A Packet-Layer Model with Content Characteristics for Video Quality Assessment of IPTV

Qian Zhang[1], Lin Ma[2(✉)], Fan Zhang[1], and Long Xu[3]

[1] School of Information and Control Engineering,
Xi'an University of Architecture and Technology, Xi'an, China
[2] Huawei Noah's Ark Lab, Hong Kong, Hong Kong
forest.linma@gmail.com
[3] Key Laboratory of Solar Activity, National Astronomical Observatories,
Chinese Academy of Sciences, Beijing, China

Abstract. Due to the lightweight measurement and no access to the media signal, the packet-layer video quality assessment model is highly preferable and utilized in the non-intrusive and in-service network applications. In this paper, a novel packet-layer model is proposed to monitor the video quality of Internet protocol television (IPTV). Apart from predicting the coding distortion by the compression, the model highlights a novel loss-related scheme to predict the transmission distortion introduced by the packet loss, based on the structural decomposition of the video sequence, the development of temporal sensitivity function (TSF) simulating human visual perception, and the scalable incorporation of content characteristics. Experimental results demonstrate the performance improvement by comparing with existing models on cross-validation of various databases.

Keywords: Video quality assessment · Temporal sensitivity function (TSF) · Packet-layer model · Content characteristics

1 Introduction

With the increasing popularity of communication through IP network in recent decades, it has witnessed an extensive expansion of its applications such as Internet access, Internet protocol television (IPTV), and voice-over-IP (VoIP). Since the IPTV deliveries the television services through internet protocol suite over a packet-switched network, the perceived service quality may be degraded by data compression before transmission, as well as channel distortion during transmission. Therefore, the assessment of quality of service (QoS) or quality of experience (QoE) is highly demanded, in order to monitor the video quality and make the services meet the users' expectation.

The video quality can be predicted based on the compressed bitstream or the video signal itself. Compared with the quality assessment methods performed in the video signal domain, the bitstream-based video quality assessment methods allow the light demand of computational resources, which is highly preferable for non-intrusive and in-service networked video services, such as IPTV. Such bitstream-based quality assessment methods are expected to show good agreements with human visual perception,

Y.-S. Ho et al. (Eds.): PCM 2015, Part II, LNCS 9315, pp. 488–496, 2015.
DOI: 10.1007/978-3-319-24078-7_49

which are believed to provide good QoS or QoE for users. According to the different levels of accesses to the bitstream and availability of information, the bitstream-based quality assessment models can be categorized into three types, which are the parametric model, packet-layer model, and bitstream-layer model [1], respectively. The packet-layer model solely utilizes the packet header information to predict the video quality. It thus reveals more insights of the content characteristics than the parametric models which only use a few general parameters on the sequence level, and involves less computation load than the bitstream-layer model where media related payload information is necessary in addition to the packet header. Therefore, the packet-layer model is extensively investigated with the development of amount of standardization activities. For example, a packet-layer model called P.NAMS [2] has been standardized as a Recommendation in ITU-T SG12 for the quality assessment of IPTV.

Since packet-layer models allow the access to the packet header of video streams, content-dependent characteristics, such as spatial and/or temporal complexities, can be extracted to a certain extent, which can facilitate the performance improvements of the model prediction ability. Garcia et al. [3] proposed to extract the loss-related features, which describe the spatial-extent and duration of the loss by considering "frame-layer" information. Literature in [4] proposed a novel video quality monitoring model by estimating the spatio-temporal complexity and exploiting the interaction between content features. The model in [5] takes into account the video content, using an objective estimation of the spatio-temporal activity. To reveal the visible artifacts and perceived quality, the efficient loss-related parameters are proposed in [6] by estimating the error propagation in both spatial and temporal domain.

In this paper, we propose a quality assessment model with loss-rated features exploiting the content characteristics in a finer level of the video stream. The contributions lie in the following three aspects.

(1) A structural decomposition method is proposed to segment the video sequence in a coarse-to-fine manner.
(2) A temporal sensitivity function (TSF) is proposed to depict the human visual perception on the temporal complexity of the video content.
(3) The proposed loss-related feature incorporates the TSF in a scalable manner, which is combined with coding-related features to formulate the quality prediction model.

The remainder of the paper is presented as follow. The proposed packet-layer quality assessment model is described in Sect. 2. The experimental settings and results are shown in Sect. 3. In Sect. 4, we summarize our model with its advantages.

