Hindawi Publishing Corporation
Advances in Multimedia
Volume 2014, Article ID 606493, 15 pages
http://dx.doi.org/10.1155/2014/606493

Research Article

No-Reference Video Quality Assessment Model for Distortion Caused by Packet Loss in the Real-Time Mobile Video Services

Jiarun Song and Fuzheng Yang

State key laboratory of ISN, School of Telecommunications Engineering, Xidian University, Taibai Road, Xian 710071, China

Correspondence should be addressed to Jiarun Song; sjrxidian@126.com

Received 20 August 2014; Revised 13 November 2014; Accepted 13 November 2014; Published 11 December 2014

Academic Editor: Deepu Rajan

Packet loss will make severe errors due to the corruption of related video data. For most video streams, because the predictive coding structures are employed, the transmission errors in one frame will not only cause decoding failure of itself at the receiver side, but also propagate to its subsequent frames along the motion prediction path, which will bring a significant degradation of end-to-end video quality. To quantify the effects of packet loss on video quality, a no-reference objective quality assessment model is presented in this paper. Considering the fact that the degradation of video quality significantly relies on the video content, the temporal complexity is estimated to reflect the varying characteristic of video content, using the macroblocks with different motion activities in each frame. Then, the quality of the frame affected by the reference frame loss, by error propagation, or by both of them is evaluated, respectively. Utilizing a two-level temporal pooling scheme, the video quality is finally obtained. Extensive experimental results show that the video quality estimated by the proposed method matches well with the subjective quality.

1. Introduction

Nowadays, the rapid growth of multimedia and network technologies and the popularization of the smart mobile devices greatly stimulate the development of mobile multimedia applications, covering a wide range of scenarios from videophone to mobile internet protocol television (IPTV) [1]. Most of these video services allow the users to receive multimedia services such as video, audio, and graphics, through IP-based networks, whenever they want and wherever they are [2]. However, the perceived audio-visual quality cannot be guaranteed in an IP network due to its best-effort delivery strategy. Therefore, it is indispensable to employ quality evaluation of the network video for quality of service (QoS) planning or control in the involved video applications. By evaluating the video quality at network nodes, the video quality can be estimated and sent back to the video service providers or network operators. They can adjust the transmitting strategy in real time to improve the user's experience.

Generally, the RTP/UDP/IP transport protocol is used in the real-time video applications due to its low end-to-end delays. However, these RTP/UDP-based video services are likely to suffer from packet loss due to the limited network capacity or bandwidth variation. When a packet loss occurs during transmission, significant errors may appear due to the corruption of related video data. Additionally, existing video coding standards or proprietary video encoders usually employ predictive coding structures to improve coding efficiency. As a result, the transmission errors in a frame will not only cause decoding failure of itself at the receiver side, but also propagate to its subsequent frames along the motion prediction path, which will bring a significant degradation of end-to-end video quality. Therefore, it is crucial to consider the effects of packet loss for video transmitted over IP-based networks.

Much research has indicated that the overall video quality can be evaluated on basis of the video sequence, where the packet loss rate is extensively utilized [3–5]. However, using the packet loss rate alone cannot provide accurate quality assessment for a certain service. It has been widely recognized that the quality degradation at a given packet loss rate varies with respect to different video sequences, loss position, and loss pattern. In [6], the packet loss robustness factor, which expresses the degree of video quality robustness to packet loss, is introduced to estimate the video quality

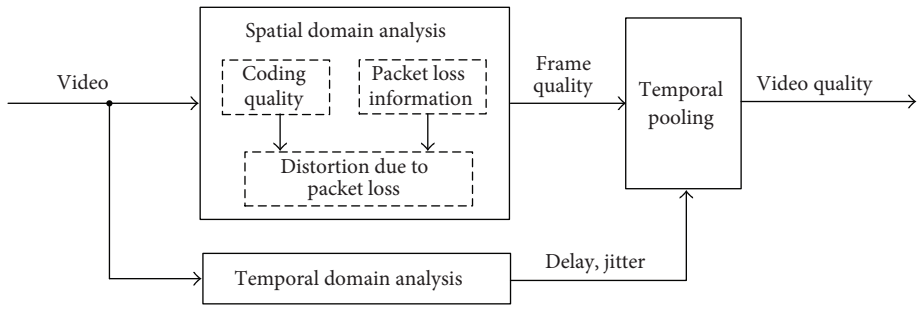

Figure 1: Framework of the video quality assessment with pooling mechanism.

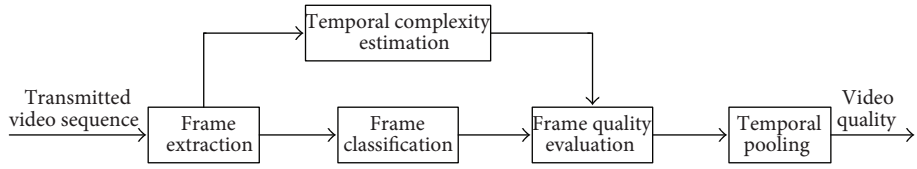

Figure 2: Framework of the proposed video quality assessment model.

apart from the packet loss rate. In ITU Recommendation P.1201.1 [7], other packet loss information is taken into consideration to evaluate the video quality, for example, the average impairment rate of video frame, the impairment rate of video stream, and the packet-loss event frequency. However, such methods cannot exactly capture the influence of lost packets on a specific video because only the statistical parameters are used.

Recently, plenty of research has paid attention to evaluate the overall video quality through pooling quality of each frame [8–11]. Such a general framework is illustrated in Figure 1. The quality of each frame is first evaluated, and then the video quality will be assessed by using a pooling strategy, conventionally resorting to averaging or weighted averaging of the frame quality. These methods can take full advantage of the characteristics of the compression and packet loss for video quality evaluation, such as the type of lost frame (I, P, B), the lost frame position in a group of pictures (GOP), and the total number of lost frames. Generally, they can obtain a superior performance than the traditional methods evaluated with the statistical parameters. It is worth noting that the frame quality plays an important role in the general framework, which directly influences the overall video quality. Therefore, how to accurately evaluate the frame quality is a key point for these methods. In [8], a perceptual full-reference video quality assessment metric is designed focusing on the temporal evolutions of the spatial distortions. The spatial perceptual distortion maps are computed for each frame with a wavelet-based quality assessment (WQA) metric. However, this method relies on full access to the original video, which limits its application in the network video service. In our previous work [9], a no-reference quality assessment framework (i.e., NR QANV-PA) was presented for monitoring the quality of networked video with the primary analysis centered on the decoded packet and the video frame header. The frame quality was measured by taking into account the impact of both the compression and packet loss. However, this method only focuses on the case of single packet loss in a GOP but is not considered the situation of multiple packet losses. Therefore, the performance of this method still needs to be improved.

Considering the masking effect of the human visual system (HVS), the temporal complexity (or motion complexity) is always employed to estimate the perceptual frame quality affected by the packet loss [11–13]. Generally, when packet loss occurs, the quality degradation of the frame in the video with high temporal complexity is larger than that with low temporal complexity. This is because some temporal error concealment methods only work well if the video content is of low motion activity and the difference between adjacent frames is insignificant [9]. In this case, the distortion caused by the packet loss usually remains quite noticeable even after error concealment operations [14]. The visibility of errors caused by packet loss will significantly depend on the video content. However, it is hard to describe and quantify the motion activity, which leads to difficulty in accurately estimating the temporal complexity.

In this paper, a no-reference objective model is proposed to evaluate the video quality affected by packet loss for the real-time mobile video service. Different from the traditional methods that estimate the video quality on basis of the video sequence, the proposed model focuses on accurately evaluating the quality of each frame and then pooling them through a proper way to better estimate the video quality. More specifically, the framework of the proposed model is illustrated in Figure 2. The frames are firstly extracted from the transmitted video sequence. Then, the characteristic of video content is estimated in terms of temporal complexity, by distinguishing the active parts from the static parts in each frame. Meanwhile, the frames are classified into different categories according to the various effects of lost packets on a frame, such as the directly packet loss and error propagation. Combined with the characteristic of the video content, the quality of frame in each category is evaluated. And finally,

the video quality is obtained by a two-level temporal pooling scheme. Particularly, it is assumed in this work that direct packet loss will lead to entire frame loss. It is valid for the real-time video services when decoders or video clients adopt the strategy of entirely discarding a frame that has been corrupted or missing information and repeat the previous video frame instead, until the next valid decoded frame is available [15].

The remainder of this paper is organized as follows. Section 2 briefly reviews some related research on frame quality evaluation and temporal complexity prediction. Section 3 describes the details of our proposed method for estimating the temporal complexity. In Section 4, the impacts of packet loss on the frame quality are analyzed and evaluated, and then a two-level temporal pooling scheme is introduced to estimate the video quality. Performance evaluation and conclusion are given in Sections 5 and 6, respectively.

2. Related Works

In recent years, much attention has been paid to the frame quality evaluation, especially the quality degradation caused by packet loss. This section will briefly review some classic methods for frame quality evaluation. Meanwhile, the temporal complexity estimation methods will also be summarized from the viewpoint of employed information.

2.1. Frame Quality Evaluation. According to [16, 17], conventional methods for video quality assessment (VQA) can be classified into three categories: full-reference (FR) methods, reduced-reference (RR) methods, and no-reference (NR) methods, depending on the level of information available with respect to the original video. Because a video sequence is composed of frames, the methods for evaluating frame quality can be classified in the same way. In the following part of this subsection, the evaluation methods in each category will be discussed in detail.

FR methods rely on full access to the original video. There are many quality metrics widely used for FR objective image/video assessment, such as mean-square error (MSE), peak-signal-to-noise ratio (PSNR), and video quality metric (VQM) [18–21]. For instance, in [18], the sum of the MSEs of all the loss-affected frames is used to express the distortion caused by direct frame loss and the subsequent error propagation. In [19], the authors propose a PSNR-based model of video quality metric to evaluate the quality degradation due to coding artifacts. Based on this study, the overall quality degradation is modeled as a weighted sum of encoding degradation and packet-loss incurred distortion in [20], where the quality degradation caused by packet loss is calculated with the sum over the PSNR drops of all erroneous frames. Despite their popularity, these metrics only have an approximate relationship with the perceived image/video quality, simply because they are based on a byte-by-byte comparison of the data without considering what they actually represent [21]. To solve this problem, Liu et al. propose an FR method to evaluate the perceptual distortion of each individual frame due to packet loss and its error propagation, considering both the luminance masking and activity masking effects of the HVS [22]. However, such FR evaluation methods require source signals as a reference, which limits their application in the network video service.

For RR methods, they provide a solution that lies between FR and NR models and only have access to certain portion of the original video. The video quality experts group (VQEG) has included RR image/video quality assessment as one of its directions for future development. In [23], a RR image quality assessment method is proposed based on a natural image statistic model in the wavelet transform domain. Authors in [24] propose a generalized linear model for predicting the visibility of multiple packet losses, where content-independent factors and content-dependent factors in each frame are considered. However, a successful RR quality assessment method must achieve a good balance between the data rate of RR features and the accuracy of image quality prediction. On one hand, with a high data rate, one can include a large amount of information about the reference image, leading to more accurate estimation of distorted image quality, but it becomes a heavy burden to transmit the RR features to the receiver. On the other hand, a lower data rate makes it easier to transmit the RR information but harder for accurate quality estimation [23]. The cost of maintaining a reliable alternate channel to a central facility may be prohibitive even for RR methods. Moreover, the RR methods are not possible if for any reason the original video is unavailable.

In NR methods, no information from the original video is available [25]. They just extract the information from the bitstream or the reconstructed frames to evaluate the video/frame quality. For instance, an analytical framework for no-reference perceptual quality assessment is proposed to evaluate the impact of packet loss on the video sequence in [26], where the blockiness of each frame is calculated to check the errors that occur in the transmission process. Authors in [27] propose a method based on support vector regression (SVR) for predicting the visibility of packet losses in SD and HD H.264/AVC video sequences and modeling their impact on perceived quality. More specifically, the visibility of packet loss is calculated using the average motion vector difference of each frame. Moreover, Staelens uses the decision tree classifier to model packet loss visibility [28]. In the NR QANV-PA model, the frame quality degradation caused by both packet loss and error propagation is estimated as well as the coding quality. The temporal and spatial characteristics of the video content are also considered. The NR metric is generally seen as the most practical method for assessing network video quality because it can be carried out readily, in real-time, with no expended additional bandwidth.

2.2. Temporal Complexity Estimation. As an inherent characteristic of the video, temporal complexity has been widely studied in both video compression and video quality assessment. Some of studies estimated the motion activity utilizing the information of pixel domain. For example, ITU-T Recommendation P.910 describes the temporal information (TI) of a video sequence using the pixel information [29], where the difference between the pixel values (of the luminance plane) at the same location in space but at successive frames is employed. Based on this analysis, frame difference, normalized frame difference, and displaced frame difference

in [30] are proposed to estimate the temporal complexity. Wolf proposes the concept of "motion energy" to estimate the temporal complexity, utilizing the pixel information as well [31]. However, the pixel values are closely related to the texture information of the video. The video sequences with low motion activity but high texture information may have drawn the wrong conclusion on the temporal complexity estimation by using these pixel-based metrics.

Apart from these metrics, some studies estimate the temporal complexity utilizing the information extracted from the video bitstream. For instance, in the NR QANV-PA model [9], the quantization parameter (QP) and the bit rate are used to estimate the temporal complexity. In [11], the temporal complexity is predicted using the number of bits for coding I frames and P frames. However, the parameters in these models cannot exactly describe the motion characteristics of the video content because the information between frames is not used. Compared to these parameters, the motion vector (MV) related factors can reflect the acuteness of temporal changes of video sequences more effectively. Some MV-based methods have been proposed to estimate the motion activity. Feghali et al. and Wan et al. use the average magnitude of motion vectors (MVM) in each frame to interpret the temporal complexity [32, 33]. Moreover, the standard deviation of MVM and the motion direction activity were employed in [30, 34], respectively. However, the HVS tends to pay more attention to moving objects and the distortion is more noticeable for the active part in each frame [33]. It cannot distinguish the influence of different parts in a frame just by averaging MVM of all blocks. Therefore, how to accurately estimate the temporal complexity is still a great challenge.

3. Temporal Complexity Estimation

The visibility of errors caused by packet loss significantly depends on the video content, especially its temporal complexity or motion complexity. Figure 3 illustrates the distortion caused by packet loss in the sequences of *BasketballDrive*, *Sunflower,* and *Optis*, respectively, where the sequences are compressed with a constant QP. It is obvious that the distortion due to packet loss (in the red box) in *BasketballDrive* is more visible than that in *Sunflower* and *Optis*. This is because the motion activity of *BasketballDrive* is much higher than that of other sequences. When packet loss occurs, the errors will be larger due to employing the inter prediction techniques in the codecs. In order to better estimate the video quality, it is indispensible to take the temporal complexity into consideration. In this section, a superior method for temporal complexity evaluation will be proposed.

