EMOS-BASED SPATIO-TEMPORAL RATE CONTROL TECHNIQUE FOR 4K60P VIDEOS

Akira YAMANAKA and Toshiyuki YOSHIDA

Dept. of Information Science, Graduate school of Eng., University of Fukui, Fukui 910-8507, Japan

Abstract - The authors have proposed a spatio-temporal rate control technique based on a maximization of the estimated Mean Opnion Score(EMOS), and applied it to video coding. This paper extends the technique to 4k60P videos. To evaluate the effectiveness of the apporach, the technique is applied to a 4K60P sequence comprising four scenes with different spatio-temporal activity. The results suggest that the maximum frame rate 60fps is insufficient for some videos, and that 120 fps or higher is desireble in order to make the best of the 4K resolution.

Keywords : Video coding, 4K60P video, Video quality, Mean opinion score, Rate control

I. INTRODUCTION

In video coding, the bit rate r [bit/s] is expressed as

$$r = s \cdot t \tag{1}$$

in terms of the average bit amounts per frame (BPF) s [bit/frame] and the frame rate (FPS) t [frame/s]. The BPF s determines spatial quality of every single frame while the FPS s dominates temporal quality, i.e., motion smoothness. Under a bit-rate constraint, the two parameters should be controlled in a well-balanced manner in order to maximize spatio-temporal total video quality.

Videos have been encoded mainly with its frame rate fixed in broadcasting services, package media, and so on. In a software-based encoding in network applications, however, the frame rate can be varied to adapt spatio-temporal activities of a target scene. Such a variable-frame-rate video encoding requires a video quality metric that can evaluate spatial and temoral video qualities equivalently on a unified scale. In some literature, the mean squared error (MSE) or equivalently the peak signal-to-noise ration (PSNR), has been utilized to adaptively control the frame rate [1]. However, it is widely known that the PSNR cannot be utilized for comparing qualities of a group of videos although it is quite useful in evaluating spatial quality within a single target video sequaence. Such a property of the PSNR suggests that it cannot evaluate spatial and temporal video qualities on a unified scale.

Based on the backgound, the authors have proposed rate control techniques based on a spatio-temporal estimated Mean Opinion Score (EMOS) [2–5]. The MOS is a direct subjective metric obtained by averaging scores of subjects who evaluate target video quality on a prescibed rating scale, and thus can evaluate the spatial and temporal qualities on a unified scale. The five-point MOS scale ranging from 1 to 5 is commonly utilized in Refs.[2-8] and in this paper. Table 1 shows the MOS rating and the corresponding labels.

Table 1 : MOS rating and the corresponding labels

Rating	Label
5	Excellent
4	Good
3	Fair
2	Bad
1	Poor

Since the MOS can only be measured through subjective assessment tests, an estimation for the spatio-temporal MOS is necessary so that it can be actually applied in a video rate control process. The authors have thus proposed estimation techniques for the spatio-temporal EMOS distibution, and have applied the estimation technique to a spatio-temporal rate control and coding [5].

The EMOS esitmation techniques [3,4] target starndard- and high-definition (SD and HD) videos and do not cover the ultra high-definition (UHD) standard. Although we have thus extended the estimation techniques to UHD 4k60P videos, they have applied to neigher an actual rate control nor coding. This paper therefore applies the 4k60P EMOS estimation technique to actual rate conrol and encoding frameworks, and evaluates its effectiveness on a real 4k60P video sequence.

In what follows, Sect. II reviews our EMOS-based sptio-temporal rate control and the EMOS estimation techniques extended for 4k60P videos. Sect. III applies our techniques to a 4k60P test sequence comprising 4 scenes in order to evaluate its actual behavior and effectiveness. Finally, Sect. IV concludes this paper.

