EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

A packet-layer video quality assessment model with spatiotemporal complexity estimation

Ning Liao* and Zhibo Chen

Abstract

A packet-layer video quality assessment (VQA) model is a lightweight model that predicts the video quality impacted by network conditions and coding configuration for application scenarios such as video system planning and in-service video quality monitoring. It is under standardization in ITU-T Study Group (SG) 12. In this article, we first differentiate the requirements for VQA model from the two application scenarios, and state the argument that the dataset for evaluating the quality monitoring model should be more challenging than that for system planning model. Correspondingly, different criteria and approaches are used for constructing the test datasets, for system planning (dataset-1) and for video quality monitoring (dataset-2), respectively. Further, we propose a novel video quality monitoring model by estimating the spatiotemporal complexity of video content. The model takes into account the interactions among content features, the error concealment effectiveness, and error propagation effects. Experiment results demonstrate that the proposed model achieves robust performance improvement compared with the existing peer VQA metrics on both dataset-1 and dataset-2. It is noted that on the more challenging dataset-2 for video quality monitoring, we obtain a large increase in Pearson correlation from 0.75 to 0.92 and a decrease in the modified RMSE from 0.41 to 0.19.

Keywords: video quality assessment, quality of experience, packet-layer model, spatiotemporal complexity estimation

1. Introduction

With the development of video service delivery over IP networks, there is a growing interest in low-complexity no-reference video quality assessment (VQA) models for measuring the impact of transmission losses on the perceived video quality. No-reference VQA model generally uses only the received video with compression and transmission impairment as model input to estimate the video quality. No-reference model fits better with the real-world situation where customers usually watch IPTV or streaming video without the original video as reference.

In ITU-T Study Group (SG) 12, there is a recent study [1] on the no-reference objective VQA models (e.g., P. NAMS [2], G. Opinion Model for Video Streaming (OMVS), P.NBAMS [3]) considering impairment caused by both transmission and video compression. In literatures, depending on the inputs, the no-reference models can be classified as packet-layer model, bitstream-level model, media-layer model, and hybrid model, as shown in Figure 1.

A media-layer model employs with pixel signal. Thus, it can easily obtain content-dependent features that influence video quality, such as texture-masking effects and motion-masking effects. However, a media-layer model usually needs special solutions (e.g., [4]) for locating the impaired parts in the distorted video because of the lack of information on packet loss.

A packet-layer model (e.g., P.NAMS) utilizes various packet headers (e.g., RTP header, TS header), network parameters (e.g., packet loss rate (PLR), delay), and codec configuration information as input to the model. Obviously, this type of model can roughly locate the impaired parts by analyzing the packet headers. However, how to take the content-dependent features into account is a big challenge to this model.

A bitstream-level model (e.g., P.NBAMS, [5]) uses the compressed video bitstream in addition to the packet headers as input. Thus, it is not only aware of the location

* Correspondence: ning.liao@technicolor.com
Media Processing Laboratory, Technicolor Research & Innovation, Beijing, China

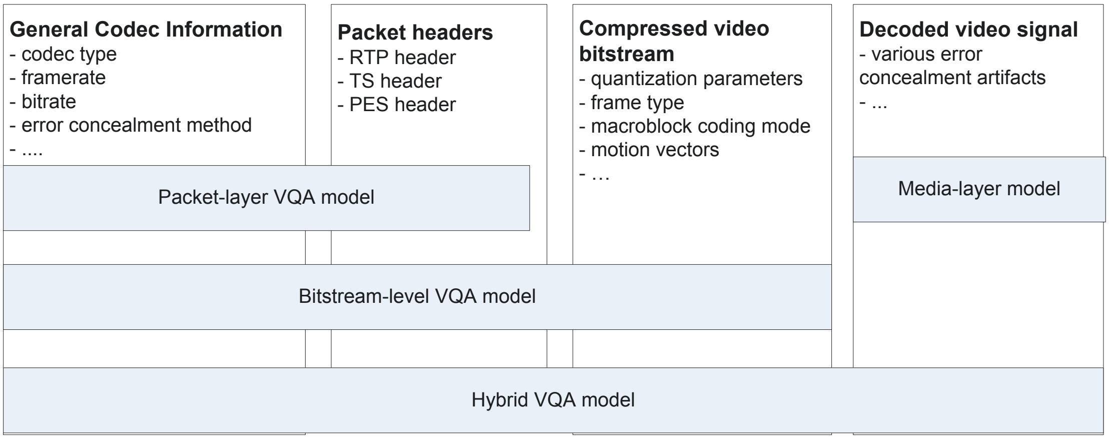


Figure 1 Scope of the four types of VQA models. The columns are four types of input information to the models.

of the loss-impaired parts of video, but also has access to video-content feature and the detailed encoding parameters by parsing the video bitstream. It is supposed to be more accurate than a packet-layer model at a cost of slightly higher computational complexity. However, in the case that video bitstream is encrypted, only packet-layer model works.

Hybrid model uses the pixel signal in addition to the bitstream and the packet headers to further improve video quality prediction accuracy. Because the various error concealment (EC) artifacts become available only after decoding video bitstream into pixel signal, in principle it can provide the most accurate quality prediction performance. However, it has much higher computational complexity.

The packet-layer model, which primarily estimates the video quality impairment caused by unreliable transmission, is studied in this article.

Two use cases of packet-layer VQA models have been identified in ITU-T SG12/Q14: video system planning and in-service video quality monitoring.

As a video system planning tool, parametric packet-layer model can help to determine the proper video encoder parameters and network quality of service (QoS) parameters. This can avoid over-engineering the applications, terminals, and networks while guaranteeing user's satisfactory QoE. ITU-T G.OMVS and G.1070 [6] for videophone service are the examples of the video system planning model.

For video quality monitoring application, usually operators or service providers need to ensure video quality service level agreement by monitoring and diagnosing video quality degradation caused by network issues. Since packet-layer model is computationally lightweight, it can be deployed in large scale along the media service chain. The video quality model of ITU-T standard P.NAMS (Non-intrusive parametric model for the Assessment of performance of Multimedia Streaming) is specifically designed for this purpose.

In general, two approaches can be followed in packet-layer modeling. One is the parameter-based modeling approach [6-9] and another is the loss-distortion chain-based modeling approach [5]. The parameter-based approach estimates perceptual quality by extracting the parameters of a specific application (e.g., coding bitrate, frame rate) and transmission packet loss, then building a relationship between the parameters and the overall video quality. Obviously, the parametric packet-layer model is in nature consistent with the requirement of system planning. However, it predicts the average video quality over different video contents. The coefficient table of this model needs to change with the codec type and configuration, the EC strategy of a decoder, the display resolution, and the video content types. Noticeably, the models in [6,8,9] were claimed to achieve a very high Pearson correlation above 0.95, and the RMSE lower than 0.3 on the 5-point rating scale or 7 on the 0-100 rating scale, even if the video content features were not considered in the models. This motivated us to verify the results and look into the ways of setting up training and evaluation dataset on which the model performance directly depends.

Loss-distortion chain-based approach [5] has the merit of accounting in error propagation, content features, and EC effectiveness. Since iteration process is generally involved in, it is suitable for quality monitoring, not for system planning model. Keeping low computational complexity, which is very important to in-service monitoring, is one challenge for this approach. Another challenge is to estimate the video content and compression information at packet layer. Our proposed model follows this approach and deals with the challenges.

The main contributions of this article are in two aspects. First, we differentiate the requirements for packet-layer model from two application scenarios: video system planning and video quality monitoring. We design the respective criteria and methods to select the processed video sequences (PVSs) for subjective evaluation when setting up the subjective mean opinion score (MOS) database. This helps us to explain why the abovementioned parametric packet-layer models had a high performance even if the video content feature was not taken into consideration. Furthermore, we state the argument that the dataset for evaluating the video quality monitoring model should be more challenging than that for video system planning model.

Second, we propose a novel quality monitoring model, which has low complexity and fully utilizes the video spatiotemporal complexity estimation at packet layer. In contrast to the parametric packet-layer models, it takes into consideration the interaction among video content features and EC effect and error propagation effect, thus improves estimate accuracy.

The rest of the article is organized as follows. In Section 2, we review several literatures that motivated this study. The novelty of this study is then discussed. In Section 3, two different criteria and methods are used to set up respective datasets for monitoring and planning scenarios. In Section 4, the proposed VQA model is described. Experimental results are discussed in Section 5. Conclusions and future work are discussed in Section 6.

2. Related work

The recent studies [10-13] are somehow related to the idea of our proposed model. In [10,11], the contributing factors to the visibility of artifacts caused by lost packet(s) were studied; video quality metrics based on the visibility of packet loss were developed in [12,13].

The factors to the visibility of a single packet loss were studied in [10] for MPEG-2 compressed video. The top three most important factors were the magnitude of overall motion which is the average across all macroblocks (MBs) initially affected by loss, the type (I, B, or P) of the frame (FRAMETYPE) in which packet loss occurred, and the initial MSE (IMSE) of the error-concealed pixels. Further, the visibility of multiple packet losses in H.264 video was studied in [11]. Again, the IMSE and the FRAMETYPE are identified as the most important factors to the visibility of losses. Besides, it was shown that the IMSE is very different because of the different concealment strategies [11]. It can be seen that the accurate detection of the initial visible artifacts (IVA) and the error propagation effects are two important aspects to be considered in a packet-layer VQA model. Furthermore, the different EC effects should be considered when estimating the annoyance level of IVA.

Yamada et al. [12] developed a no-reference hybrid video quality metric based on the count of the MBs for which the EC algorithm of a decoder is identified as ineffective. Classifying lost MBs based on the error-concealment effectiveness can be essentially regarded as an operation to classify the visibility of the artifacts caused by packet loss(s). Suresh [13] reported that the simple metric of mean time between visible artifacts has an average correlation of 0.94 with subjective video quality.

There are two major novel points in our proposed model. First, the IVA of a frame suffering from packet loss and EC is estimated based on the EC effectiveness. Unlike [12], the EC effectiveness is determined based on the spatiotemporal complexity estimation with packet-layer information; and the different EC effects are considered. Second, the IVA is incorporated into an error propagation model to predict the overall video quality. The estimate of spatiotemporal complexity is employed to modulate the propagation of the IVA in the error propagation model. The performance gain resulting from the spatiotemporal complexity-based IVA assessment and from using the error propagation model is analyzed in the experiment section.

3. subjective dataset and analysis

As described above, the packet-layer video QoE assessment model has two typical application scenarios, video system planning and in-service video quality monitoring, each of which has different requirements. The video system planning model is for network QoS parameter planning and video coding parameter planning, given a target video quality. It predicts average perceptual quality degradation, ignoring the impact of different distortion and content types on the perceived quality. Therefore, it should predict well the quality of the loss-affected sequences with large occurrence probability. Whereas, the VQA model for monitoring purpose is expected to give quality degradation alarm with high accuracy and should be able to estimate as accurate as possible the quality of each specific video sequence distorted by packet losses. Correspondingly, the respective subjective dataset for training and evaluating the planning model and the monitoring model should be built differently. Further analysis of the PVSs in Sections 3.3 and 3.4 illustrates that the different EC effects and the different error propagation effects are two of the most important factors to the perceptual quality of packet-loss distorted videos.

There are mutual influences between the perception of coding artifacts and that of transmission artifacts especially at low coding bitrate [14]. In our subjective database, visible coding artifact is not considered by setting the quantization parameter (QP) to a certain smaller value. Only the video quality degradation cause by transmission impairments is discussed in this article.

3.1 Subjective test

Video QoE is both application-oriented and user-oriented assessments [15]. Viewer's individual interests, quality expectation, and service experience are among the contributing factors to the perceived quality. To compensate the subjective variance of these factors, usually MOS averaged over a number of viewers (called subjects hereafter) is used as the quality indication of a video sequence. Moreover, to minimize the variance of subjects' opinion caused by these factors, subjective test should be conducted under well-controlled environment; subjects should be well instructed about the task and video application scenario, which influences the subjects' expectation to video quality.

The absolute category rating with hidden reference method specified in ITU-P.910 [16] is adopted in our experiment. It is a single stimulus method where a processed video is present alone. The five scales shown in Figure 2 are used for evaluating the video quality. Observers are instructed to focus on watching video program instead of scrutinizing visual artifacts. Before the subjective test, observers are required to watch 20 training sequences that evenly cover the five scales, and to write down their understanding of the verbal scales in their own words. Interestingly, the most of the description of the five scales are heavily related to video content, not merely related to the amount of noticeable artifacts as described in [17]. The descriptions can be summarized as follows:

- *Imperceptible*: "no artifact (or problematic area) can be perceived during the whole video display period".

-*Perceptible but not annoying*: "artifact can be perceived occasionally, but it does not influence the interested content, or it appears in the background for an instant moment".

- *Slightly annoying*: "the noticeable artifact appearing in the region of interest (ROI) is identified, or noticeable artifacts are detected for several instant moments even if they do not appear in the ROI".

- *Annoying*: "noticeable artifact appears in ROI for several times or many noticeable artifacts are detected and last for a long time".

- *Very annoying*: "video content cannot be understood well due to artifacts and the artifacts spread all over the sequence".

Twenty-five non-expert observers are asked to rate the quality of the selected 177 PVSs of 10 s. The scores given by these subjects are processed to discard subjects who are suspected to have voted randomly. Then for each PVS, a subjective MOS and a 95% confidence interval (CI) are computed using the scores of the valid subjects. As shown in Figure 3, for PVSs of middle quality, the subjectivity variation is higher; for sequences of very good or very bad quality, the subjects tend to reach a more consistent opinion with high probability. This observation is similar to the previous report in [14]. Since the subjective MOS itself has statistical uncertainty because of the abovementioned subjective factors, it is reasonable to allow certain prediction error (e.g., less than CI_{95}) when evaluating the prediction accuracy of an objective model. Therefore, the modified RMSE [18] described later in Equation 8 is used in our experiment.

