.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE