INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 3 (September), pp. 565-573

Packet-Layer Quality Assessment for Networked Video

H. Su, F. Yang, J. Song

Honglei Su, Fuzheng Yang, Jiarun Song
State Key Laboratory of Integrated Service Networks
Xidian University, Xi'an, Shaanxi 710071, China
E-mail:{hlsu,fzhyang}@mail.xidian.edu.cn,
sjrxidian@hotmail.com

Abstract:
To realize real-time and non-intrusive quality monitoring for networked video, a content-adaptive packet-layer model for quality assessment is proposed. Considering the fact that the coding distortion of a video is dependent not only on the bit-rate but also on the motion characteristic of the video content, temporal complexity is evaluated and incorporated in quality assessment in the proposed model. Since very limited information is available for a packet-layer model, an adaptive method for frame type detection is first applied. Then the temporal complexity which reflects the motion characteristic of the video content is estimated using the ratio of the bit-rate for coding I frames and P frames. The estimated temporal complexity is incorporated in the proposed model, making it adaptive to different video content. Experimental results show that the proposed model achieves an advanced performance in comparison with the ITU-T G.1070 model.
Keywords: Packet-Layer Model, Networked Video, Video Quality Assessment, Coding Distortion, Temporal Complexity.

1 Introduction

Recently, with the development of advantage multimedia processing technologies [1], multimedia services such as videophone, mobile conference and Internet Protocol Television (IPTV) have gained significant popularity in our daily life. However, the quality of these applications cannot be guaranteed in an IP network due to its best-effort delivery. It is therefore crucial to establish an objective model for video quality assessment targeting system design, QoS (Quality of Service) planning and quality monitoring [2], [3].

Objective video quality assessment can be categorized into media-layer models, bitstream-layer models, packet-layer models, parametric models and hybrid models from the viewpoint of the input information. To estimate the perceptual quality of service (QoS) for users, the media-layer models use media signals [4], where characteristics of the video content and decoder strategies such as error concealment are usually taken into account. The bitstream-layer models, on the other hand, perform an analysis on the bitstream without resorting to a complete decoding [5], which can be used in situations where one does not have access to decoded video sequences. The packet-layer models exploit the packet headers to obtain information about the service quality [6], making them well suited for in-service non-intrusive monitoring. The parametric models employ parameters from the network or the application [7], [8]. Parameters from the network may include the packet loss rate and the delay information, while those from the application usually cover the coding bit rate, frame rate, and so on. The hybrid models use a combination of information from the bitstream and the media data, and therefore have an advanced performance as well as combined features of the other models [9].

Since the packet-layer model only utilizes information from packet headers, it is very efficient in quality monitoring due to its low complexity, especially suitable for quality monitoring at network internodes. The other advantage is that the packet-layer model does not need decryption and decoding,

making it favorable when packet payloads are encrypted. In this paper, a packet-layer model is proposed for efficient quality assessment for networked video. Utilizing the limited information which can be provided by packet headers, the frame type and temporal complexity are estimated based on the coding bit-rate. The temporal complexity is incorporated in the proposed model to make it content-adaptive.

The remainder of this paper is organized as follows. The framework for proposed packet-layer video quality assessment is introduced in Section 2. Section 3 discusses the relationship between the coding distortion and the bit-rate. The proposed packet-layer model for video quality assessment is described in Section 4. The experimental results are presented in Section 5. This paper closes with conclusions given in Section 6.

2 Packet-Layer Model for Video Quality Assessment

Packet-layer model for video quality assessment is especially suitable for application scenarios like in-service video quality monitoring and network service planning. It predicts the networked video quality from packet-header information, without resorting to any media-related payload information. Since only the packet header is exploited, the packet-layer model is very useful at network inter-nodes due to its low complexity, where it can monitor thousands of video streams at the same time. The other advantage is that the packet-layer model does not need decryption and video decoding, making it favorable when packet payloads are encrypted.

As an example, Figure 1 shows the structure of a packet in RTP/UDP/IP protocol stacks. In this case, the IP (Internet Protocol) header, the UDP (User Datagram Protocol) header, and the RTP (Real-time Transport Protocol) header can be accessed by a packet-layer model. The length of payload is easily obtained since the UDP length field indicates the length of the UDP header [10], the RTP header and the payload [11]. The marker bit in the RTP header indicates the end of a video frame, and all packets related to one video frame are with the same RTP timestamp. Using this information, the packets can be assembled to frames.