3.2 Select PVSs for dataset

Six CIF format video contents, which cover a wide range of spatial complexity (SC) index and temporal complexity (TC) index [19], are used as original sequences, namely *Foreman, Hall, Mobile, Mother, News*, and *Paris*. The six sequences are encoded using H.264 encoder with two

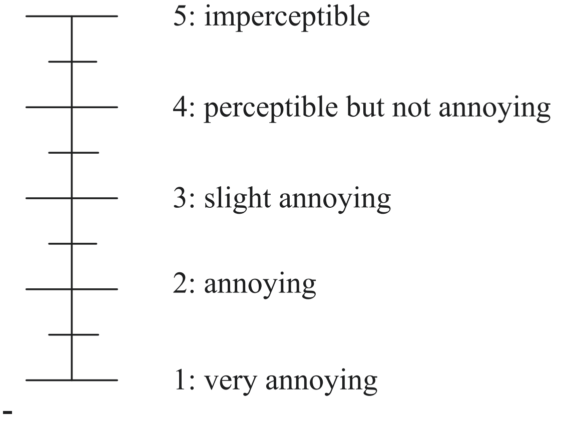


Figure 2 Five point impairment scales of perceptual video quality.

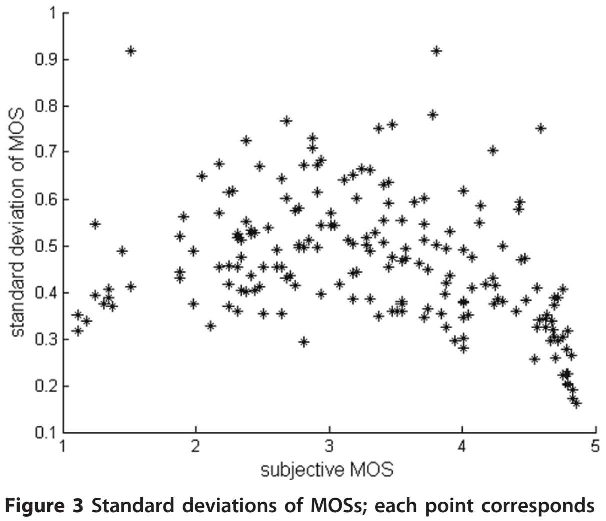


Figure 3 Standard deviations of MOSs; each point corresponds to the standard deviation of the MOS of a PVS.

sequence structures, namely, IBBPBB and IPPP. Group of picture (GOP) size is 15 frames. A proper fixed QP is used to prevent the compressed video from visible coding artifacts. Each row of MBs is encoded as an individual slice, and one slice is encapsulated into an RTP packet. To simulate transmission error, the loss patterns generated at five PLRs (0.1, 0.4, 1, 3, and 5%) in [17] are used. For each nominal PLR, 30 channel realizations are generated by starting to read the error pattern file at a random point. Thus, for each original sequence, there are 150 realizations of packet loss corrupted sequences. Before subjective evaluation test, we must choose some typical PVSs from the large numbers of realizations.

Owing to the different requirements of planning and monitoring scenarios, we choose the PVSs for subjective test according to different criteria:

1. For each video content, select the PVSs that are representatives of the dominant MOS-PLR distribution as done in [17];
2. For each video content, select the PVSs that cover the MOS-PLR distribution widely by including the PVSs of the best and the poorest quality at a given PLR level, in addition to those representing the dominant MOS-PLR distribution.

Actually, when we select the PVSs for the subjective test, the subjective MOSs of the abovementioned 150 sequences is not available before subjective test. The objective measurement PSNR is used as substitute of MOS in the initial selection of PVSs; then the PVSs selected in the initial round are watched and adjusted if necessary to make sure that the subjective qualities of the selected PVSs satisfy the above criteria. The PVSs chosen by criteria-1 and criteria-2 are collectively named as dataset-1 and dataset-2, respectively. Figure 4 shows the PLR-MOS distribution and PSNR-MOS distribution of dataset-1 and dataset-2. The PLR here is calculated as the ratio of actually lost packets to the total transmitted packets for a PVS. It can be seen that the PVSs in dataset-2 present much more diverse relationship between PLR and subjective video quality than those in dataset-1. Because the scales of “annoying” and “very annoying” are equally unacceptable in real-world applications, we selected sequences mostly of the MOSs ranging from 2 to 5, as shown in Figure 4a,b. It is noted that, in subjective test, one sequence with score one point for each video content is included in each test session to balance the range of rating scales, although they are not included in the datasets as drawn in Figure 4.

In Figure 4c, the PLR-PSNR distribution for all the six video contents spreads away from each other, whereas in Figure 4a the PLR-MOS distributions for the mostly video contents are mixed together. This phenomenon partially illustrates that the PSNR is not a good objective measurement of video quality because it fails to take into consideration the impact of video content feature on human perception of video quality.

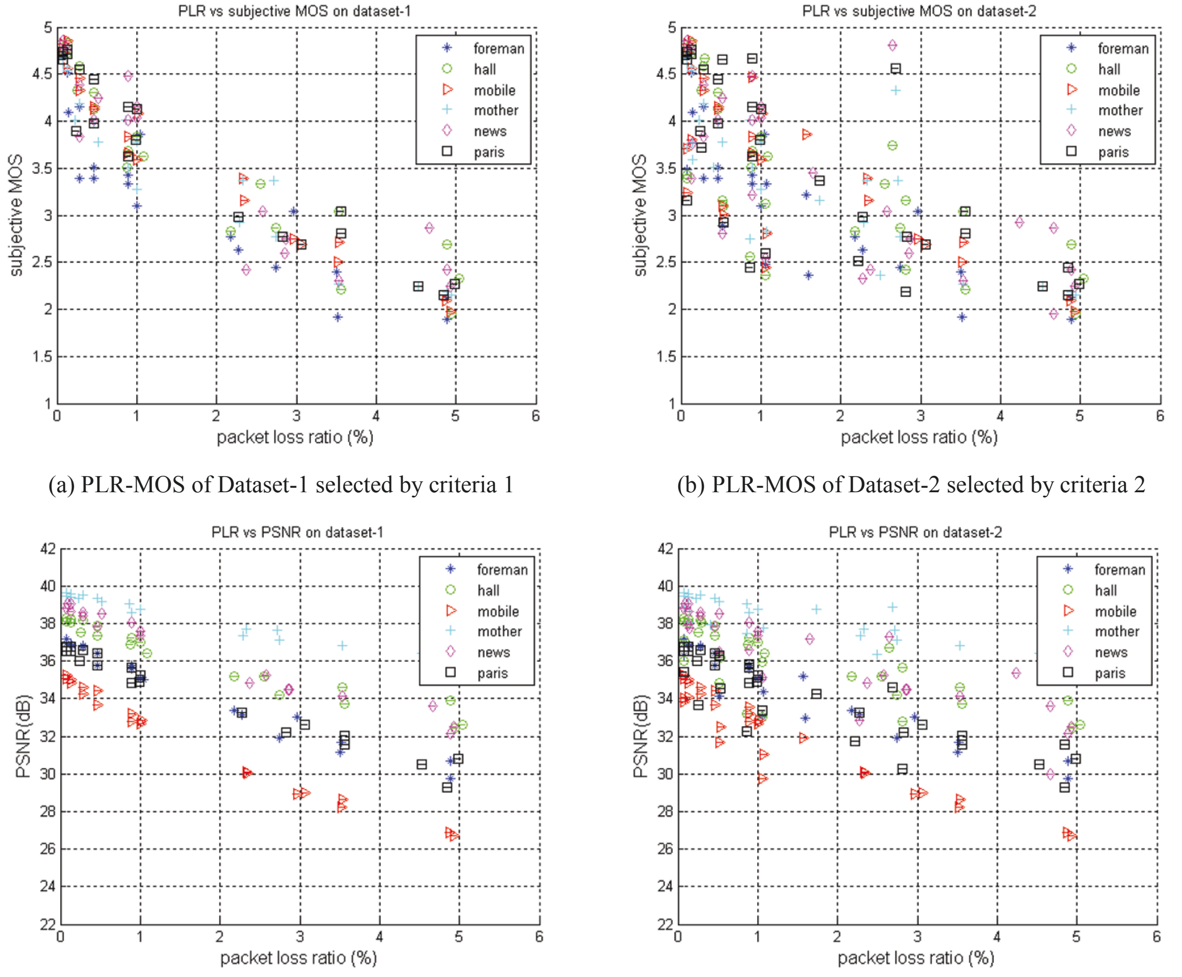
Figure 4b shows that PVSs present very different perceptual qualities in dataset-2 even under the same PLR. Taking the PLR of 0.86% for an example, the MOSs vary from Grade 2 to Grade 4. PLR treats all lost data as equal important to perceived quality, ignoring the content and compression’s influence on perceived quality. It may be an effective feature on dataset-1 as shown in Figure 4a, but is not an effective feature on dataset-2 for quality monitoring applications.

Unlike [6,8,9], our proposed objective model targets at video quality monitoring application. The objective model for monitoring purpose should be able to estimate as accurately as possible the video quality of each specific sequence distorted by packet loss. Correspondingly, the dataset for evaluating the model performance should be more challenging than that for planning model, i.e., the proposed model should work well not only on dataset-1 but also on dataset-2.

3.3 Impact of EC

Both the duration and the annoyance level of the visible artifacts contribute to the perceived video quality degradation. The annoyance level of artifacts produced by packet loss depends heavily on the EC scheme of a decoder. The goal of EC is to estimate the missing MBs in a compressed video bitstream with packet losses, in order to provide a minimum degree of perceptual quality degradation. EC methods that have been developed roughly fall into two categories: spatial EC approach and temporal EC approach. In the spatial EC class, spatial correlation between local pixels is exploited; missing MBs are recovered by interpolation from neighbor pixels. In the temporal EC class, both the coherence of motion field and the spatial smoothness of pixels along edges cross block boundary are exploited to estimate motion vector (MV) of a lost MB. In H.264 JM reference decoder, spatial approach is applied to conceal lost MBs of Intra-coded frame (I-frame) using bilinear interpolation technique; temporal approach is applied to conceal lost MBs for inter-predicted frame (P-frame, B-frame) by estimating MV of the lost MB based on the neighbor MBs’ MVs. Minimum boundary discontinuity criterion is used to select the best MV estimate.

Visible artifacts produced by spatial EC scheme and by temporal EC scheme are very different. In general, spatial EC approach produces blurred estimates of the lost MB as shown in Figure 5a, while the temporal EC approach produces edge artifacts as shown in Figure 5b, if the guessed MV is not accurate. The effectiveness of spatial EC scheme is significantly affected by SC of the frame



(a) PLR-MOS of Dataset-1 selected by criteria 1

(b) PLR-MOS of Dataset-2 selected by criteria 2

(c) PLR-PSNR on Dataset-1 selected by criteria 1

(d) PLR-PSNR on Dataset-2 selected by criteria 2

Figure 4 The processed sequences selected by criteria-1 and criteria-2.

with loss, while that of the temporal EC scheme is significantly affected by motion complexity around the lost area. In Figure 5c, although the fourth row of MBs is lost, almost no visual quality degradation can be perceived because of the stationary nature of the lost content. Whereas, in Figure 5e, slightly noticeable artifacts appear at the area near the mother's hand, because of inconsistent motion of the lost MBs and its neighbor MBs. In Figure 5d, the second row of MBs is lost, but resulting in hardly noticeable artifacts. This is because the lost content is of smooth texture.

3.4 Impact of error propagation

The duration of visible artifact depends on the error propagation effects resulting from the inter-frame prediction technique used in video compression. For the same encoder configuration and channel conditions, Figure 6 shows that the error propagation effects vary significantly depending on different video contents, in particular, on the SC and the TC of the video content. For example, the 93th frame, in which four packets are lost, is a P-frame. Because the head moves largely in the ensuing frames of sequence *foreman*, the error in the P-frame is propagated up to the 120th frame, which corresponds to about 1 s. Even if there is a correctly received I frame at the 105th frame, the error is still propagated to the 120th frame because of large motion, two reference frames, and open GOP structure. In contrast, for sequence *hall* and *mother* having small motion, propagated artifacts are almost invisible.

In general, an I-frame packet loss results in artifact duration of GOP length, or even longer if open GOP structure is used in compression configuration. The more intra-coded MBs exist in inter-coded frames, the more easily the video quality recovers from error, and the shorter the artifact duration is. In general, the

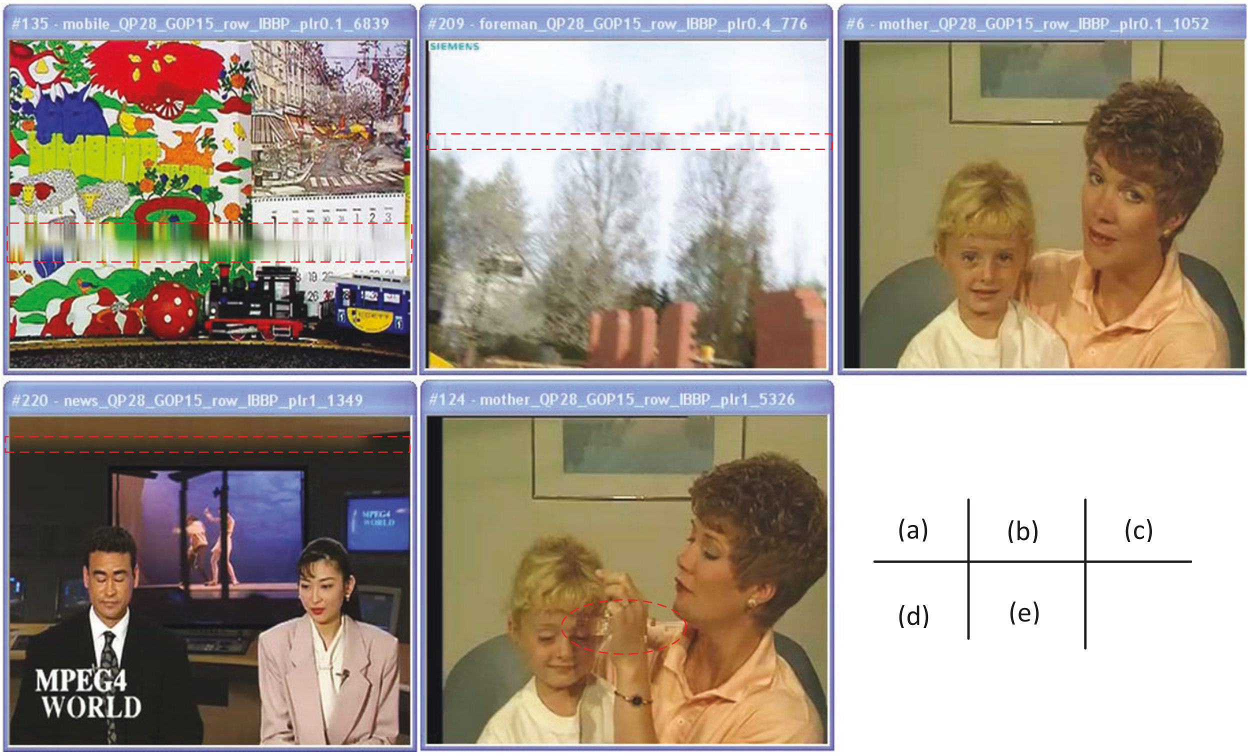

#135 - mobile_QP28_GOP15_row_IBBP_plr0.1_6839
#209 - foreman_QP28_GOP15_row_IBBP_plr0.4_776
#6 - mother_QP28_GOP15_row_IBBP_plr0.1_1052
#220 - news_QP28_GOP15_row_IBBP_plr1_1349
#124 - mother_QP28_GOP15_row_IBBP_plr1_5326
(a)
(b)
(c)
(d)
(e)

Figure 5 Illustration of EC effectiveness. **(a)** Artifacts produced by spatial EC technique; **(b)** artifacts produced by temporal EC technique in area with camera pan; **(c)** no visible artifacts due to the stationary nature of the lost MBs; **(d)** very slightly noticeable artifacts produced by spatial EC technique in area with smooth texture; **(e)** noticeable artifacts only in small area produced by temporal EC technique.

artifact duration caused by P-frame packet loss is less than that by I-frame packet loss. However, the impact of a P-frame packet loss can be significant, if large motion exists in the packet and/or the packets temporally adjacent to it. The artifacts caused by a B-packet loss, if noticeable, look like an instant glitch, because there is no error propagation from B-frame and the artifacts last merely for 1/30 s. When the motion in a lost B slice is low, there are no visible artifacts at all.

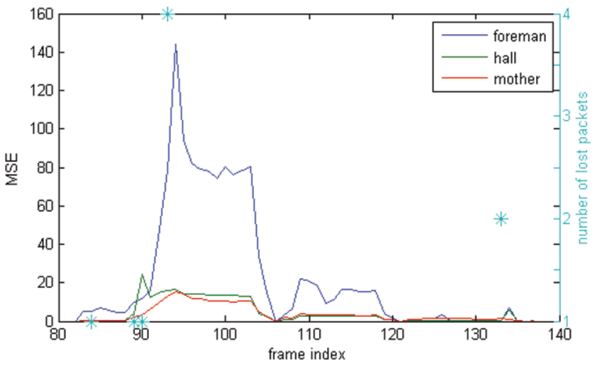


Figure 6 MSE per frame for different video sequences under the same test condition.

4. VQA model with spatiotemporal complexity estimation

Both the effects of EC and the effects of error propagation have close relationship with the spatiotemporal complexity of the lost packets and its spatiotemporally adjacent packets. To improve prediction accuracy of packet-layer VQA model in the quality monitoring case, influence from video content property, EC strategy, and error propagation should be taken into consideration as much as possible. The proposed objective quality assessment model is based on the video spatiotemporal complexity estimation.

4.1 Spatiotemporal complexity estimation

For a video frame indexed as i, the parameter set π_i including frame size s_i, number of total packets $N_{i,\text{total}}$, number of lost packets $N_{i,\text{lost}}$, and the location of lost packet in the frame is calculated or recorded. The location of lost packets in a video frame is detected with the assistance of the sequence number field of RTP header. To identify different frames, the timestamp in RTP header is used. The frame size includes both lost packet size and received packet size. For a lost I-frame packet, its size is estimated as the average of the two spatially adjacent I-frame packets that are correctly received or equal to the size of the spatially adjacent I-frame packet if there is only one spatially adjacent I-frame packet correctly received. For a lost P-frame packet, its size is estimated as the average size of the two temporally adjacent collocated P-frame packets that are correctly received. Similar method is used for size estimate of lost B-frame packet.

The SC and the TC of a slice encapsulated in a packet, which can be roughly reflected by the packet size variation, are estimated using an adaptive thresholding method as shown in Figure 7. In general, I-frame size is much larger than P-frame size, and P-frame size larger than B-frame size. However, when the texture in an I-frame is very smooth, the size of the I-frame is small, which depends on QP used. In the extreme case that the objects in a P-frame are almost stationary, the size of the P-frame can be as small as that of a B-frame; in another extreme case where the objects in a P- or B-frame is rich of texture and diverse motion, the size of the P- or B-frame can be as large as that of a I-frame. In our database, each row of MBs is encoded as a slice; therefore, each detected lost slice is classified with a SC or TC level using adaptive threshold.

For P- or B-slice, if the slice size is larger than a threshold Thrd$_r$, then the slice is classified as high-TC slice; otherwise, if the slice size is larger than a threshold Thrd$_p$, then the slice is classified as medium-TC slice; otherwise, the slice is classified as low-TC slice. The two thresholds are adapted from the empirical equations [20] below. The variable *av_nbytes* is the average frame size in a sliding window. The variant *max_iframe* is the maximum I-frame size, and *nslices* is the number of slices per frame.

$$\text{Thrd}_I = \left(\left(max\_iframe \times 0.995/4 + av\_nbytes \times 2\right)/2\right)/nslices \tag{1}$$

$$\text{Thrd}_P = \left(av\_nbytes \times 3/4\right)/nslices \tag{2}$$

For a I slice, if its size is smaller than thrd$_{\text{smooth}}$, then the slice is classified as smooth-SC slice; otherwise, as edged-SC slice. The thrd$_{\text{smooth}}$ is a function of coding bitrate. In our experiment, thrd$_{\text{smooth}}$ is set to 200 bytes for CIF format sequences coded with H.264 encoder and QP equal to 28.

4.2 Objective assessment model

The building block diagram of the proposed model is shown in Figure 8. The packet information analysis block uses the RTP/UDP header information to get a set of parameters π_i for each frame. These parameters and the encoder configuration information are used by visible artifacts detection module to calculate the level of visible artifacts (LoVA) for each frame. The encoder configuration information includes GOP structure, number of reference frames, error resilience tools like slicing mode, and intra refresh ratio. For a sequence of t seconds, we calculate the mean LoVA (MLoVA) and map the MLoVA to an objective MOS value according to a second-order polynomial function, which is trained using least square fitting technique. The results in [13] showed that the simple metric of mean time between visible artifacts has an average correlation of 0.94 with MOS. Thus, the simple averaging method is used as the temporal pooling strategy in our model.

For the *i*th frame, the LoVA is modeled as the sum of the IVA V_i^0 caused by the loss of the packets of the current frame and the propagated visible artifacts (PVA) V_i^P due to error propagation from the reference frame, as shown in Equation 3. It is assumed here that the visible artifacts caused by current-frame packet loss and by the reference-frame packet loss are independent.

$$V_i = V_i^0 + V_i^P \tag{3}$$

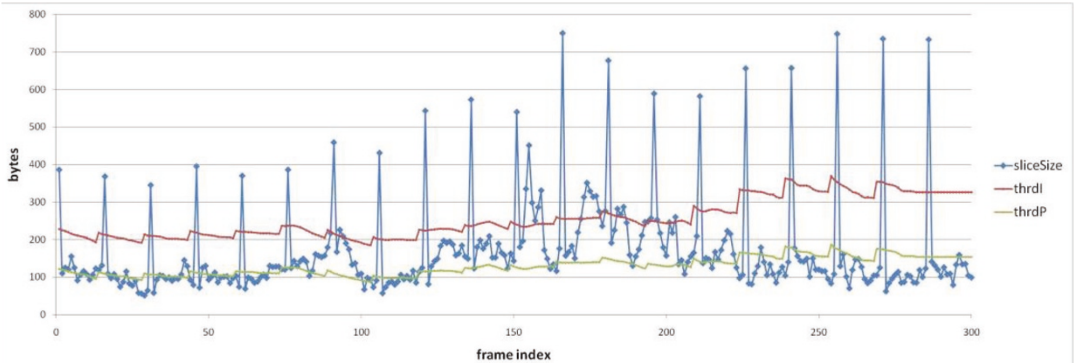


Figure 7 Illustration of the frame-by-frame slice complexity classification based on the adaptive thresholds. The 14th slice of foreman bitstream coded with IPPP GOP structure.