For the coded video frame, the motion activity of each part can be expressed by MVs. Therefore, the MV field (MVF) will be employed to determine the activity degree of the frame. Given an MVF with $W \times H$ macroblocks (MBs), the MVM value of each MB $B_{i,j}$ $(0 \le i \le W, 0 \le j \le H)$ can be calculated as follows:

$$M(i,j) = \sqrt{\left|\mathrm{MV}_x(i,j)\right|^2 + \left|\mathrm{MV}_y(i,j)\right|^2}, \quad (1)$$

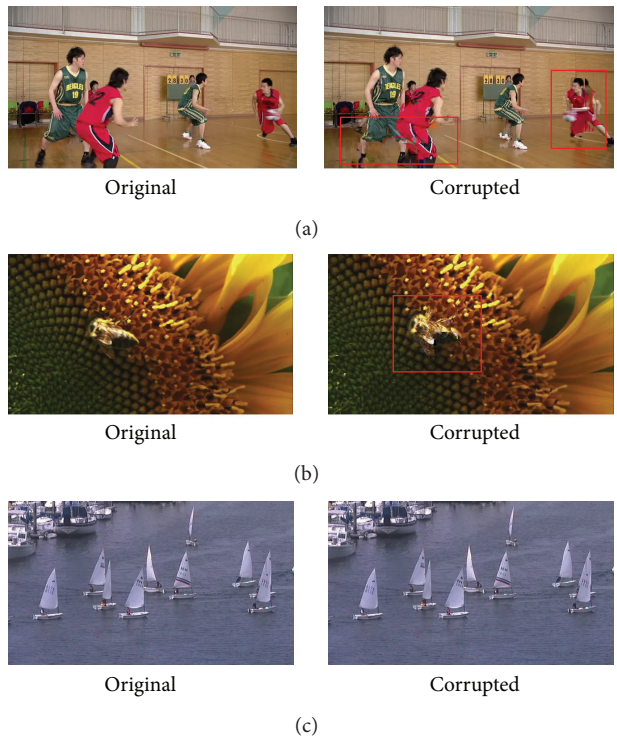

Figure 3: Packet loss in different video sequences. (a) *BasketballDrive,* (b) *Sunflower*, and (c) *Optis.*

where $\mathrm{MV}_x(i,j)$ and $\mathrm{MV}_y(i,j)$ are the x-axis and y-axis components of the absolute motion vectors, respectively. $\mathrm{MV}(i,j)$ represents the magnitude of motion vectors for MB $B_{i,j}$. It is notable for some special situations that if the MB is further divided into subblocks, the MVM of the MB will be obtained by averaging all the MVM of the subblocks. For the intracoding blocks in the frame, the MVM will be regarded as zero. Moreover, the MV of a direct or skip mode block is predicted from its adjacent blocks, so the MVM value of a direct or skip mode block is equal to its prediction.

As mentioned in Section 2.2, the traditional methods to estimate the temporal complexity always average the MVM of all blocks in a frame. These methods will greatly reduce the influence of the active region in a frame to the overall frame quality. When faced with visual stimuli, HVS cannot process the whole scene in parallel. Due to the limited visual angle, human eyes actually pay most of the attention to objects with high temporal changes, and therefore static background or smoothly moving objects in peripheral regions are partly ignored. For example, the sailing boats in sequence *Optis* are all smoothly moving, and the average MVM value of the frame is quite small and the distortion in the frame is not visible. However, the distortion of the bee in sequence *Sunflower* is more visible than that of other regions. It may greatly affect the quality of the frame even though the average MVM value of the frame is small. Therefore, the temporal complexity should be estimated considering the different weightiness of temporal masking effect for high motion regions and low motion regions.

Considering the difference in the MVM values, the type of an MB can be classified into two categories: a static MB

for background and a moving MB for moving object, where the MVM value is zero in a static MB and correspondingly nonzero in a moving MB. To distinguish the different degrees of motion activity for moving objects, the moving MB is further divided into the following three types in accordance with the MVM values: the one with slight motion activity, the one with moderate motion activity, and that with intense motion activity. Therefore, the partition results of the moving MBs in a frame can be described as

$$\begin{aligned}
&(i,j) \in \Omega_S, \quad F_S(i,j) = 1, \quad \text{when } M(i,j) \in (0, T_1] \\
&(i,j) \in \Omega_M, \quad F_M(i,j) = 1, \quad \text{when } M(i,j) \in (T_1, T_2] \\
&(i,j) \in \Omega_I, \quad F_I(i,j) = 1, \quad \text{when } M(i,j) \in (T_2, +\infty),
\end{aligned} \tag{2}$$

where (i, j) is the index of the block in a frame and Ω_S, Ω_M, and Ω_I denote the sets of integer indexes to blocks belonging to slight motion regions, moderate motion regions, and intense motion regions, respectively. F_S, F_M, and F_I are the flags to identify different types of blocks. T_1 and T_2 are the thresholds of MVM equal to $\sqrt{2}$ and $5\sqrt{2}$, respectively. Here, the values of T_1 and T_2 are determined through empirical observations. For the video sequence with low or medium resolutions, we found that when $\mathrm{MV}_x(i, j)$ and $\mathrm{MV}_y(i, j)$ are not larger than 1, the motion activity of the MB is quite slight. Therefore, the value of T_1 can be calculated by $\sqrt{1^2 + 1^2}$. Moreover, when $\mathrm{MV}_x(i, j)$ and $\mathrm{MV}_y(i, j)$ are larger than 7 and both of them are not smaller than 1, the motion activity of the MB is quite intense. Therefore, the value of T_2 can be achieved by $\sqrt{7^2 + 1^2}$.

Generally, most video sequences consist of some objects and background, and the MVs belonged to the same moving object or background which usually have similar magnitudes. To better indicate the temporal complexity of each intercoded frame, the percentage of the blocks with slight motion activity (denoted as R_s), with moderate motion activity (denoted as R_m), and with intense motion activity (denoted as R_i) in a frame is jointly introduced. Taking the video sequences in Figure 3 as an example, in the sequence *BasketballDrive*, the players can be regarded as the R_i part due to their intense motion activity and the background is regarded as the R_s part. In the sequence *Sunflower*, the bee is with moderate motion activity, so it can be regarded as the R_m part, while the background flower is the R_s part. In the sequence *Optic*, all the objects can be regarded as the R_s part due to their slight motion activity. It is found that the stimulus of each part to human is usually different from others and the temporal complexity of the sequence should not be estimated by simply averaging all the parts of objects or background. Therefore, a new method combining the characteristics of HVS is proposed to estimate the temporal complexity $\delta_{t,f}$ of video sequence as

$$\delta_{t,f} = w_s \cdot R_s \cdot M_s + w_m \cdot R_m \cdot M_m + w_i \cdot R_i \cdot M_i, \tag{3}$$

where R_s, R_m, and R_i can be achieved by

$$\begin{aligned}
R_s &= \frac{\sum_{(i,j)\in\Omega_S} F_s(i,j)}{N_B}, \\
R_m &= \frac{\sum_{(i,j)\in\Omega_M} F_M(i,j)}{N_B}, \\
R_i &= \frac{\sum_{(i,j)\in\Omega_I} F_I(i,j)}{N_B},
\end{aligned} \tag{4}$$

where N_B is the total number of MBs in a frame. For the temporal complexity $\delta_{t,f}$, it consists of three parts: the first part $R_s \cdot M_s$ means the contribution of the slight motion regions to the motion activity, and the second part $R_m \cdot M_m$ means the contribution of the moderate motion regions to the motion activity, while the third part $R_i \cdot M_i$ means the contribution of the intense motion regions to the motion activity. w_s, w_m, and w_i are weight factors in a range from 0 to 1. Since humans are more sensitive to the significantly active region in a frame, the weightiness of the temporal masking effect for significantly active regions should be higher than that of other regions. Therefore, there should be $w_s < w_m < w_i$, $w_s + w_m < w_i$, and $w_s + w_m + w_i = 1$. In this paper, the values of w_s, w_m, and w_i are set as 0.1, 0.3, and 0.6, respectively, to make $\delta_{t,f}$ reflect the temporal complexity effectively and accurately.

For a video sequence without scene change, the motion activity between successive frames will not change sharply. Therefore, the temporal complexity of the video sequence δ_t can be evaluated by averaging the $\delta_{t,f}$ values of all the intercoded frames. In the following part, the values of δ_t will be employed to evaluate the quality of each frame.

4. Evaluation of Frame Quality and Video Quality

Packet loss seriously impairs the frame quality of the video transmitted over lossy channels, not only because of the data corruption due to packet loss, but also owing to error propagation along the path of motion compensated prediction (MCP). The distortion in a frame caused by a lost packet will propagate to the subsequent frames until an I-frame is introduced for update [33].

Considering the different effects of lost packets on a frame, the frames in a video sequence can be classified into four categories: frames without any distortion caused by packet loss, frames affected only by the loss of their reference frame, frames affected by error propagation without reference frame loss, and frames affected by both loss of their reference and error propagation. An example is shown in Figure 4, where pictures P0–P8 stand for nine successive frames. P0 and P1 are I-frame and P-frame, respectively, without any influence by packet loss, and the quality of these frames can be evaluated by considering compression only. P2 and P5 are P frames that suffer from packet loss, and the data of these frames are completely lost in our assumed situation. Based on

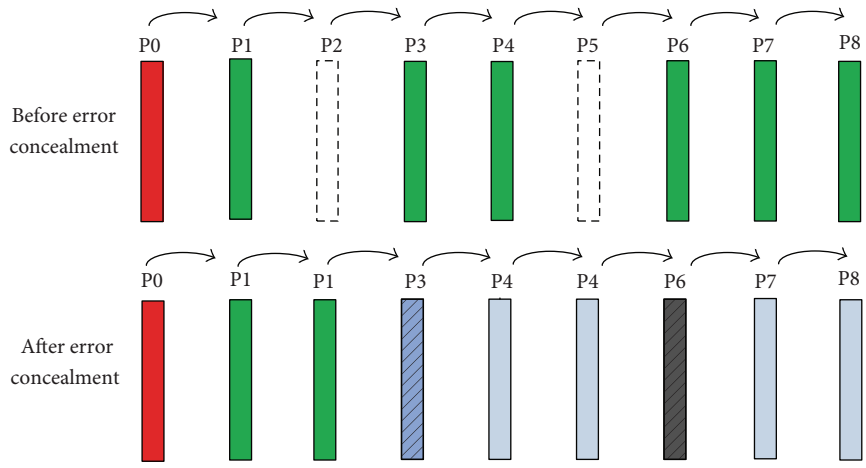

Figure 4: Video frames affected by packet loss.

the zero-motion error concealment mechanism (ZMEC), P2 and P5 are replaced by their reference frames P1 and P4. P3 is a P frame whose reference frame is lost. It is the first frame that is affected by error propagation. P4, P7, and P8 are P frames whose reference frames are not lost but affected by error propagation, while P6 is influenced by another lost packet which makes its reference lost apart from error propagation.

In the following parts of this section, the remaining three types of distortion in the contaminated frames will be discussed in terms of distortion caused by loss of the reference frame, by error propagation without reference frame loss, and by both of them, respectively. And the corresponding evaluation models of frame quality will be derived by combining with the temporal complexity. Utilizing a two-level temporal pooling method, the video quality will be obtained as well.

4.1. Distortion Caused by Loss of the Reference Frame. As illustrated above, P3 is the frame directly following the lost frame, and errors will be introduced in P3 due to prediction. However, P3 is different from the other frames affected by error propagation because the distortion in P3 appears in the sequence for the first time, but in other frames it is just aggravated by error propagation. The stimulus of these frames to human may be quite different in the video sequence. Therefore, the quality of frame P3 will be analyzed separately.

To obtain the perceived quality of the frame directly following the lost frame, a number of experiments have been carried out using a series of video sequences with different content. The involved sequences include *BasketballDrive*, *ParkJoy*, *Crowdrun*, *Sunflower,* and *Optis*. The resolution of each sequence is 640 × 360 (16 : 9). There are 250 frames in each sequence in total. The frames are numbered from 1 to 250. All the sequences are compressed by the ×264 coded at 25 frames per second (fps), with the QP at 24, 30, 36, and 42, respectively. The number of reference frame is set as 1. The GOP of each sequence is 30 with the structure of "IPPP." In practice, ZMEC is widely used in decoders for its effectiveness and simplicity and is hence adopted in this paper. An in-depth investigation into other types of error concealment methods or a more general case is currently underway.

Table 1: Lost frame index of each sequence.

Sequence	Index of the lost frame	Index of the frame to be evaluated
BasketballDrive	127	128
ParkJoy	97	98
Crowdrun	217	218
Sunflower	37	38
Optis	67	68

The subjective tests are carried out following the guidelines specified by VQEG. The test method is based on double-stimulus continuous quality-scale (DSCQS) procedure [29]. A 5-point scale is used to obtain the mean opinion score (MOS) of video frames. All the monitors used for display are with 22-in LCD flat panel and equipped in a medium illuminance environment as indicated in [35]. Twenty nonexpert viewers participate in each test with a viewing distance of 4H and test subjects have been screened for visual acuity and color blindness. The test subjects are between 20 and 30 years old, including 9 women and 11 men. All the subjects are naive test subjects in the sense that none of them works in the field of video coding or visual quality evaluation.

In the experiment, the index of the lost frame in each sequence is set randomly, as shown in Table 1. The successive frame of the lost frame is extracted from the impaired sequence and the original coded sequence, respectively, and evaluated under different QPs. Each frame is presented in its uncompressed version for 4 seconds, then a gray screen is shown for 1 second, and the distorted frame is shown for another 4 seconds. Each session lasts less than 30 minutes.

The quality of the extracted frame is obtained by subjective test and all the values of frame quality form the training set TR1. Then, the distortion of the frame caused by reference frame loss, that is, D_l, can be calculated as

$$D_l = Q_c - Q_l, \tag{5}$$

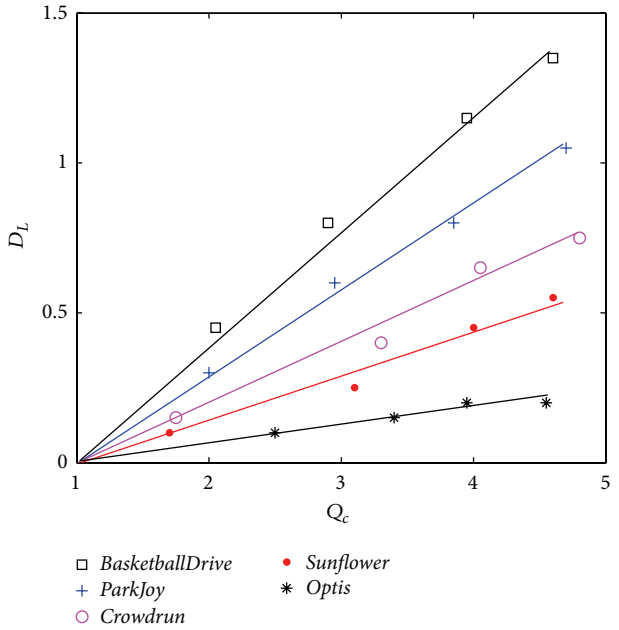

FIGURE 5: Relationship between Q_c and D_l.

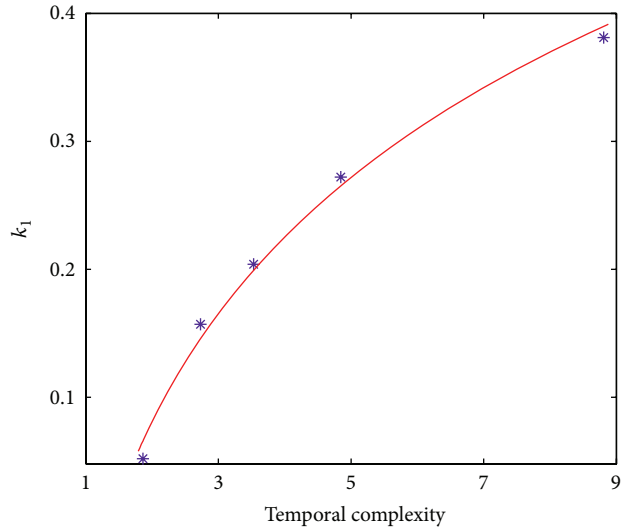

FIGURE 6: Relationship between δ_t and the slope k_1.

where Q_c is the coding quality of the frame and Q_l is the quality of the frame suffering from reference frame loss. Generally, the quality degradation caused by packet loss relies on the coding quality of the reconstructed frame. The distortion of the frame with high quality is usually larger than that in the frame with low quality when a packet loss occurs, because it is more visible in the frame with high quality. More specifically, the relationship between D_l and Q_c of frames with different content is analyzed, as shown in Figure 5. It is observed that the values of D_l increase linearly with the increasing of Q_c, which can be expressed as

$$D_l = k_1 \cdot (Q_c - 1), \tag{6}$$

where k_1 is a parameter to indicate the increasing degree of the distortion, which can be obtained by using linear regression. However, the values of k_1 for different video sequences in Figure 5 are quite different. This is because the distortion caused by packet loss is also correlated with the video content. For the video sequences with the same coding quality (under the same QP), if the video content is with larger motion activity, the distortion will be more visible and the values of k_1 will be larger correspondingly. Therefore, to determine the values of k_1, the relationship between the temporal complexity and k_1 for different sequences is further analyzed. As illustrated in Figure 6, the values of k_1 will increase with the increasing of the temporal complexity, but the increase of k_1 will be slow because the temporal masking effect will play a part to reduce the difference of the distortion between different video sequences [9]. By fitting the curve in Figure 6, a natural logarithm function is introduced to capture the variations of k_1 as

$$k_1 = a_1 \cdot \ln(\delta_t) + b_1, \tag{7}$$

where δ_t is the estimated temporal complexity and parameters a_1 and b_1 are obtained by regression. Consequently, substituting (6) and (7) into (5) and transformming the form of the formula, the quality of the frame suffering from reference frame loss can be objectively achieved by

$$Q_l = \mathrm{Max}\left(\mathrm{Min}\left(Q_c - (Q_c - 1) \cdot (a_1 \cdot \ln(\delta_t) + b_1), 5\right), 1\right). \tag{8}$$

Considering the video quality is smaller than 5 and larger than 1, a constraint is added here to make the value of video quality in a rational interval.

With regard to the coding quality of each frame Q_c, it has been well estimated using the QP or bit-rate information extracted from the bit stream in our previous work [9, 11]. Therefore, we will not pay much attention on how to estimate the frame coding quality but just obtain it in virtue of the subjective test for simplicity.

4.2. Distortion Caused by Error Propagation. When the ith frame is subject to errors caused by reference frame loss, its subsequent frames are usually error-prone owing to the employment of the ith frame as the reference directly or indirectly through the use of temporal prediction, for example, frames P4, P7, and P8 in Figure 4. In this way, evaluation of the frame quality should take the factor of error propagation into account. To evaluate the frame distortion caused by error propagation, the sequences with a reference frame loss in training set TR1 are employed, and the frames with different error propagation lengths 1, 2, 4, 7, 10, and 13 are extracted from the impaired sequence and the original coded sequence, respectively, as shown in Table 2. The quality of each frame is obtained by subjective test, which forms the training set TR2.

