The Application of Equivalent Mean Square Error Method in Scalable Video Perceptual Quality

Daxing Qian(✉), Ximing Pei, and Xiangkun Li

Dalian Neusoft University of Information,
Software Park Road 8, Dalian 116023, Liaoning, China
{qiandaxing, peiximing, lixiangkun}@neusoft.edu.cn

Abstract. Scalable video is a stream video over heterogeneous networks to different clients. To provide the better quality of service (QoS) or quality of experience (QoE) to customer, we propose an Equivalent Mean Square Error (Eq-MSE) method which is developed based on spatial and temporal frequency analysis of input video content. Eq-MSE is used to calculate minimal frame rate (MinFR) for different videos to guarantee motion without jitter. The proposed scheme in this paper can provide better perceptual video quality than without considering the video content impact.

Keywords: SVC · Eq-MSE · MinFR · Perceptual quality

1 Introduction

With the advances of semi-conductor and access network technologies, real-time video streaming becomes more and more popular in our daily life. We can enjoy the videos service at famous website through different networks using heterogeneous devices. How to provide the high quality of service (QoS) or quality of experience (QoE) to different users over heterogeneous networks is a crucial problem for the success of video streaming application. Scalable video coding (SVC) [1, 2] is a full resolution scalable video stream which can be truncated to adapt different requirements imposed by the subscribed users and underlying access networks.

SVC includes temporal, spatial, SNR and combined scalabilities. Temporal scalability is realized by the hierarchical-B prediction [3]. Spatial scalability is achieved by encoding each supported spatial resolution into one layer. SNR scalability includes coarse grain scalability (CGS) and medium grain scalability (MGS) [4]. To achieve the SNR refinement, we usually use different quantization steps at different SNR layers. In this paper, we study the temporal and SNR joint scalability, and the spatial scalability is not mentioned.

Video content have a significant impact on the perceptual quality. For example, a motion intensive video need a larger frame rate to maintain the continuity of the object movement and avoid jitter and guarantee the motion smoothness, while for stationary video, a relatively lower frame rate is enough to provide the decent video quality. For motion-intensive content, bit stream extracted at higher frame rate is favored. On the

X. Gu et al. (Eds.): MLICOM 2017, Part II, LNICST 227, pp. 3–7, 2018.
https://doi.org/10.1007/978-3-319-73447-7_1

other hand, if there are larger high-frequency components (i.e., rich texture) in a single frame of the video, a finer quantization to reach better spatial quality is typically preferred. To solve the problem, we study the spatial and temporal frequency of the input video content, and propose an Equivalent Mean Square Error (Eq-MSE) scheme to derive the minimal frame rate (MinFR) for different video sources to guarantee the motion smoothness and excellent QoE of the decoded video.

This paper is organized as follows. Section 2 introduces the temporal frequency in a video sequence (i.e., motion). In Sect. 3 we introduce the Eq-MSE method to derive the minimal frame rate without jitter for different input video sources. Subjective test evaluation and experimental results are shown in Sect. 4. Section 5 concludes the paper and discusses the future directions.

2 Temporal Frequency

The concept of spatial frequency is introduced in [5].

We can use the function [6, 7]:

$$\begin{aligned}
&\Psi(f_x,f_y,f_t)\\
&= \iiint \psi(x,y,t)\exp\left(-j2\pi\left(f_x x+f_y y+f_t t\right)\right)dxdydt\\
&= \iint \psi_0(x-v_x t, y-v_y t)\cdot \exp\left(-j2\pi\left(f_x(x-v_x t)+f_y\left(y-v_y t\right)\right)\right)dxdy\\
&\quad\cdot \int \exp\left(-j2\pi\left(f_t+f_x v_x+f_y v_y\right)t\right)dt\\
&= \Psi_0(f_x,f_y)\int \exp\left(-j2\pi\left(f_t+f_x v_x+f_y v_y\right)t\right)dt\\
&= \Psi_0(f_x,f_y)\,\delta\left(f_t+f_x v_x+f_y v_y\right)
\end{aligned} \tag{1}$$

where $\Psi_0(f_x,f_y)$ indicates the 2D CSFT of $\psi_0(x,y)$. This function means that a spatial pattern characterized by (f_x,f_y) in the object will lead to a temporal frequency, i.e.,

$$f_t = -f_x v_x - f_y v_y \tag{2}$$

For a video signal, the temporal frequency is 2D position dependent. For a fixed 2D position (*x, y*), its temporal frequency is defined as the number of cycles per second usually denoted by Hertz (Hz).

From (2) we can draw a conclusion that the temporal frequency depends on not only the motion, but also the spatial frequency [6] of the object.

3 Equivalent Mean Square Error (Eq-MSE)

We propose an Eq-MSE method to calculate the SF of general objects in a picture and find the appropriate frame rate [6, 7].