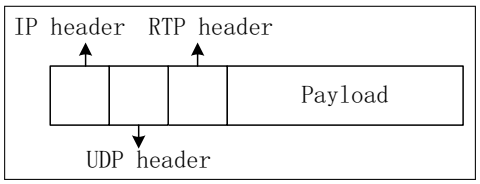


Figure 1: The structure of a packet under RTP/UDP/IP protocol stacks

By analyzing packet headers, the information needed by the parametric model can be obtained and used to estimate the video quality [12]. Apart from the parameters of bit-rate, frame rate, and packet loss rate which are usually employed in parametric models [13], [14], other information can also be employed in a packet-layer model, such as coding parameters (e.g., the frame type, and the bit-rate of each frame), information about the video content characteristics (e.g., the ratio of the bit-rate for coding I frames and P frames), and the detailed positions of lost packets in a video.

The framework of the proposed packet-layer model is shown below in Figure 2. Firstly, after packet header analysis, the bit-rate for coding each frame can be obtained. Then, it is employed to detect the frame type and calculate the ratio of the bit-rate for coding I frames and P frames. This ratio is employed in the proposed model to estimate the temporal complexity which reflects the motion characteristic of the video content. Finally, the coding distortion of networked video is evaluated using the bit-rate information and the estimated temporal complexity.

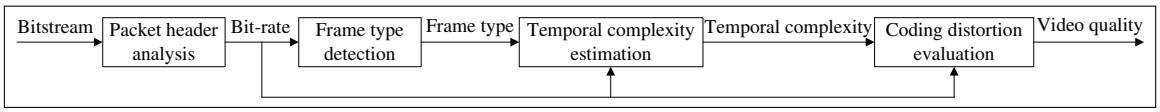


Figure 2: Framework of the proposed packet-layer model

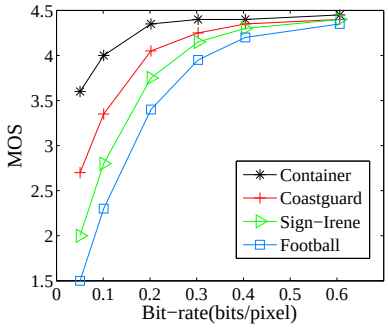


Figure 3: Relationship between the MOS and the bit-rate for each sequence

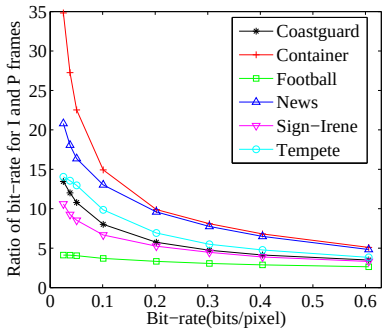


Figure 4: The ratio of the bit-rate for coding I frames and P frames

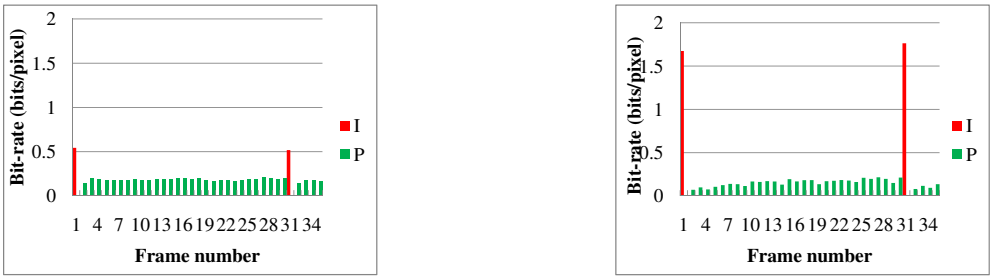
3 Coding Distortion and Bit-Rate

The bit-rate is a key parameter for estimating the coding distortion. It has been well recognized that there is a relationship between the bit-rate and the average Mean Opinion Score (MOS) for different video sequences. Therefore, several functions have been proposed to approximate the relationship to predict the average coding distortion using the bit-rate, such as the computational model proposed in the ITU-T Recommendation G.1070 [15] and its enhancement [16]. Other model forms can also be found, such as the "m-n model" [17], as well as the exponential model [12], [18]. A detailed performance comparison of those models is provided in [19], where superior performance of the enhanced G.1070 model is observed.

Although the average MOS can be predicted using models, the subjective quality of individual videos cannot be well formulated when provided only with the coding bit-rate. Considering the video quality for each sequence, the relationship between the bit-rate and the MOS is shown in Figure 3. It is observed that there are obvious differences in the video quality at a same bit-rate for different sequences. Therefore using the bit-rate only is not suitable for estimating the quality of a certain video service.

It has now been widely acknowledged that content features must be taken into account for an accurate prediction of the perceived video quality [20]. Building on this argument, video clips are classified into three classes according to the subjective movement content (High, Medium and Low movement content), and the model parameters are calculated for each class [18], [19]. However, it is not described how to obtain the information about movement content for each video clip based on objective parameters [18]. Although the average SAD (sum of absolute differences) can be employed in [19] to reflect the motion characteristic of video content, this value is not available for a packet-layer model.

According to Figure 3, it can be seen that the video clip which has a higher motion complexity such as "Football" has a comparatively lower quality at the same bit-rate. Correspondingly, "Container" having a lower motion has a higher quality over the others under the same coding bit-rate. Therefore, the temporal complexity estimated from the packet headers is expected to reflect the motion extent of video clips. How to measure this variable and then establish a packet-layer model based on video content is proposed in the next section.



(a) Football (b) News

Figure 5: Bit-rate histogram of different frame type at 0.202(bits/pixel)

4 Packet-Layer Model Based on Video Content

Figure 4 shows the values of the ratio of the bit-rate for coding I frames and P frames for different video clips. This ratio is always comparatively lower for the "Football" sequence due to its high temporal complexity which results in more bit-rates in coding the P frames. On the other hand, the values of this ratio for the "Container" sequence are consistently larger because of its lower temporal complexity. As the observation, low values of this ratio may correspond to a high temporal complexity. Therefore, the temporal complexity can be roughly estimated using the ratio of the bit-rate for coding I frames and P frames. For a packet-layer model, however, the frame type information is not readily available. Consequently, a frame type detection method based on the bit-rate distribution for each frame is introduced as follows.

4.1 Frame Type Detection

As a general principle, video coding exploits spatial redundancy using intra coding and temporal redundancy using inter coding, where the inter coding modes are usually more efficient in removing redundancy. Accordingly, the bit-rate for coding an I frame is usually much higher than that for a P frame, as shown in Figure 5.

Consequently, a threshold-based method is proposed to detect the frame type using the information about the coding bit-rate. However, Figure 5 also shows that the bit-rate related to a certain frame type varies with the video content. Because of the high motion complexity of "Football", more bit-rates are distributed to P frames. As the result, at a same bit-rate, the values of bit-rate coding I frames of "Football" are lower than the values of "News" which has a lower motion complexity under the same coding bit-rate. Therefore, to make the detection more effective, the threshold for video clips of higher temporal complexity should be lower than the threshold for clips of lower temporal complexity. So fixed thresholds may fail in detecting for different video clips. Dynamic thresholds which are adaptively adjusted [21] are applied in this paper.

4.2 Temporal Complexity Estimation

After frame type detection, the ratio of the bit-rate for coding I frames and P frames can be calculated using the information about the frame type and bit-rate of each frame. This ratio is defined as:

$$r = \frac{R_I}{R_P}, \tag{1}$$

where R_I is the average bit-rate for coding I frames in a certain duration and R_P is the average bit-rate for coding P frames in the same duration. However, it can be seen from Figure 4 that there is obvious

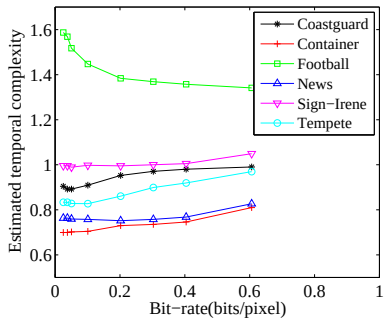


Figure 6: Relationship between estimated temporal complexity and bit-rate

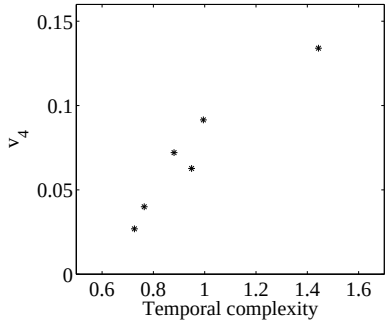


Figure 7: Relationship between v_4 and temporal complexity

differences in the values of r at different bit-rates for each sequence. Due to the fact that the temporal complexity should be evaluated for the sequence with a given related value, r can not be directly applied as the measure of temporal complexity.

Consequently, a mathematical mapping is used to unify the values of r for all the range of bit-rates for each sequence. Firstly, the natural logarithm function acts on the values of r to eliminate the enormous differences due to their distribution of different orders of magnitude. Then, adaptively adjusted factors which are related to the average bit-rate are introduced to make the values generated by the first step for each sequence as near as possible. Both of these two steps are implemented without influencing the relatively ranking of the curves of different sequences. However, to give the temporal complexity a practical significance (a high value corresponds to a higher motion extent sequence), the inverse function is employed at last. So the temporal complexity is formulated as:

$$\sigma_T = \frac{a_1 \cdot \ln(R) + b_1}{\ln(R_I) - \ln(R_P)}, \tag{2}$$

where R is the average bit-rate, and a_1 and b_1 are constants obtained by the experiments.

As shown in Figure 6, the values of estimated temporal complexity calculated by Formula 2 for each sequence are roughly consistent for all bit-rate. And different sequences have different average values which can reflect the motion characteristic of the video content. So the estimated temporal complexity can be employed for the packet-layer model based on video content.

4.3 Proposed Model for Quality Assessment

It is obvious that the average MOS for different video sequences increases as the bit-rate increases and saturates at the maximum MOS, which has been formulated in ITU-T Recommendation G.1070 as:

$$V_q = 1 + v_3 \cdot (1 - \frac{1}{1 + (\frac{R}{v_4})^{v_5}}), \tag{3}$$

where V_q is the video quality, R is the bit-rate, and v_3, v_4, v_5 are empirical parameters. This model can estimate the average video quality for different contents at each bit-rate.

However, video quality strongly depends on the video content. Best values for v_3, v_4 and v_5 are calculated for the video sequences are presented in Table 1. It can be found from Figure 6 and Table 1 that a higher value of v_4 corresponds to a video sequence with higher temporal complexity (e.g., the "Football" sequence). On the contrary, a video sequence whose temporal complexity is low usually has a low value of v_4 (e.g., the "Container" sequence). Figure 7 shows the relationship between v_4 and the temporal complexity, and a linear model approximates this relationship well as follows,

$$v_4 = a_2 \cdot \sigma_T + b_2, \tag{4}$$

where a_2, b_2 are obtained by the experiments, and v_4 is not a constant but a variable varied with σ_T.

Table 1: The values of v_3, v_4 and v_5 for each video sequence

Video sequence	v_3	v_4	v_5
Container	3.546	0.027	1.237
News	3.371	0.040	2.232
Coastguard	3.464	0.063	1.733
Tempete	3.456	0.072	1.687
Sign-Irene	3.546	0.091	1.884
Football	3.481	0.134	2.230

In addition, Table 1 shows that there is relative small difference between the values of v_3 for different sequences, which is the maximum MOS of the sequence. Though there is a relative large difference between the values of v_5 for different sequences, the difference influences the value of V_q slightly. Therefore, v_3 and v_5 are set as constants for all video clips in the proposed model and both of them are obtained by the experiments.

Consequently, submitting Formula 4 into Formula 3, the proposed model is established as:

$$V_q = 1 + v_3 \cdot (1 - \frac{1}{1 + (\frac{R}{a_2 \cdot \sigma_T + b_2})^{v_5}}). \tag{5}$$

Apart from the bit-rate, the temporal complexity, which reflects the motion characteristic of the video content, is considered in this model to make the evaluation more accurate.