II. REVIEW OF OUR APPROACH

A. EMOS-based spatio-temporal rate control framework

Since video quality depends on spatial and temporal ones, the MOS is also dependent on the two qualities, and thus the EMOS should be estimated spatio-temporally in terms of s and t as EMOS(s,t). Ref. [2] has confirmed through numbers of subjective assessment tests that EMOS(s,t) can

978-1-5386-2615-3/18/$31.00

be estimated independenly in terms of s and t, which leads to

$$\text{EMOS}(s,t) = \frac{\text{EMOS}_s(s) \cdot \text{EMOS}_t(t)}{\text{EMOS}_{max}}, \quad (2)$$

where $\text{EMOS}_s(s)$ and $\text{EMOS}_t(t)$ are the spatial EMOS estimated for s and the temporal one for t, respectively, and EMOS_{max} is the maximum value in the MOS rating scale ($\text{EMOS}_{max} = 5$ in this paper).

The estimation techniques for $\text{EMOS}_s(s)$ and $\text{EMOS}_t(t)$ have been proposed in Refs. [3] and [4], respectively, and a combination of these techniques in Eq. (2) enables us to estimate the $\text{EMOS}(s,t)$ surface for a given video scene. Fig. 1 illustrates the $\text{EMOS}(s,t)$ surface for a video scene. Once $\text{EMOS}(s,t)$ is estimated, it can be utilized in several applications, one of which includes the spatio-temporal rate control that finds the optimal frame rate for a video scene.

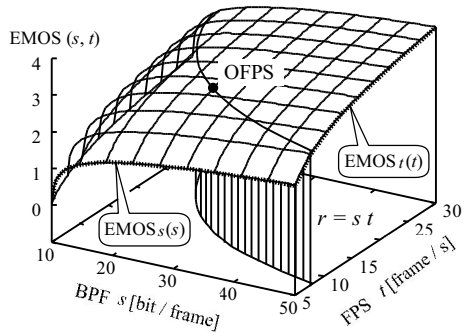


Fig. 1 : EMOS distribution for BPF s and FPS t

Eq. (1) indicates that s is invesely proportional to t in encoding a video scene for a prescribed bit rate r. and that the spatial and temporal video qualities are in a tradeof relationship: in particular for a relatively low bit rate, a large (small) t leads to smooth (jaggy) motion but leads to low (hight) frame quality. This fact suggests that the optimal frame rate exists that gives the highest spatio-temporal total quality, which can be estimated by using $\text{EMOS}(s,t)$. Since (s,t) satisfying Eq. (1) forms a hyperbolic curve as in Fig. 1, the optimal frame rate (OFPS) can be readily estimated as the maximum point along with the hyperbolic curve. By encoding the target scene with the OFPS t_{opt} [frame/s] and the corresponding BPF $s = r/t_{opt}$, we can obtain the optimal bitstream that has the highest total quality for the prescribed bit rate r [5]. Actually, such an approach is applied to video sequences preceded by dividing them into a single "scene" whose spatio-temporal video activity is substantially constant.

Another application of the EMOS estimation is a realization of constant MOS (CMOS) variable bit rate (VBR) coding. Since the MOS can be compared over video scenes unlike the PSNR, CMOS coding can be readily realized by varying the bit rate of each scene so that the resulting EMOS remains constant over scenes.

B. Extention to 4k60P videos

As the EMOS estimation techniques in Refs. [3,4] target upto HD videos, they have been extended to UHD 4k60P videos [6-8]. This section summarizes the techniques.

The spatial $\text{EMOS}_s(s)$ is directly affected by encoder efficiency and encoder parameters such as the search range of motion vectors. To avoid such a dependency, Ref. [3] introduces the PSNR in estimating the EMOS from s, and divides the esimation into those from s to PSNR and from PSNR to EMOS; a variation of encoder parameters directly affects the former, while the latter remains unchanged over the variation. The same strategy has been extended to 4k60P videos [6,7].