The distortion of the frame due to error propagation D_e can be achieved by

$$D_e = Q_c - Q_e, \tag{9}$$

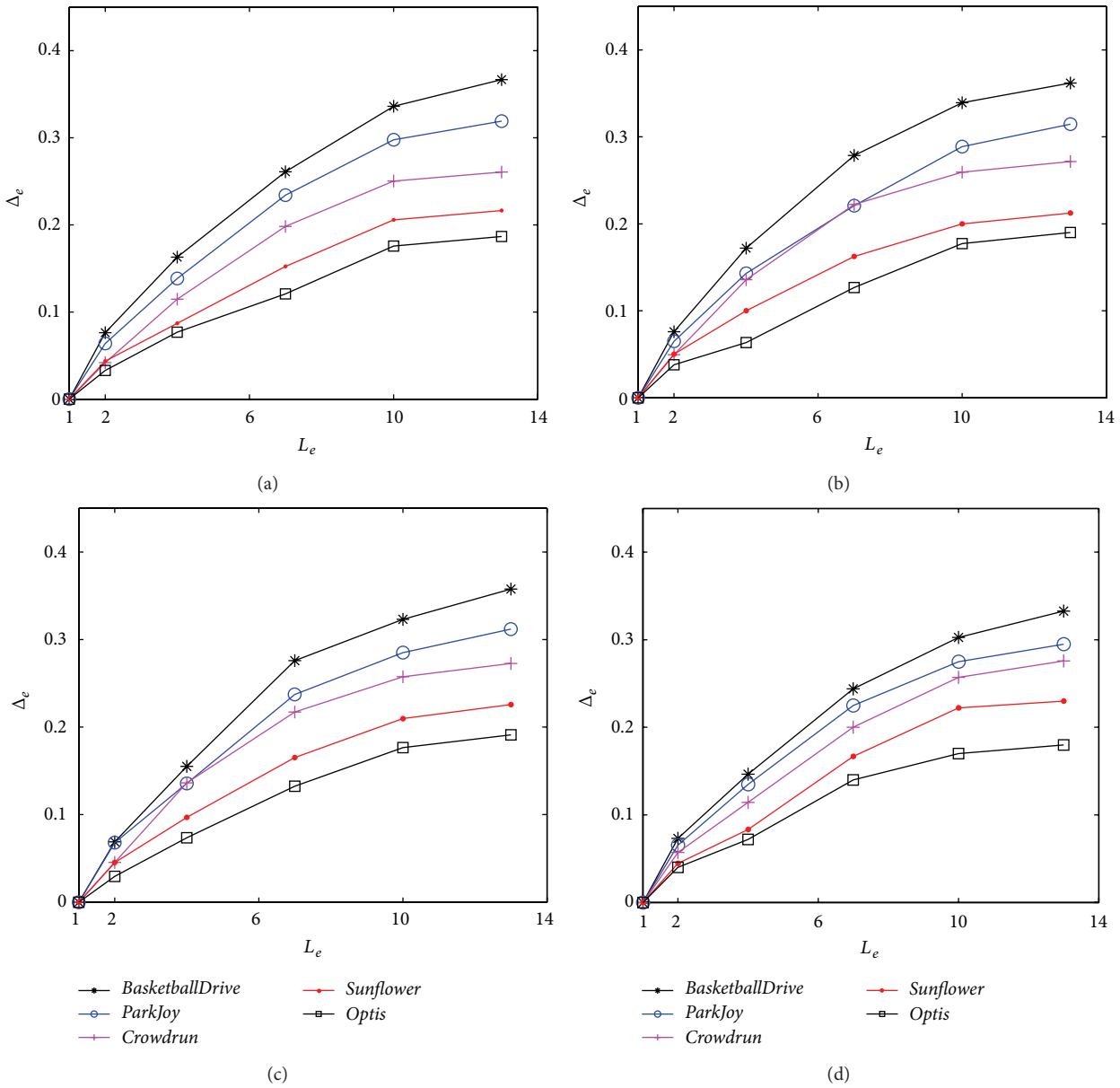

FIGURE 7: Relationship between length of error propagation and corresponding frame distortion under different QPs. (a) QP = 24, (b) QP = 30, (c) QP = 36, and (d) QP = 42.

TABLE 2: Lost frame index of each sequence.

Sequence	Index of the lost frame	Index of the frame to be evaluated
BasketballDrive	127	128, 129, 131, 134, 137, 140
ParkJoy	97	98, 99, 101, 104, 107, 110
Crowdrun	217	218, 219, 221, 224, 227, 230
Sunflower	37	38, 39, 41, 44, 47, 50
Optis	67	68, 69, 71, 74, 77, 80

where Q_e is the quality of the frame suffered from error propagation. Moreover, the normolized net increase in distortion caused by error propagation Δ_e for the ith frame can be calculated as

$$\Delta_e = \frac{D_e - D_l}{Q_c}. \tag{10}$$

Generally, D_e significantly depends on the length of error propagation L_e. This is because the distortion caused by packet loss will be aggravated by error propagation. Therefore, the relationships between L_e and Δ_e under different QPs are analyzed, as shown in Figure 7. It is discovered that the values of Δ_e for each sequence increase continuously with the increasing of L_e, which can be expressed as

$$\Delta_e = a_2 \cdot \left(\exp\left(b_2 \cdot (L_e - 1)\right) + \frac{c_2}{a_2} \right), \tag{11}$$

where the parameters a_2, b_2, and c_2 can be obtaind through regression by using the data in Figure 7. Moreover, it is also

Table 3: Values of a_2, b_2, and c_2 for each video sequence.

Sequence	δ_t	a_2	b_2	c_2	R-square	RMSE
BasketballDrive	8.81	−0.45	−0.16	0.46	0.981	0.037
ParkJoy	4.85	−0.34	−0.16	0.34	0.986	0.034
Crowdrun	3.53	−0.31	−0.17	0.30	0.982	0.031
Sunflower	2.73	−0.26	−0.15	0.26	0.991	0.018
Optis	1.86	−0.22	−0.16	0.22	0.987	0.029

Table 4: Values of b_2, d_2, and f_2 for each video sequence.

QP	b_2	d_2	f_2
24	−0.16	0.03	0.18
30	−0.17	0.04	0.16
36	−0.16	0.03	0.17
42	−0.15	0.03	0.18

obvious in Figure 7(a) that the frame distortion is closely related to the temporal complexity and the distortion will propagate fiercely when the temporal complexity is high. Therefore, to check the relationship between Δ_e and δ_t, the values of a_2, b_2, and c_2 for each sequence in Figure 7(a) are fitted, which are listed in Table 3. It can be found that there is a relative small difference between the values of b_2 for different sequences. However, there are relative large differences between the values of a_2 and c_2, respectively, for different sequences, and the values of a_2 and c_2 are nearly opposite with each other for the same sequence. Therefore, b_2 is considered as a constant and set as the average value of the video clips, while a_2 and c_2 are content-adaptive parameters which may be expressed using the content information like temporal complexity. Then, we will pay attention to predicting the parameters a_2 and c_2.

Figure 8 shows the relationship between c_2 and the estimated temporal complexity δ_t for different video sequence, where a linear model well approximates this relationship as

$$c_2 = d_2 \cdot \delta_t + f_2, \tag{12}$$

where constants d_2 and f_2 are obtained through regression using the data in Figure 8. Because a_2 is nearly opposite to c_2, it can be indicated as $-c_2$ in the proposed model. Substituting (11) and (12) into (10), D_e can be expressed as follows:

$$D_e = D_l + Q_c \cdot (d_2 \cdot \delta_t + f_2) \cdot (1 - \exp(b_2 \cdot (L_e - 1))). \tag{13}$$

More specially, when L_e is 1, D_e will be equal to D_l. That is, the frame with the reference frame loss is also the first frame of error propagation.

Moreover, to check D_e under different coding quality, the values of b_2, d_2, and f_2 under each QP are fitted using the same method as mentioned above, which are listed in Table 4. It can be found that the values of b_2, d_2, and f_2 are nearly constants under different QPs. That is to say, the D_e can be estimated in a consistent way using (13); even the QPs are various. Taking

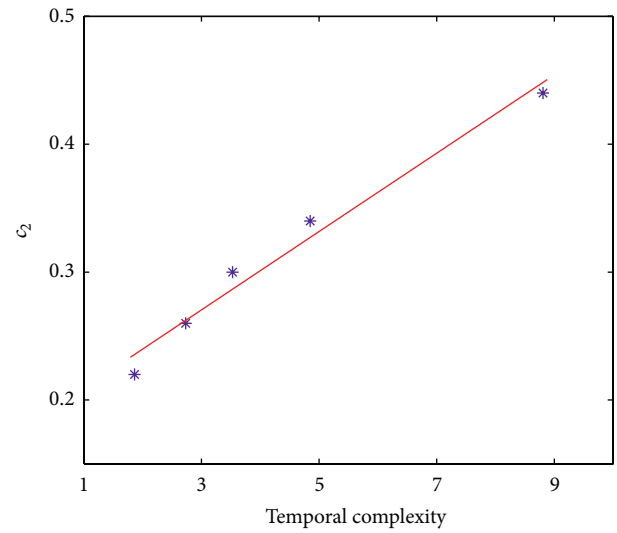

Figure 8: Relationship between temporal complexities δ_t and c_2.

(13) into (9), the quality of the frame suffering from error propagation can be evaluated as

$$Q_e = \mathrm{Max}(\mathrm{Min}(Q_c \cdot (1 - (d_2 \cdot \delta_t + f_2) \cdot (1 - \exp(b_2 \cdot (L_e - 1)))) - D_l, 5), 1). \tag{14}$$

4.3. Distortion Caused by Both Reference Frame Loss and Error Propagation. The impacts of reference frame loss and error propagation on frame quality have been analyzed above, respectively. Nevertheless, there is another case that the frame quality distortion is introduced by both the reference frame loss and error propagation together. When there has been a frame lost far ahead and at the same time the previous frame is also lost, the distortion introduced into the current frame should be considered as a joint influence of loss of reference frame and error propagation, such as the frame P6 shown in Figure 4.