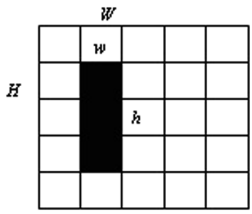

Fig. 1. Illustrative figure for object in general picture.

Figure 1 illustrates that the size of black column is $w \times h$ and picture size is $W \times H$. We use f_{tB} to represent the induced frame rate by the object, which is defined as:

$$f_{tB} = -\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf(1,0)} \cdot v_x - \frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf(0,1)} \cdot v_y \tag{3}$$

where $MSEf(f_x,0)$ is $MSE(x,y)$ when the picture SF is $(f_x,0)$, and $MSEf\left(0,f_y\right)$ is $MSE(x,y)$ when the picture SF is $\left(0,f_y\right)$. v_x and v_y are velocities in horizontal and vertical directions.

We regard the SF of a general picture is:

$$\left(f_x,f_y\right) = \left(\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf(1,0)},\frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf(0,1)}\right) = \left(\frac{h}{H},\frac{w}{W}\right) \tag{4}$$

The objects in a picture that induce the frame rate from moving from arbitrary directions are:

$$f_t = \sum f_{tB} = \sum\left(-\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf(1,0)}v_x - \frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf(0,1)}v_y\right) = \sum\left(-\frac{h}{H}v_x - \frac{w}{W}v_y\right) \tag{5}$$

where $\sum$ is all the MBs in the picture. v_x and v_y are velocities in horizontal and vertical directions of corresponding MB. We get the mode and number of MB in a picture, and then choose the other picture within the same GOP to get MVs according to every MB. The ratio between MVs number in MB and the time interval between two frames are v_x and v_y. For example, the $\sum\frac{h}{H}v_x$ and $\sum\frac{w}{W}v_y$ of sequence Mobile are 12.2, 9.4, respectively. Its MinFR is 12.2 + 9.4 = 21.6. Note that with a real signal, the CSFT is symmetric, so that for every frequency component at $\left(f_x,f_y\right)$, there is also a component at $\left(-f_x,-f_y\right)$ with the same magnitude. The corresponding temporal frequency caused by this other component is $f_xv_x+f_yv_y$ [5].

Equation (5) is the function of minimal frame rate (MinFR) that makes the video motion smoothness without jitter.

4 Experimental Results

We invite 15 experimenters to give the decoded video subjective ratings for evaluate the subjective quality. Sub0 is the default scalable video adaptation without considering the video content impact, while sub1 is scalable adaptation with dependent video content. We use 11 ranks (i.e., 0–10) for the subjective tests ranging. The worst is 0 and the best is 10. The subjective assessment follows [8]. The results show in Table 1.

Table 1. Sequences subjective test comparative results

Sequences	Sub0	Sub1
Akiyo	6.6	6.8
City	4.9	7.7
Mobile	5.7	7.1
Football	*6.1*	7.9

Table 1 depicts the subjective test results of four sequences. It is obviously that "City", "Mobile" and "Football" have better perceptual rating for sub1 session, while "Akiyo" is quite similar between sub1 and sub0. We can draw a conclusion that the Eq-MSE method is providing better-decoded video quality at a given bit rate.

5 Conclusions

In this paper, we propose the Eq-MSE scheme, which is developed based on the spatial and temporal frequency analysis of the video content. This scheme is used to derive the MinFR for different videos and in consequence, so as to guarantee the motion smoothness for decent decoded video quality. Compared with the default scalable video adaptation without considering the video content impact, our proposed scheme can provide better perceptual video quality at the same bit rate according to the subjective quality assessments.

References

1. Text of ISO/IEC 14496-10:2005/FDAM 3 Scalable Video Coding, Joint Video Team (JVT) of ISO-IEC MPEG and ITU-T VCEG, Lausanne, N9197 (2007)
2. ISO/IEC ITU-T Rec. H264: Advanced Video Coding for Generic Audiovisual Services, Joint Video Team (JVT) of ISO-IEC MPEG and ITU-T VCEG, International Standard (2003)
3. Schwarz, H., Marpe, D., Wiegand, T.: Hierarchical B pictures. In: Joint Video Team, Doc. JVT-P014 (2005)
4. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans. Circ. Syst. Video Technol. **17**(9), 1103–1120 (2007)
5. Wang, Y., Ostermann, J., Zhang, Y.-Q.: Video Processing and Communications (2001)

6. Qian, D., Wang, H., Niu, F.: Scalable video coding bit stream extraction based on equivalent MSE method. In: Advanced Materials Research, vol. 204–210, pp. 1728–1732 (2011)
7. Qian, D., Wang, H., Sun, W., Zhu, K.: Bit stream extraction based on video content method in the scalable extension of H.264/AVC. J. Softw. **6**, 2090–2096 (2011)
8. ITU-R Rec. BT.500-11: Methodology for the subjective assessment of the quality of television pictures (2002)