5 Experimental Results

The video sequences chosen for experiments covered a wide range of scenes from high motion to low motion events. Specifically, standard video sequences of "Carphone", "Coastguard", "Container", "Football", "Foreman", "Hall_Monitor", "Mother&Daughter", "News", "Paris", "Sign_Irene", "Silent", "Soccer" and "Tempete" were used for performance evaluation. The sequences were all in the Common Intermediate Format (CIF) at 25 frames per second (fps), and encoded using x264 coder [22] with a GOP (Group of Picture) structure of "IPPP" sized of 30. For each sequence, the first 8 seconds were used for evaluation.

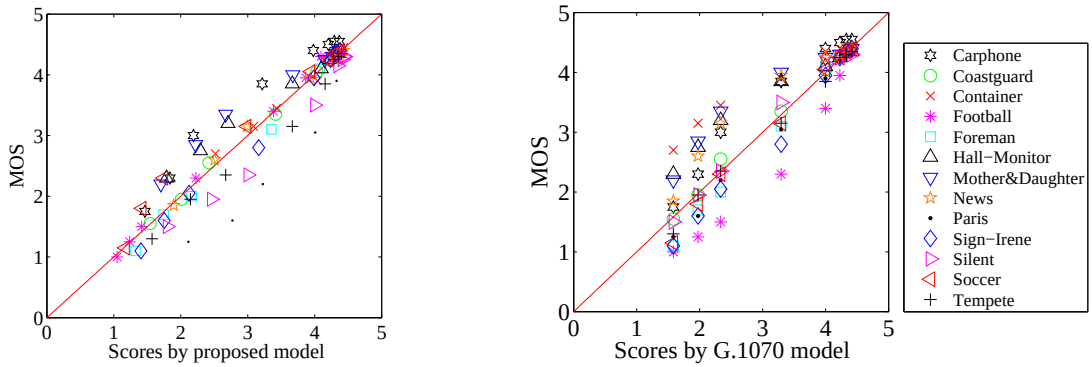
The subjective scores were collected for comparison purposes. The guidelines specified by the Video Quality Experts Group (VQEG) in [23] were followed for the subjective tests. Twenty-five non-expert viewers were involved in these tests, using the Absolute Category Rating (ACR) with a 5-point scale to obtain the MOSs of reconstructed sequences [24], [25].

The parameters were obtained according to the experiments, and their values are shown in Table 2. These parameters were set fixed for all carried experiments. However, if applied to videos generated by the other codecs, they may need to be adjusted.

Table 2: Parameter values

v_3	v_5	a_1	b_1	a_2	b_2
3.477	1.834	-0.334	1.137	0.142	-0.065

Pearson correlation coefficient (PCC) and the root-mean-squared error (RMSE) were used to evaluate the performance of the proposed model. By comparison with the G.1070 model, the proposed model gets an increment about 0.024 in PCC and a decrement about 0.082 in RMSE, as shown in Table 3. The



(a) Proposed model (b) G.1070 model

Figure 8: Scatter plot of MOSs vs objective scores

scatter plots of the objective scores versus the subjective scores are shown in Figure 8, from which the same conclusion can be drawn that using the proposed model the perceived coding distortion can be more accurately measured.

Table 3: Performance comparison of proposed model and G.1070 model

Video quality assessment model	PCC	RMSE
Proposed model	0.9582	0.3250
G.1070 model	0.9338	0.4067

6 Conclusions

A packet-layer model based on characteristics of the video content is proposed in this paper to measure the perceived coding distortion for networked video. Without resorting to the payload information, the temporal complexity is estimated using the ratio of the bit-rate for coding I frames and P frames to reflect the motion characteristic of the video content. Based on analysis of the parameters in the original G.1070 model, the measure of temporal complexity is integrated in the proposed model. Extensive experimental results have demonstrated that the proposed model shows an advanced performance in comparison with the G.1070 model. Further work may include the application of the proposed model to practice by considering both coding distortion and packet loss.

Acknowledgement

This work was supported by the National Science Foundation of China (60902081, 60902052), the Fundamental Research Funds for the Central Universities (72004885), the International Science and Technology Cooperation Program of China (2010DFB10570), and the 111 Project (B08038).