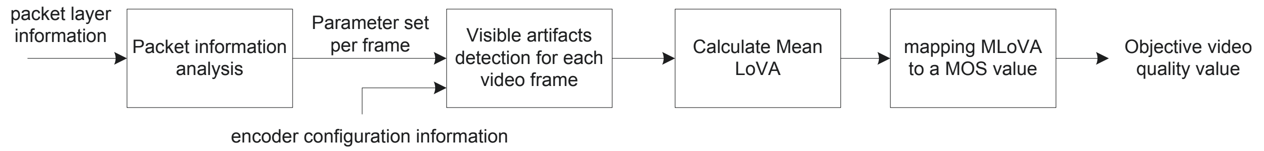

packet layer information
Packet information analysis
Parameter set per frame
Visible artifacts detection for each video frame
encoder configuration information
Calculate Mean LoVA
mapping MLoVA to a MOS value
Objective video quality value

Figure 8 Building block diagram of the proposed model.

The IVA V_i^0 is calculated by

$$V_i^0 = \frac{\sum_{j=1}^{N_{i,\text{ lost}}} w_{i,j}^{\text{location}} \times w_{i,j}^{\text{EC}}}{N_{i,\text{ total}}} \qquad (4)$$

Depending on the location of the lost packets in one frame, different weight $w_{i,j}^{\text{location}}$ is assigned to the lost packet (i.e., lost slice because one coded slice is encapsulated in one RTP packet in our dataset). The location weight allows us differentiating the slice with attention focus from others. In experiments, we found that the contribution of location weight to performance gain is small as compared to EC and EP weights. Thus, simply set location weight to 1. $w_{i,j}^{\text{EC}}$ is the EC weight which reflects the effectiveness of EC technique. As discussed in Section 3.3, the visible artifacts produced by temporal EC approach and spatial EC approach are quite different, correspondingly present different level of annoyance. The blurring artifacts of spatial EC are visibly more annoying than the edged artifacts of temporal EC generally. Further, the EC effectiveness depends on the SC and the TC of the lost slices. For the lost I-slice having smooth texture, the loss can be concealed well with little visible artifacts by the bilinear interpolation-based spatial EC technique. For the lost P- or B-slice having zero MV or same MV as its adjacent slices, it can be recovered well with little noticeable artifacts by the temporal EC technique. It is reported in [10] that, when IVA is above the medium, increasing the distance between the current frame with packet loss and the reference fame used for concealment increases the visibility of packet loss impairment. Therefore, we applied different weights for P-slices of IBBP GOP structure and those of IPPP GOP structure. In summary, the weight $w_{i,j}^{\text{EC}}$ is set according to EC method used and spatial-TC classification as in Table 1. As shown in Figure 5a,b, the perceptual annoyance of the artifacts produced by spatial EC method and temporal EC method is almost at the same level, so we applied the same weight for lost slices of edged-SC type and those of H-TC type. In experiment, the values α_1 to α_5 are set empirically to 0.01, 1, 0.01, 0.1, and 0.3, in order to reflect the relative annoyance of the respective typical artifacts on the artifacts scale ranging from 0 to 1.

The PVA is zero for I frame, because I frame is coded with intra-frame prediction only. For the inter-frame predicted P/B frames, the PVA V_i^P is calculated as

$$V_i^P = \frac{\sum_{j=1}^{N_{i,\text{ total}}} E_{i,j}^{\text{prop}} \times w_i^{\text{EP}}}{N_{i,\text{ total}}} \qquad (5)$$

E_i^{prop} denotes the amount of visible artifacts of reference frames. Its value depends on the encoder configuration information, i.e., GOP structure and the number of reference frames. Taking IPPP structure and two reference frames for an example, the E_i^{prop} is calculated as

$$E_{i,j}^{\text{prop}} = (1-b) \times V_{i-1,j} + b \times V_{i-2,j} \qquad (6)$$

where b is weight for the propagated error from respective reference frames. For our datasets, b = 0.75 for P frames, and b = 0.5 for B frames.

Weight w_i^{EP} modulates the propagation effects of reference frames' artifacts to current frame. The reference frames' artifacts may attenuate because of error resilience tool like Intra MB Refresh or more prediction residual left in the ensuring frames. No matter more Intra-MBs are used or more prediction residual information remains in the compressed bitstream of current slice, the bytes of current slice will be larger than the slice that have fewer Intra-MBs and easy-to-predict content. Therefore, the value of w^{EP} is set according to the spatiotemporal complexity of the frame as in Table 2. In experiment, β_1 is set to 1 which means no artifacts attenuation, and β_2 is set to 0.5, which means visible artifacts attenuates by half.

Finally, clip the value of V_i to [0,1]. Record the value of the LoVA of the frame in a frame queue, and put the frame in the queue according to its displaying order.

Table 1 The value of $w_{i,j}^{\text{EC}}$ depending on EC method and SC/TC classification

Spatial EC method			Temporal EC method		
Smooth-SC	Edged-SC	L-TC	M-TC & IPPP structure	M-TC & IBBP structure	H-TC
α_1	α_2	α_3	α_4	α_5	α_2

Table 2 The value of w_i^{EP} depending on TC classification

L-TC & M-TC	H-TC
β_1	β_2

When time interval of t seconds is reached, the algorithm will calculate the mean LoVA by

$$\mathrm{MLoVA} = \left(\frac{1}{M}\sum_{i=1}^{M} V_i\right)\Big/ f_{\mathrm{r}} \tag{7}$$

where M is the total number of frames in t seconds; f_x is the frame rate of a video sequence.

5. Experimental results

First, we compare the correlation between the subjective MOS and some affecting parameters that are used in the existing packet-layer models. These parameters include PLR [6], burst loss frequency (BLF) [8], and invalid frame ratio (IFR) [21]. In the existing work, these parameters and other video coding parameters like coding bitrate, frame rate, are modeled together. In order to fairly compare the performance of the above parameters that reflect transmission impairment, the coding artifacts are prevented by properly setting QP in our datasets.

Two metrics, Pearson correlation and the modified RMSE, shown in Equation 8 are used to evaluate performance. In the ITU-T test plan draft [18], it is recommended to take the modified RMSE as primary metric and Pearson correlation as informative. The scope of modified RMSE is to remove from the evaluation the possible impact of the subjective scores' uncertainty. The modified RMSE is described as:

$$P_{\mathrm{error}}(i) = \max(0, \left|\mathrm{MOS}(i) - \mathrm{MOS}p(i)\right| - \mathrm{CI}_{95}(i)) \tag{8}$$

The final modified RMSE* is calculated as usual, but based on P_{error} with the equation below.

$$rmse^* = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}\left(P_{error}(i)\right)^2} \tag{9}$$

where the index i denotes the video sample; N denotes the number of samples; and d the number of freedoms. The degree of freedom d is set to 1 because we did not apply any fitting method to the predicted MOS score before comparing it with the subjective MOS.