2 Packet-Layer Video Quality Assessment Model

2.1 Framework

The framework for the proposed packet-layer video quality assessment model is illustrated in Fig. 1. The input of the model is the encoded bitstream and the output is the predicted value indicating the video perceptual quality. By analyzing the packet header with the encrypted payload, the information such as bit-rate, frame-rate, packet loss rate, can be

obtained. Also based on the packet header information, some features, which are closely related with video quality distortion and human perception, can be further computed. Such features are extracted to represent the compression-related and loss-related distortions. Finally, with such extracted features, the compression and loss degradation are estimated simultaneously to predict the perceptual quality of the video, where the parameters of the quality prediction model are obtained from the training samples.

2.2 Structural Decomposition

The sequence-level features such as bit-rate, frame-rate and packet loss rate are commonly used in the packet-layer quality assessment models. However, these features can only provide a general and crude reflection, while quality may vary on different video content even with the same features. Even for one video sequence with same features, the perceptual quality may vary dramatically with the content complexity changes. Thus the content-dependent characteristics incorporated with the sequence structural decomposition should be taken into consideration to improve the model effectiveness.

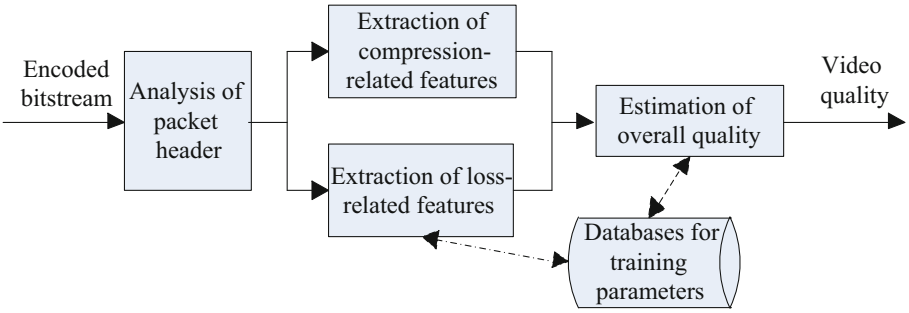

Fig. 1. The framework of the proposed packet-layer video quality assessment model

The video coding technique encodes a sequence with specific configurations before transmission, for example "IPBB" GOP structure with a GOP length of 25 in Fig. 2. The inter-frame coding modes, specifically the "P" and "B" frames, are usually employed to remove temporal redundancies. According to the coding structure, a video sample can be structurally decomposed from the coarse "SEQUENCE" level to the "GOP" level, and then the finest "GROUP" level, as shown in Fig. 2. Generally speaking, the temporal complexity may vary between GOPs, while it may keep more consistent in a same GROUP. The reason is that the GROUP consists of only a few frames, which present the same content with a large probability. In order to structurally decompose the sequence in different levels, the frame type estimation algorithm in [6] is adopted, where the "scene-cut frame" and "non scene-cut I frame" are not differentiated and denoted as I frame for simplicity. As such, the frame type is estimated, based on which the GOP and GROUP structures are identified. Consequently, the GOP-level and GROUP-level features are further computed for the quality assessment model.

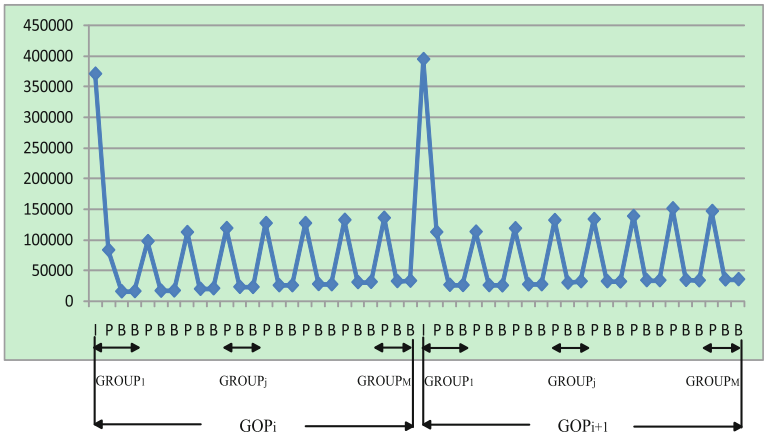

Fig. 2. Frame type estimation and sequence decomposition: *y*-axis is frame bytes and *x*-axis is the estimated frame type. A sequence is decomposed into *N* GOPs, and one GOP is further decomposed into *M* GROUPs, which starts from the I/P frame till the next I/P frame.