Bibliography

[1] C. Grava, A. Gacs¨˜di, I. Buciu, "A homogeneous algorithm for motion estimation and compensation by using cellular neural networks", *International Journal of Computers Communications & Control*, ISSN 1841-9836, Vol. 5, No. 5, pp.719-726, 2010.

[2] H. R. Wu, K. R. Rao, Eds., Digital video image quality and perceptual coding, *CRC Press*, 2005.

[3] A. Marchand, M. Chetto, "Quality of service scheduling in real-time systems", *International Journal of Computers Communications & Control*, ISSN 1841-9836, Vol. 3, No. 4, pp. 353-365, 2008.

[4] ITU-T Recommendation J.148, "Requirements for an objective perceptual multimedia quality model", 2003.

[5] O. Verscheurei, X. Garcia, "User-oriented QoS in packet video delivery", *IEEE Network*, pp. 12-21, Nov. 1998.

[6] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality", *IP Telephony Workshop*, 2001.

[7] K. Yamagishi, T. Hayashi, "Analysis of psychological factors for quality assessment of interactive multimodal service", *Electronic Imaging 2005*, pp. 130-138, Jan. 2005.

[8] K. Yamagishi, T. Hayashi, "Opinion model using psychological factors for interactive multimodal services", *IEICE Trans. Commun.*, Vol. E89-B, No. 2, pp. 281-288, Feb. 2006.

[9] A. Takahashi, A. Kurashima, H. Yoshino, "Objective assessment methodology for estimating conversational quality in VoIP", *IEEE Trans.on SALP*, Nov. 2006.

[10] RFC 768, UDP, User datagram protocol, 2003.

[11] RFC 3550, RTP, A transport protocol for real-time applications, 2003.

[12] A. Raake, M. Garcia, J. Berger, F. Kling, P. List, J. Johann, C. Heidemann, "T-V-Model: parameter-based prediction of IPTV quality", *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 1149-1152, Mar. 2008.

[13] M. N. Garcia, A. Raake, "Parametric packet-layer video quality model for IPTV", *Proc. Information Sciences Signal Processing and their Applications*, Kuala Lumpur, Malaysia, May 2010.

[14] K. Yamagishi, T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services", *Proc. International Communications Conference*, Beijing, China, May 2008.

[15] ITU-T Recommendation G.1070, "Opinion model for video-telephony applications", Apr. 2007.

[16] J. Joskowicz, J. C. Lopez-Ardao, "Enhancements to the opinion model for video-telephony applications", *Proc. the International Latin American Networking Conference*, Pelotas, Brazil, Sep. 2009.

[17] J. Joskowicz, J. C. Lopez-Ardao, M. A. G. Ortega, C. L. Garcia, "A mathematical model for evaluating the perceptual quality of video", *Proc. International. Workshop on Future Multimedia Networking*, Coimbra, Portugal, June 2009.

[18] H. Koumaras, A. Kourtis, D. Martakos, J. Lauterjung, "Quantified PQoS assessment based on fast estimation of the spatial and temporal activity level", *Multimedia Tools and Applications*, Vol. 34, No. 3, Sep 2007.

[19] J. Joskowicz, J. C. Lopez-Ardao, "A general parametric model for perceptual video quality estimation", *Proc. Communications Quality and Reliability*, Vancouver, BC, June 2010.

[20] M. N. Garcia, A. Raake, P. List, "Towards content-related features for parametric video quality prediction of IPTV services", *Proc. Acoustics, Speech and Signal Processing*, Las Vegas, USA, pp. 757-760, April 2008.

[21] N. Liao, Z. Chen, "A packet-layer video quality assessment model based on spatiotemporal complexity estimation", *Proc. Visual Communications and Image Processing*, Huangshan, China, July 2010.

[22] VideoLAN, X264 CODEC, http://www.videolan.org/x264.html.

[23] VQEG, "Hybrid perceptual/bitstream group TEST PLAN 1.1", http://www.its.bldrdoc.gov/vqeg/, Sep. 2007.

[24] ITU-T, Recommendation P. 910, "Subjective video quality assessment methods for multimedia applications", April 2008.

[25] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", 2002.