To study the joint distortion, the sequences with a reference frame loss in training set TR1 are employed, and then a second frame loss occurs in the same GOP when the lengths of error propagation are 4 and 8 frames, respectively, as shown in Figure 9. The frames affected by both the reference frame loss and error propagation are extracted from the impaired sequence and original coded sequence, respectively. Then, the quality of each extracted frame is obtained by subjective test

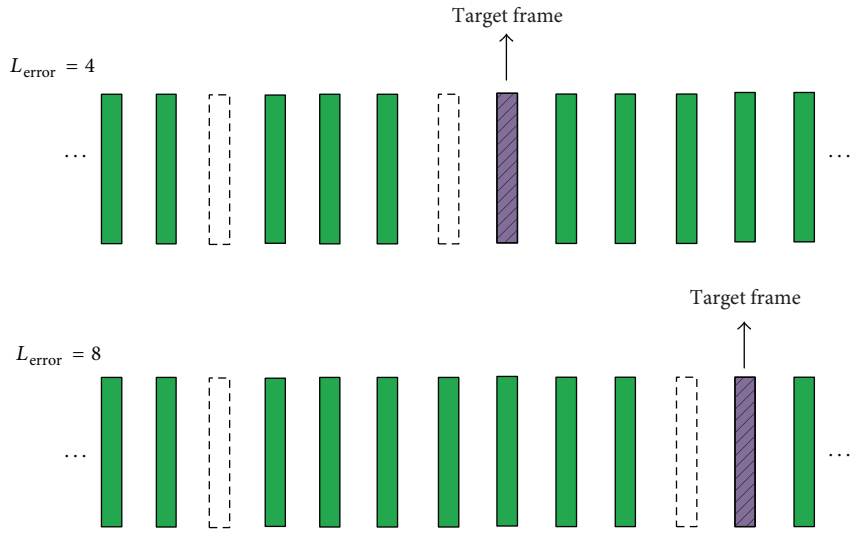

FIGURE 9: Multiple packet loss occurs with different error propagation length.

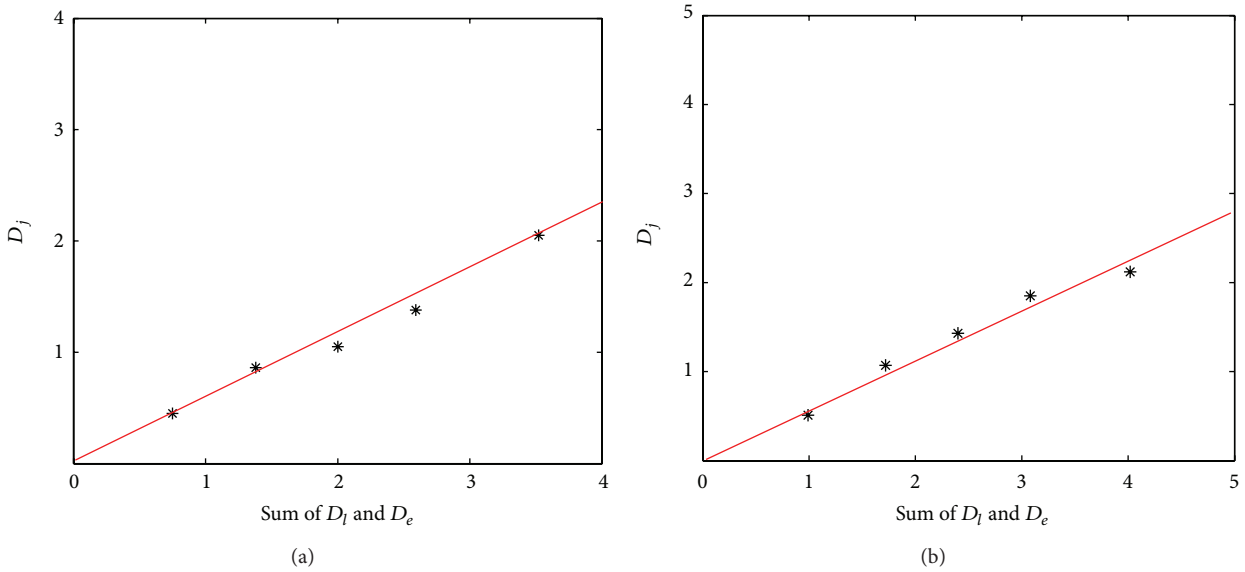

FIGURE 10: Relationship between the D_j and the sum of D_l and D_e under QP 20. (a) Error length is 4. (b) Error length is 8.

and all the values of frame quality form the training set TR3. The joint distortion of a frame D_j can be calculated as

$$D_j = Q_c - Q_j, \tag{15}$$

where Q_j is the quality of frame that is affected by the reference frame loss and error propagation and Q_c is the coding quality of each frame. Figure 10 gives the relationship between the D_j and the sum of D_l and D_e for the frames in different sequences, where the QP is 20 and the error length is 4 or 8. D_l and D_e are determined by (8) and (13), respectively. It is obvious that the joint distortion D_j is smaller than the sum of D_l and D_e and D_j is linearly related to the sum of D_l and D_e. Thus the joint distortion is estimated by

$$D_j = a_3 \cdot (D_l + D_e), \tag{16}$$

where the parameter a_3 is a proportional factor and can be obtained through regression. Then, the experimental values of a_3 with various QPs and lengths of error propagation are fitted using the same way and listed in Table 5. It is noticed that the parameter a_3 keeps almost constant in any case; therefore, in the experiment the value of a_3 is set as the average values. Consequently, combining (15) and (16), the frame quality affected by the joint loss of reference frame and error propagation can be calculated by the following equation:

$$Q_j = \text{Max}\left(\text{Min}\left(Q_c - a_3 \cdot (D_l + D_e), 5\right), 1\right). \tag{17}$$

4.4. Video Quality Evaluation by Temporal Pooling. Given the frame quality, the video quality can be assessed by

Table 5: Value of a_3 at the different QPs and lengths.

QP	Length of error propagation	a_3
20	4	0.56
	8	0.54
32	4	0.55
	8	0.54
38	4	0.53
	8	0.55
44	4	0.57
	8	0.56

using a pooling strategy, conventionally resorting to averaging or weighted averaging of the frame quality. Generally speaking, the quality degradation of videos in the temporal domain is mainly associated with the display duration of each frame. Accordingly, we have proposed a concept of "Quality Contribution" to describe the effect of each frame on the video quality, taking account of its spatial quality and display duration [36]. The temporal pooling is a function of the "Quality Contribution" of every frame weighted by the corresponding display duration.

Specifically, the "Quality Contribution" is derived from the logarithm relationship empirically found between the MOS and the display duration:

$$C_q = Q_f \cdot \left(p_1 + p_2 \cdot \delta_t' + p_3 \cdot \delta_t' \cdot \log(T)\right), \quad (18)$$

where C_q is the contribution of the frame, Q_f is the spatial quality of the frame, δ_t' is the normalized temporal complexity, defined as $\mathrm{Max}(\mathrm{Min}(\delta_t/\mathrm{Max}(\delta_t), 1), 0)$, T is the display duration of each frame, and p_1, p_2, and p_3 are parameters. The display duration of the frame is set as $T = \mathrm{Max}(T, 40)$, which means that the temporal discontinuity will be ignored when the displayed frame rate is not less than 25 fps; that is, the display frame interval $T = 40$ ms.