The estimation technique for the EMOS from the PSNR in Ref. [3] is directly extended to 4k videos in Ref. [6]. Let x [dB] denote the PSNR of a target video scene. Then, EMOS(x) is estimated by

$$\text{EMOS}_s(x) = aT(x-T)^3 + bT(x-T) + c, \quad (3)$$

where the constants $a = 0.00001151$, $b = 0.0009652$, and $c = 5$ have been obtained in Ref. [6]. T represents the PSNR that gives the highest EMOS at $x = T$, which can be estimated as

$$T = c_0 + c_1 \ln(\epsilon+1) + c_2 \ln(E+1) + c_3 \ln(L+1) \quad (4)$$

where $c_0 = 85.99$, $c_1 = -4.333$, $c_2 = -1.775$, and $c_3 = -3.044$ have been obtained in Ref. [6]. ϵ, E, and L are the feature values extracted from the target scene, each of which is edge statistic, the average residual in motion estimation between successive frames, and the average luminance over frames [6].

An estimation technique of the PSNR from s has been proposed for HEVC encoders in Ref. [7]. The PSNR of a video scene encoded by an HEVC encoder with the BPF s can be estimated by an encoding it in a single R-D point by the technique [7]. Let p [dB] be the PSNR of the target scene encoded at the prescribed bit rate 5.0 [Mbit/s] with an HEVC encoder. Then, the PSNR can be estimated by the model

$$\text{PSNR} = a(s-c)^b, \quad (5)$$

where the constants a, b, and c can be estimated by substituting p into

$$a = a_0(p-a_1)^{a_2} \quad (6)$$
$$b = b_0(p-b_1)^{b_2} + b_3 \quad (7)$$
$$c = c_2 \tanh c_0(p-c_1) - c_3. \quad (8)$$

a_0–c_3 are the constants given as [7]

$$a_0 = 30.13,\ a_1 = 24.06\ , a_2 = 0.1678$$
$$b_0 = -0.003406,\ b_1 = 28.45,\ b_2 = 1.109$$
$$b_3 = 0.1174,\ c_0 = 3.585,\ c_1 = 34.04$$
$$c_2 = 0.003829,\ c_3 = -0.01630$$

A combination of Eqs.(5) and (3) enables us to esimate $\text{EMOS}_s(s)$ for a target scene.

On the other hand, Ref. [8] has proposed an estimation technique for $\text{EMOS}_t(t)$ by using the succesive frame difference D as its feature value. In the technique, the temporal EMOS is estimated by

$$\text{EMOS}_t(t) = \frac{4}{e^{f(t,D)}+1} + 1, \quad (9)$$

where $f(t, D)$ is given by

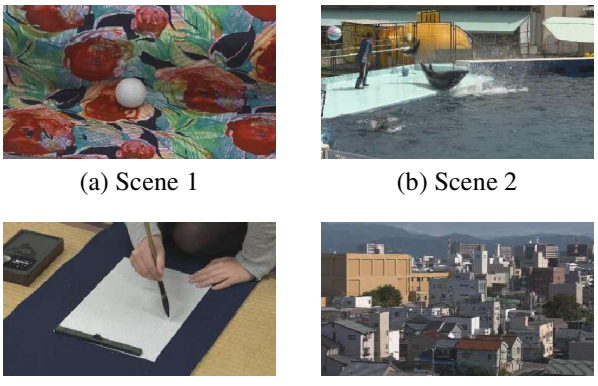
$$f(t, D) = (at + b) \ln D + ct + d \tag{10}$$

with $a = 0.02625$, $b = 0.1726$, $c = -0.1022$, and $d = 1.729$ as constants. By extracting the feature value D from a target scene, the EMOS can be estimated for a given value of t by Eqs. (9) and (10).

By combining the techniques in this section, the EMOS surface EMOS(s, t) can be estimated for a 4k60P video scene.

III. EXPERIMENTS

To evaluate actual behavior and effectiveness of the spatio-temporal rate control and the CMOS coding technique extended to 4k60P videos, they were applied to a 4k60P test sequence comprising four scenes with different spatio-temporal activities. Each scene comprises 96 frames, whose first one is shown in Fig. 2.