2.3 Temporal Sensitivity Function

As mentioned before, the perceived video quality is not only predicted by the extracted features but also dependent on the video content. The temporal complexity reflects the acuteness of temporal changes of a video sequence, which is usually high for sequences with high motions. Generally, the motion vectors can be employed to measure the motion activities [8]. However, it is not available in the packet-layer model where the payload is encrypted. Yang et al. [9] proposed a measurement of temporal complexity using the ratio of number of bit encoding P and I frames in "IPPP" GOP structure. Inspired by this, we propose a temporal complexity measurement in the GOP and GROUP level, which are calculated in (1) and (2) respectively:

$$\alpha_{GOP}^{i} = R_{PB}/R_I \tag{1}$$

$$\alpha_{GROUP}^{j} = \begin{cases} R_{PB}/R_I & if\ GRUOP_j\ contains\ I\ frame \\ R_B/R_P & otherwise \end{cases} \tag{2}$$

where α_{GOP}^{i} and α_{GROUP}^{j} are the measurement of temporal complexity in the i-th GOP and j-th GROUP respectively. R_I is the number of bits coding I frame, and R_{PB} is the average number of bits coding each frame (including the P and B frames) in the same decomposition level. R_P is the number of bits coding each P frame, and R_B is average number of bits coding each B frame in the same GROUP. If α tends to be large, each B/P frame tends to consume more bits for compression, which means that the motion estimation can hardly remove the temporal redundancies. Therefore, a high temporal complexity with fast motion is expected in the video sequence. For small α value, a low temporal complexity with smooth motion is expected.

The work in [7] implies that visual content with different motion strengthes will influence the evaluation of perceptual quality in human vision system. Human eyes will pay more attention to the distortions with median motion acuteness, whereas pay less attention to the distortions with high or low motion acuteness. Based on this psychophysical study, we propose a temporal sensitivity function (TSF) in (3) to simulate the human vision perception with temporal complexity, which forms a normal distribution:

$$TSF(\alpha) = N\left(\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right) \tag{3}$$

where $\mu(0 < \mu < 1)$ and σ^2 are the expectation and variance receptively, which can be obtained by offline training from the collected samples.

2.4 Loss-Related Feature

An efficient loss-related method named ALAE [6] is proposed by estimating the error propagation in spatio-temporal domain. ALAE is the averaged loss artifact extension (LAE), which is calculated for each frame as the sum of initial artifact caused by the loss in the current frame and propagated artifact caused by the loss in reference frames. By considering the incorporation of sequence structural decomposition in Sect. 2.2 and human vision perception reflected by temporal complexity in Sect. 2.3, an improved parameter named as averaged loss artifact extension with human vision perception (ALAEhvp) is proposed to predict the loss-related distortion:

$$ALAEhvp = \frac{\sum_{i=1}^{N} TSF\left(\alpha_{GOP}^{i}\right) \times \frac{\sum_{j=1}^{M} TSF\left(\alpha_{GROUP}^{j}\right) \times LAE_{sum}^{j}}{\sum_{j=1}^{M} TSF\left(\alpha_{GROUP}^{j}\right)}}{\sum_{i=1}^{N} TSF\left(\alpha_{GOP}^{i}\right) * (f * \sqrt{s})} \tag{4}$$

where $LAE_{sum}{}^{j}$ is the sum of LAE in the j-th GROUP. M is number of GROUPs in i-th GOP. N is number of GOPs in the sequence. The LAE is weighted by temporal complexity which reflects the importance of human perception and is accumulated at GROUP, GOP and SEQUENCE levels. The quality value is then averaged by the number of frames f and a function of number of slices per frame S. TSF is employed to calculate the human visual perception to the video temporal complexity at the GROUP and GOP levels.

2.5 Video Perceptual Quality Prediction

Considering a video sequence may be corrupted simultaneously by compression and transmission distortion, the overall quality prediction model is capable of predicting the video quality combining the coding artifacts with loss artifacts, which can be obtained by a logistic function [6]:

$$V_q^N = \frac{1}{1 + a * Br^b * ALAEhvp^c} \tag{5}$$

In (5), Br is the bit-rate used to model coding artifacts and $ALAEhvp$ is used to model slicing channel artifacts. a,b,c are the constants obtained from curve-fitting using a least square fitting method through training databases. V_q^N is the normalized mean opinion score (NMOS) within [0,1], which is transformed from MOS by the linear mapping. It should be noticed that the overall model can be reduced to predict the compression distortion only by setting $c = 0$.