When evaluating the performance of the features on dataset-1 or dataset-2, the dataset is partitioned into the training sub-dataset and the validation sub-dataset in 50% versus 50% proportion to perform the cross-evaluation process. The Pearson correlation and the modified RMSE in Tables 3 and 4 are the average performance over 100 runs of the cross-evaluation process.

Table 3 The correlation and modified RMSE between different artifact features and subjective MOS

Feature	RMSE*		Pearson correlation	
	Dataset-1	Dataset-2	Dataset-1	Dataset-2
PLR	0.1636	0.4094	0.9397	0.7544
BLF	0.1622	0.4082	0.9409	0.7558
IFR	0.2456	0.4185	0.8973	0.7388
MLoVA	0.1158	0.1932	0.9591	0.9174

The results using least square curve fitting are shown in Table 3. From Figure 9, it can be seen that the correlation between the subjective MOSs and the PLR/BLF/IFR reaches up to 0.94 on dataset-1, but is only 0.75 on dataset-2. This shows that the features PLR/BLF/IFR are effective for video system planning modeling, but are not effective for quality monitoring model.

It can be seen in Figure 9 that our model proposed a better metric, MLoVA, which is more consistent with subjective MOS. When we use second-order polynomial function to fit the curve, the correlation and RMSE pair of predicted MOS versus subjective MOS is (0.96, 0.12) and (0.93, 0.17) on dataset-1 and dataset-2, respectively, Figure 10 shows the predicted MOS as compared with the subjective MOS. This demonstrates that the proposed model has robust performance on both datasets.

Second, the contributions of two factors, namely EC effectiveness and EP model, are quantified on dataset-2. If we set the weights for EC effectiveness to one in Equation 4 and ignore the second item of propagated artifacts by setting it to zero in Equation 3, then the MLoVA regresses to PLR, where the data losses are regarded as equally important to perceptual quality. As described in Section 3, the EC strategy employed at decoder can hide the visible artifacts caused by packet loss to a degree that depends on the spatiotemporal complexity of the lost content. When the complexity estimation-based EC weights are applied to calculate IVA and still ignore the item of propagated error, it is shown in Figure 10b that the correlation of mean IVA (MIVA) with subjective MOS is 0.86, and the modified RMSE is reduced to 0.27. The performance is significantly improved as compared with PLR. Further, the improvement brought by incorporating the error propagation model of Equation 5 was evaluated. As we know,

Table 4 Quantitative analysis of the contribution from EC effectiveness estimation and EP model

Feature	RMSE*		Pearson correlation	
	Dataset-1	Dataset-2	Dataset-1	Dataset-2
PLR	0.1647	0.4095	0.9396	0.7511
MIVA	0.1559	0.2897	0.9408	0.8504
MLoVA$_0$	0.1478	0.2375	0.9490	0.8929
MLoVA	0.1400	0.1909	0.9516	0.9185

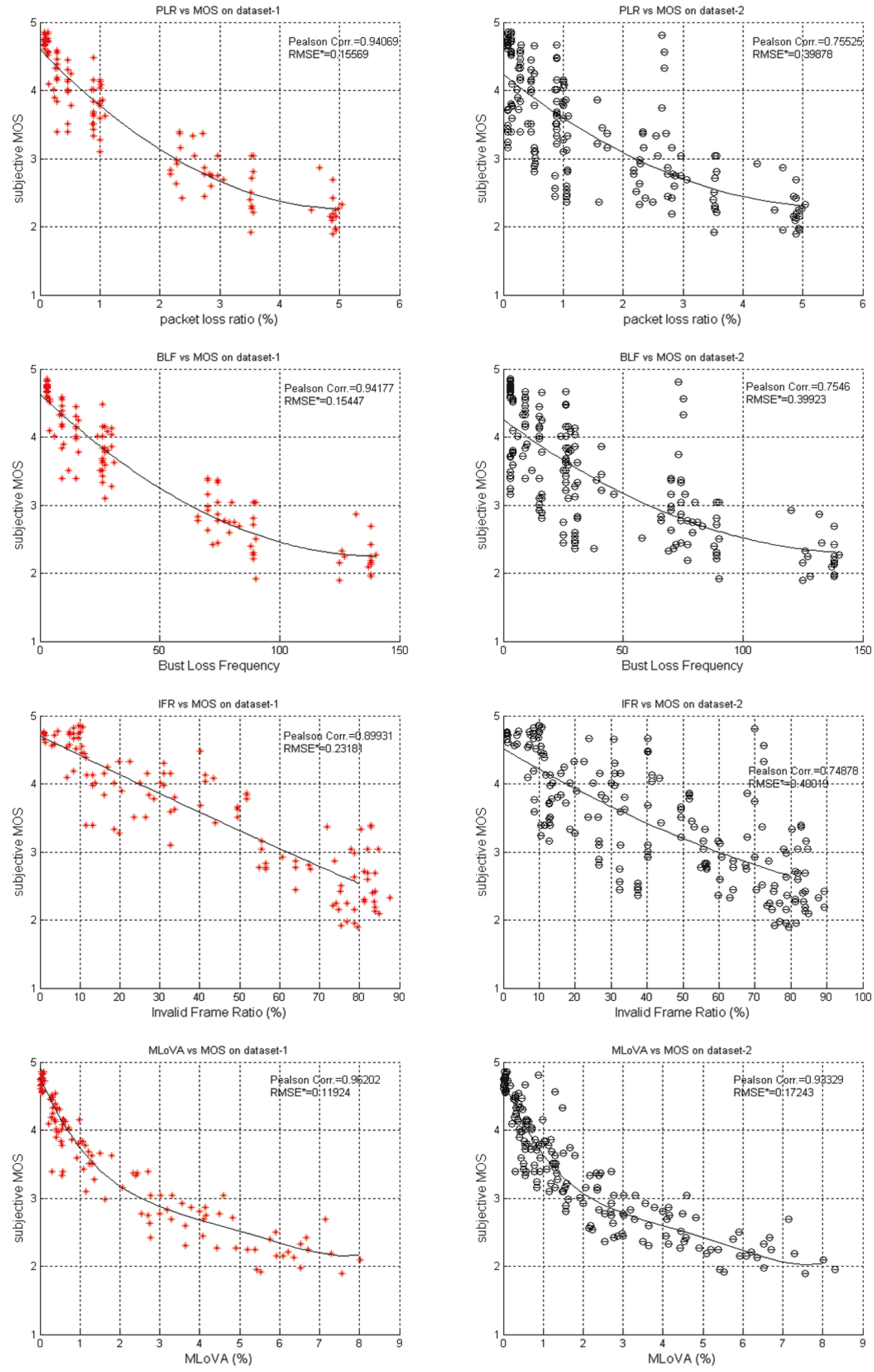
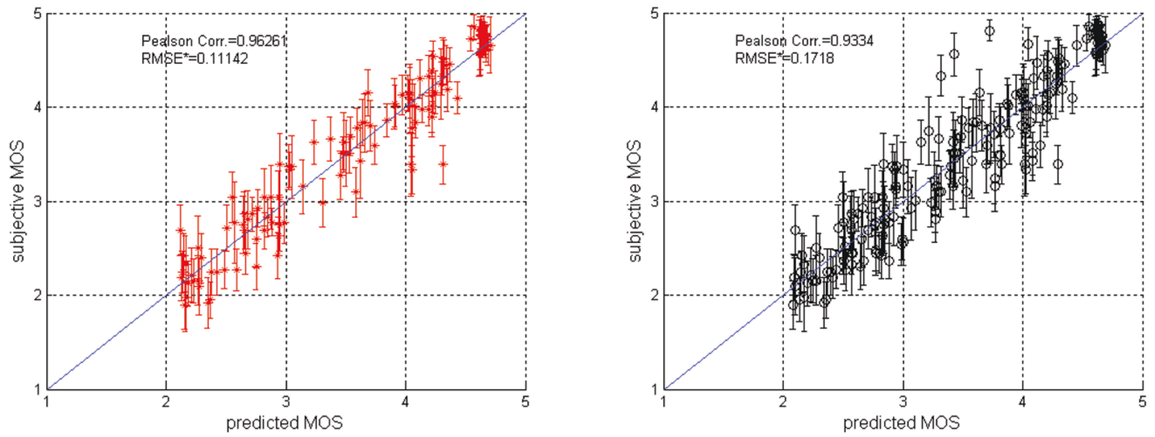