(a) Scene 1 (b) Scene 2

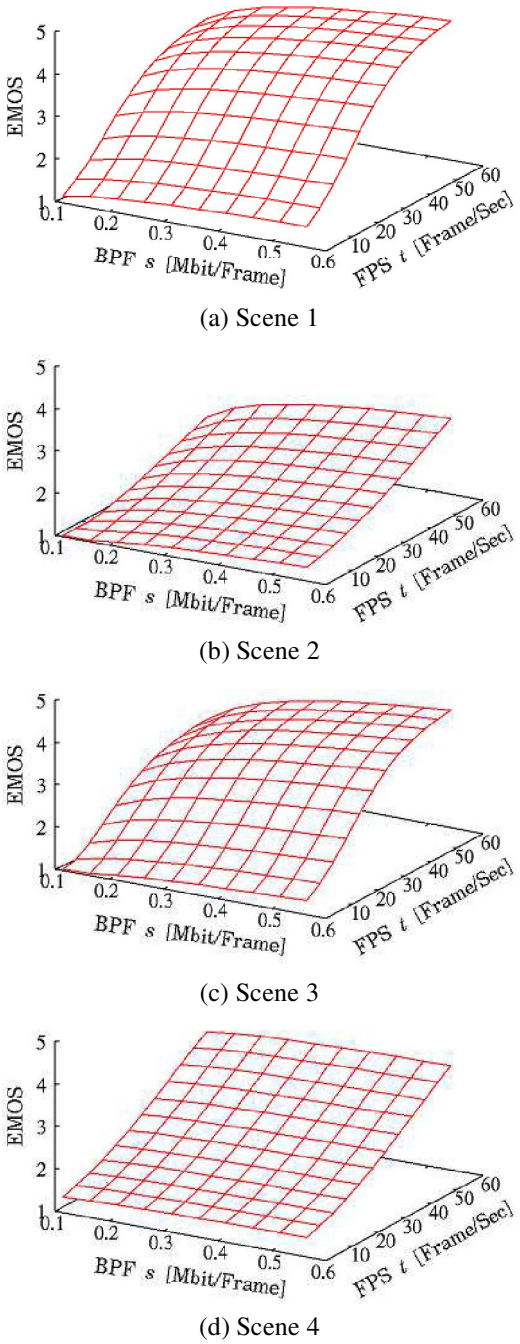
(c) Scene 3 (d) Scene 4

Fig. 2 : First frame of the four scenes.

The feature values ϵ, F, L, D were first extracted from each scene to estimate EMOS(s, t) surface, which is given in Fig. 3. Note in Fig. 3 that the EMOS sustantially saturates along with the FPS t axis for the scenes 1 and 3 while it is not the case for the scenes 2 and 4, i.e., the EMOS remains less than 5 at 60 fps, which greatly affects the results in the following experiments.

A. Spatio-temporal rate control

The spatio-temporal rate control was applied to each scene to estimate its OFPS for the bit rate $r = 15$ [Mbit/s]. Table 2 summarizes the estimated OFPS together with the corresponding EMOS, where the value in "Maximu 60 [fps]" shows the OFPSs estimated within the range 15–60 [fps].



(a) Scene 1

(b) Scene 2

(c) Scene 3

(d) Scene 4

Fig. 3 : Estimated EMOS(s, t) for scenes 1–4.

It can be scene from Table 2 that the EMOSs at the OFPS for the scenes 2 and 4 are relatively low. This is because the EMOS$_t(t)$ does not saturate at $t = 60$ [fps] indicating that even higher frame rate is necessary to make the best of 4k video capability. In fact, if we assume that an extrapolation EMOS$_t(t)$ over 60 fps is valid and that a selection of 120 fps is allowed, OFPS $= 120$ is actually selected for the scenes 2 and 4 with the corresponding EMOS 3.56 and 4.20, respectively, as shown in the last column in Table 2.

These observations suggest that a frame rate of 60 fps is not sufficient and that 120 or 240 [fps] may be required in some video sequences in order to improve temporal quality to be equivalent to spatial one in the 4k resolution. The results also shows that the spatio-temporal rate control is

useful for 4K60P videos, similarly for SD and HD videos.

Table 2 : Estimated OFPSs for scenes 1–4.