Table 1. The database configuration: Df: display format (p-progressive; i-interlace); Br: bitrate (Mbps); Fr: frame rate (fps); Ns: no. of the slices per frame.

	Training	Validation
Df	1080p/i,720p,576i	1080p/i,720p,576i,480i
Br	15,9,7,6,2.5,2,1,0.5	15,7,6,5,4,3,3.5,2.5,2,1.5,0.5
Fr	50,30,25	60,50,30,25
Ns	1, 18, 68	1, 15, 18, 34, 45, 68

3 Experimental Results

3.1 Experimental Setting

In order to develop the model and conduct the experimental comparison, 5 training databases are built for training the parameters μ and σ^2 in Eq. (3) and determining the coefficients a,b,c in Eq. (5), and 6 validation databases for testing the model performance. Each database contains 8 video contents with 10 second duration of high dimension (HD) or standard dimension (SD). The hypothetical reference circuits (HRCs) are encoded by H. 264 with several different GOP structures of fixed or adaptive GOP length. The packet-loss-concealment (PLC) mode is slicing and packet-loss-duration is random or burst. The configurations of training and validation databases are summarized in Table 1. The testing environment is conformed to ITU-R BT.500 [10] and the subjective test is performed using the absolute category rating with hidden reference method in ITU-T Rec. P.910 [11]. The MOS value per HRC is the averaged rating values from 24 subjects. The subjective quality of the video content is categorized into 5 scale: 1 (very annoying), 2 (annoying), 3 (slightly annoying), 4 (perceptible but not annoying), 5 (imperceptible). The subjects are required to watch the sequence and provide their opinions ranging from 1 to 5.