Figure 9 Performance evaluation compared with existing metrics in dataset-1 and dataset-2.

inter-frame prediction is used in video compression, as a result, the influence of an I-packet loss, or a P-packet loss appearing early in a GOP, is quite different from that of a B-packet loss or a P-packet loss appearing later in a GOP. By setting $\beta_1 = \beta_2 = 1$, we did not consider the error attenuation effects during propagation, and denoted the corresponding result of Equation 7 as MLOVA_0. It can be seen that introducing the EP model and the complexity estimation-based EP attenuation weight can further improve the prediction accuracy on dataet-2.



(a) performance on dataset-1

(b) performance on dataset-2

Figure 10 Subjective MOS versus predicted MOS by proposed model in different dataset.

6. Conclusion and future work

In this study, the different requirements of two application scenarios of a parametric packet-layer model are discussed. We provide the insight that different criteria and methods should be used to select processed sequences for subjective evaluation when setting up the evaluation dataset. It is shown that the parameters PLR/BLF/IFR used in existing models are effective for video system planning modeling, but are not effective for video quality monitoring applications.

Further, a model is proposed for video monitoring scenario, taking into consideration the interaction between video content features and EC effects and error propagation effects. It achieves much better performance on both types of datasets for planning and monitoring applications. The result also shows that, for the encoding configuration given in this article, the packet-layer model taking packet header information and encoder configuration information as inputs is able to estimate video quality with enough accuracy for practical use. However, there are many error-resilience tools (e.g., flexible MB order) in H.264 to combat the video quality degradation in case of transmission losses and different EC strategies that may be employed at a decoder. A packet-layer model must be tailored to the specific video application configuration. For future study, the influence of the distribution of visible packet losses on the overall perceived quality will be studied.

Competing interests

The authors declare that they have no competing interests.

Received: 1 November 2010 Accepted: 22 August 2011
Published: 22 August 2011

References

1. Takahashi A, Hands D, Barriac V: **Standardization activities in the ITU for a QoE assessment of IPTV.** *IEEE Commun Mag* 2008, **46(2)**:78-84.
2. ITU-T document, **Draft terms of reference (ToR) for P.NAMS.** 2009 [http://www.itu.int/md/meetingdoc.asp?lang=en&parent=T09-SG12-091103-TD-GEN-0146].
3. ITU-T document, **Draft Terms of Reference (ToR) for P.NBAMS.** 2009 [http://www.itu.int/md/T09-SG12-110118-TD-GEN-0521].
4. Rui H, Li C, Qiu S: **Evaluation of packet loss impairment on streaming video.** *J Zhejiang Univ Sci* 2006, **A7**:131-136.
5. Reibman AR, Vaishampayan VA, Sermadevi Y: **Quality monitoring of video over a packet network.** *IEEE Trans Multimedia* 2004, **6(2)**:327-334.
6. Yamagishi K, Hayashi T: **Video-quality planning model for videophone services.** *Inf Media Technol* **4(1)**:1-9.
7. Mohamed S, Rubino G: **A study of real-time packet video quality using random neural networks.** *IEEE Trans Circ Syst Video Technol* 2002, **12(12)**:1071-1083.
8. Yamagishi K, Hayashi T: **Parametric packet-layer model for monitoring video quality of IPTV services.** *IEEE International Conference on Communications* 2008, 110-114.
9. Raake A, Garcia M-N, Moller S, Berger J, Kling F, List P, Johann J, Heidemann C: **T-V-model: parameter-based prediction of IPTV quality.** *Proc ICASSP* 2008, 1149-1152.
10. Kanumuri S, Cosman PC, Reibman AR, Vaishampayan VA: **Modeling packet loss visibility in MPEG-2 video.** *IEEE Trans Multimedia* 2006, **8(2)**:341-355.
11. Reibman AR, Poole D: **Predicting packet-loss visibility using scene characteristics.** *Proceedings of the International Workshop in Packet Video* 2007, 308-317.
12. Yamada T, Miyamoto Y, Serizawa M: **No-reference video quality estimation based on error-concealment effectiveness.** *IEEE Packet Video Workshop* 2007, 288-293.
13. Suresh N: **Mean time between visible artifacts in visual communications.** *PhD thesis, Georgia Institute of Technology* 2007.
14. Winkler S, Dufaux F: **Video quality evaluation for mobile applications.** *Proc VCIP* 2003, 593-603.
15. Winkler S, Mohandas P: **The evolution of video quality measurement: from PSNR to hybrid metrics.** *IEEE Trans Broadcast* 2008, **54(3)**:660-668.
16. ITU-T Rec. P.910, **Subjective video quality assessment methods for multimedia applications.** Geneva; 2008.
17. Simone FD, Naccari M, Tagliasacchi M, Dufaux F, Tubaro S, Ebrahimi T: **Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel.** *Proc International Workshop on Quality of Multimedia Experience (QoMEx)* 2009, 204-209 [http://mmspl.epfl.ch/].
18. ITU-T document, **Qualification test plan for P.NAMS.** [http://www.itu.int/md/meetingdoc.asp?lang=en&parent=T09-SG12-091103-TD-GEN-0150], accessed on October 2009.

19. ITU-R Rec. BT.500-10, Methodology for the subjective assessment of the quality of the television pictures. 2000.
20. Clark A: **Method and system for viewer quality estimation of packet video streams.** 2009 [http://www.freepatentsonline.com/y2009/0041114.html], U.S. Patent 2009/0041114A1.
21. Hayashi T, Masuda M, Tominaga T, Yamagishi K: **Non-intrusive QoS monitoring method for realtime telecommunication services.** *NTT Tech Rev* 2006, **4(4)**:35-40.

doi:10.1186/1687-5281-2011-5
Cite this article as: Liao and Chen: **A packet-layer video quality assessment model with spatiotemporal complexity estimation.** *EURASIP Journal on Image and Video Processing* 2011 **2011**:5.