Scene	Maximum 60[fps]		Maximum 120[fps]	
	OFPS [fps]	EMOS	OFPS [fps]	EMOS
1	60	3.89	60	3.89
2	60	3.13	120	3.56
3	60	3.75	60	3.75
4	60	3.06	120	4.20

B. CMOS VBR coding

In this section, CMOS VBR coding was realized by using the EMOS(s, t) surface for each scene. The bit rate r in each scene was varied to find r on which the OFPS gives the prescribed EMOS 3 or 4. Table 3 summarizes the results, where the required bit rate r together with the corresponding OFPS are listed for EMOS $= 3$ and 4. Note that the maximum frame rate 120 fps was also allowed in Table 3.

Table 3 shows that a frame rate of 60 fps is insufficient for the scenes 2 and 4 while the scene 4 requires the lowest bit rate in the four scenes, indicating that the scene 4 gives high spatial quality with a relatively low bit rate while higher frame rate is necessary to achieve the highest total video quality.

From the viewpoint of the bit rate, it can be observed that the bit rate required for the prescribed MOS greatly fluctuates over scenes, which means that constant bit rate (CBR) coding never relieves fluctuations of total video quality over scenes. In view of the average bit rate, more than three times of the average bit rate is necessary to improve total quality from MOS $= 3$ to 4, which can only be confirmed with such a unified approach for evaluating spatio-temporal video quality.

Table 3 : Results in CMOS VBR coding

Scene	EMOS = 3	EMOS = 4
1	6.0[Mbps] / 60[fps]	18.0[Mbps] / 60[fps]
2	9.6[Mbps] / 120[fps]	26.4[Mbps] / 120[fps]
3	4.8[Mbps] / 60[fps]	27.0[Mbps] / 60[fps]
4	3.6[Mbps] / 120[fps]	9.6[Mbps] / 120[fps]
average	6.0[Mbps]	20.3[Mbps]

IV. CONCLUSIONS

In this paper, we have extended the EMOS-based rate control and coding techniques to 4k60P videos. These techniques were applied to a 4k60P video sequence comprising four scenes, and their behavior and efectiveness were evaluated. The results have confirmed that the techniques also work as expected for 4k60P videos, while they clearly indicate that a frame rate of 60 fps is insufficient for videos with high temporal activity.

The results suggest the necessity of over 60 fps frame rate for UHD videos. We are going to extend our techniques for even higher frame rate.

This work was supported by JSPS KAKENHI Grant Number JP15K00150.

REFERENCES

[1] H. Song and C.-C. J. Kuo, “Rate control for low-bit-rate video via variable encoding frame rates”, IEEE Trans. Circuits & Syst. for Video Tech. vol.11, pp.512–521, 2001.

[2] Y. Inazumi, T. Yoshida, Y. Sakai, Y. Horita, “Video coding technique on an optimal framerate estimation”, IEICE Trans. Comm., vol.J87-B, no.2, pp.292–304, Feb. 2004 (in Japanese).

[3] T. Miyata and T. Yoshida, “Estimation for the subjective image quality based on SNR”, IEICE Trans. fund. vol.J88-A, no.11, pp.1292–1296, Nov. 2005 (in Japanese).

[4] M. Degura and T. Yoshida, “Adaptive frame interval control for image sequences and its subjective quality estimation”, J. of IIEEJ, vol.35, no.5, pp.497–508, Sep. 2006 (in Japanese).

[5] J. Yamgataga, S. Tanaka, and T. Yoshida, “Spatio-temporal rate control based on a maximization of a subjective assessment value and its application to video coding”, J. of ITE, vol.62, no.11, Nov. 2008 (in Japanese).

[6] T. Yoshida, “Estimation of subjective quality for compressed 4K videos”, IEICE Trans. Inf.& Syst., vol. J100-D, no.9, Sep. 2017 (in Japanese).

[7] A. Yamanaka and T. Yoshida, “Precise estimation for RD curves based on measurement of a single RD Point”, Proc. PCSJ2016, P-4-07, Nov. 2016 (in Japanese).

[8] S. Yamaguchi, “Research on improvement of temporal MOS estimation for 4K video sequences”, Univ. of Fukui, Graduation Thesis, Feb. 2017 (in Japanese).