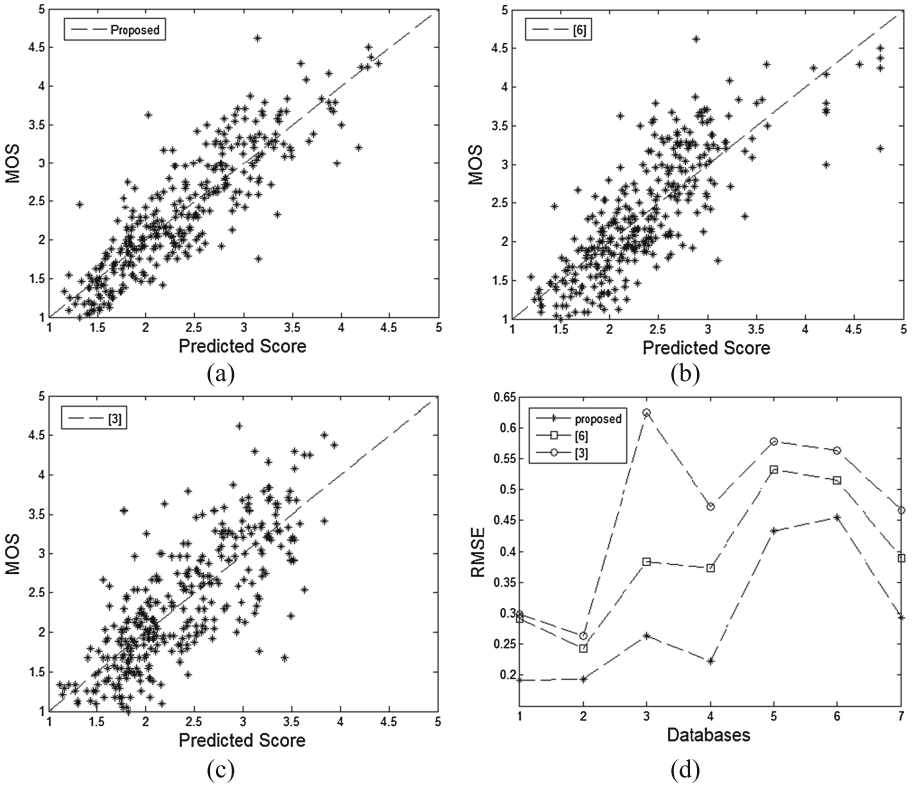

Fig. 3. Experimental results and comparisons

3.2 Experimental Results

To demonstrate the performance of the proposed quality assessment model, we carry out the experiments and make the comparisons with two related models in [6, 3]. The results are shown in Fig. 3. Figure 3(a)–(c) illustrate the predicted quality scores versus MOS values from the subjective rating, by the proposed model, the model in [6], and the model in [3], respectively. It can be observed that the data points are more closely distributed around diagonal line in the proposed model (Fig. 3(a)) than those in [6] (Fig. 3(b)) and in [3] (Fig. 3(c)), which highlights the superiority of the proposed method to the other two models. Figure 3(d) illustrates the root mean square error (RMSE) between the predicted and perceived quality using our model, the model in [6], and the model in [3]. For evaluation, the smaller the RMSE value, the better is the performance of the quality metric. The RMSE value generated by our model is outperformed by other models in all the databases from index 1-6, and clearly better in the mean value in index 7, which demonstrates its good performance. From the experimental results, we believe that our model outperforms the model in [3], because of the introduction of content-related characteristics, which are not considered in [3]. Furthermore, the outper-

formance over model in [6] lies in the fact that in the proposed model, the temporal complexity is used to reflect the human vision perception by the TSF and incorporated into assessment model based on structural decomposition, which is not taken into account in [6]. It also should be noticed that the content characteristics, the GOP structure, bit-rate, frame-rate, number of the slices per frame, and frame-type are considered in the calculation of model parameters. The trained one set of coefficients is sufficiently used in other 6 databases for cross-validation, which shows the generalization of the proposed mode as well.

4 Conclusion

In this paper, a packet-layer model for video quality assessment is proposed in non-intrusive and in-service network applications, such as IPTV. Apart from predicting the compression distortion, the model mainly focuses on the development of a loss-related feature which is capable of predicting the slicing type distortion during video transmission. The novelty of the feature extraction is the incorporation of human vision perception with content characteristics. The contributions lie in the following three aspects: (1) based on the essential principle of video coding technique, a video stream is structurally decomposed from coarse "SEQUENCE" to "GOP", and finest "GROUP" level, which makes the content characteristics of frames within the same level exhibit highly consistency; (2) a TSF of the video temporal complexity is proposed, which is then used to reflect the human vision perception; (3) the human vision perception is incorporated into the loss-related features in a scalable manner. The performance comparison with related existing methods in the experimental results demonstrates the superiority of the proposed model.

Acknowledgement. The work described in this paper was partially supported by Young Scientist Foundation QN1304, and Talent Technology Foundation RC1349 of Xi'an University of Architecture and Technology, and Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2014JM2-6127), and National Natural Science Foundation of China under Grant 61202242.

References

1. Yang, F., Wan, S.: Bitstream-based quality assessment for networked video - a review. IEEE Commun. Mag. **50**(11), 203–209 (2012)
2. P.NAMS (Parametric non-intrusive assessment of audiovisual media streaming quality). Recommendation in ITU-T SG12. http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=6441
3. Garcia, M.N., Raake, A.: Frame-layer packet-based parametric video quality model or encrypted video in IPTV services. In: International Workshop on Quality of Multimedia Experience (QoMEX) (2011)
4. Liao, N., Chen, Z.: A packet-layer video quality assessment model with spatiotemporal complexity estimation. EURASIP J. Image Video Process. **5**(5), 1–13 (2011)
5. Joskowicz, J. Ardao, J.C.L.: A general parametric model for perceptual video quality estimation. In: Proceedings of Communication Quality and Reliability, Vancouver, June 2010

6. Zhang, Q., Zhang, F., Ma, L.: Packet-layer model for quality assessment of encrypted video in IPTV services. In: APSIPA Annual Summit and Conference (2013)
7. Stocker, A.A., Simoncelli, E.P.: Noise characteristics and prior expectations in human visual speed perception. Nat. Neurosci. **9**, 578–585 (2006)
8. Xu, L., Ma, L., Ngan, K.N., Lin, W., Weng, Y.: Visual quality metric for perceptual video coding. In: IEEE Visual Communications and Image Processing (2013)
9. Yang, F.Z., Song, J.R., Wan, S., Wu, H.R.: Content adaptive packet layer model for quality assessment of networked video services. IEEE J. Sel. Topics Signal Process. **6**(6), 672–683 (2012)
10. ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunications Union, Technical report (2000)
11. ITU-T Recommendation P. 910, Subjective video quality assessment methods for multimedia applications. Available Online: http://www.videoclarity.com/PDF/