The perceived distortion is an effect of the distortion of successive frames, not just that of a single frame [37], therefore, a two-level temporal pooling method has been proposed in our earlier work [36]. It divides the video sequence into short-term groups of frames (GOFs) from the eye fixation level and then uses the quality of a GOF as the basic unit of long-term temporal pooling to evaluate the video quality. Given the contribution of each frame, where frame loss is well considered by the display duration, the quality of each GOF can be computed as

$$Q_{\mathrm{GOF}} = \frac{\sum_{n\in\mathrm{GOF}} \left(C_q(n) \cdot T(n)\right)}{\sum_{n\in\mathrm{GOF}} T(n)}, \quad (19)$$

where Q_{GOF} is the quality of a GOF. We have conducted subjective tests to find out the number of frames by which human observers can provide an accurate and steady quality judgment, that is, the duration of eye fixation. The results show that it is reasonable to select 0.8–1 second to get the steady response of video quality and the related number of frames can be used as the length of a GOF to make a quality judgment [36]. The method for segmenting GOFs in [36] is used in temporal pooling.

Next, a long-term temporal pooling will be presented combining the GOF quality to obtain the video quality. The quality of a significantly impaired frame would influence the perceptual quality of its neighboring frames, and the worst part of the video predominately determines the perceptual quality of the entire sequence. Therefore, we propose to select GOFs with quality computed by (19) lower than 75% of the average and compute the average quality of the selected GOFs as the video quality.

5. Performance Evaluation

In this section, both the performances of the proposed frame quality and video quality evaluation model will be validated. For the frame quality evaluation, the distortion caused by reference frame loss, by error propagation only, and by both of them will be estimated. For the video quality evaluation, the video quality will be evaluated under different packet loss rates.

Operational parameters given in Table 6 are trained following the descriptions in the previous sections by using the data in the training set. These parameters are fixed for all the experiments reported in this section. It is worth noting that the optimal values of parameters may be subjected to changes with other codecs or error concealment techniques.

The performance of the objective quality model is characterized by the prediction attributes: accuracy, linearity, monotonicity, and consistency. Accordingly, four metrics for performance evaluation suggested by the VQEG are used to evaluate the performance of the proposed model [38], that is, the root-mean-squared error (RMSE) for the accuracy, the Pearson correlation coefficient (PCC) for the linearity, the Spearman rank order correlation coefficient (SCC) for the monotonicity, and the outlier ratio (OR) for the consistency. The calculation of each statistical metric is performed along with its 95% confidence intervals. And the statistically significant differences among the performance metrics of various models are also calculated [39]. The larger value of PCC or SCC means the two variables are more linear or monotonic. The smaller value of RMSE or OR means the model's prediction is more accurate or consistent.

5.1. Performance Evaluation for the Frame Quality Assessment Model. To verify the efficiency and accuracy of the proposed frame quality assessment model, we use the test sequences sized of 640×360 (16 : 9) with different contents, respectively: *Cactus*, *PackC*, *Stockholm*, *Traffic,* and *Cyclists*. The video sequences chosen here are different from the sequences in training set. All the sequences are encoded by the ×264 codec with the frame rate 25 fps. The QP is set as 24, 32, 38, and 44, respectively, for all frames in a sequence. The GOP length is set as 30 with the structure of "IPPP," and the number of reference frame is set as 1. For all the sequences, the ZMEC error concealment strategy will be adopted at the decoder.

The NR QANV-PA model [9] and the CAPL model [11] are simulated for comparison purposes. The CAPL model employs the number of bits for coding I frames and P frames

Table 6: Parameters in the proposed method.

Parameter	a_1	b_1	b_2	d_2	f_2	a_3	p_1	p_2	p_3
Value	0.17	−0.02	−0.16	0.03	0.18	0.55	1	0.81	−0.51

Table 7: Lost frame index of each sequence.

Sequence	Index of the lost frame	Index of the frame to be evaluated
Cactus	91	92, 93, 96, 99, 103, 106
PackC	151	152, 153, 156, 159, 163, 166
Stockholm	61	62, 63, 66, 69, 73, 76
Traffic	121	122, 123, 126, 129, 133, 136
Cyclists	31	32, 33, 36, 39, 43, 46

Table 8: Performance comparison between objective data and subjective data.

Data set V1	PCC	RMSE	SCC	OR
Proposed model	0.983	0.204	0.968	0.046
NR QANV-PA model	0.944	0.291	0.923	0.058
CAPL model	0.926	0.347	0.895	0.058
Baseline model	0.951	0.272	0.928	0.054

to estimate the temporal complexity, and the NR QANV-PA model uses QP and bit rate to predict the temporal complexity. Both models evaluate the frame quality affected by the packet loss combined with the temporal complexity. Moreover, we also use the average MVM as the temporal complexity to evaluate the frame quality, which is named "baseline model" in this paper.

Firstly, the performance of the proposed model is evaluated with regard to a single lost packet. In this situation, two kinds of frames are affected by the packet loss, namely, frames affected by reference frame loss and frames affected by error propagation only. In the experiments, the lost frame indexes of each video sequence are set randomly, and the length of error propagation are chosen as 1, 2, 5, 8, 12, and 15, respectively, as shown in Table 7. Twenty subjects participate in the validation test, including 9 females and 11 males. The procedures of subjective test are the same as those in the training set, as described in Section 4.1. These sets of data constitute the validation set V1.

Table 8 lists the detailed performance of each model using the data of V1 ($P < 0.05$ in statistical significance test). It is observed that the proposed model demonstrates a superior MOS prediction performance to that of the NR QANV-PA, the CAPL, and the baseline models. Therefore, the proposed model can well estimate the frame quality under the condition of single packet loss in a GOP.

Secondly, the experiment with two frames loss in a GOP is carried out to check the performance of the proposed evaluation method for the joint distortion. The lost frame indexes are randomly set, as shown in Table 9. The error propagation length of the second frame loss is set as 1, 2, 5, 8, and 12, respectively. All the subjective values constitute the validation set V2.

Table 9: Lost frame index of each sequence.

Sequence	Index of the lost frame	Index of the frame to be evaluated
Cactus	91, 94	95, 96, 99, 102, 106
	91, 98	99, 100, 103, 106, 110
PackC	151, 153	154, 155, 158, 161, 165
	151, 157	158, 159, 162, 165, 169
Stockholm	61, 64	65, 66, 69, 72, 76
	61, 67	68, 69, 72, 75, 79
Traffic	121, 125	126, 127, 130, 133, 137
	121, 128	129, 130, 133, 136, 140
Cyclists	41, 43	44, 45, 48, 51, 55
	41, 47	48, 49, 52, 55, 59

Table 10: Performance comparison between objective data and subjective data.

Data set V2	PCC	RMSE	SCC	OR
Proposed model	0.941	0.283	0.927	0.047
NR QANV-PA model	0.917	0.355	0.881	0.063
CAPL model	0.843	0.403	0.824	0.095
Baseline model	0.922	0.368	0.907	0.054

The corresponding values of PCC, RMSE, SCC, and OR are summarized in Table 10 ($P < 0.05$), which confirms the advantage of proposed frame quality evaluation model over the NR QANV-PA model, the CAPL model, and the baseline model in terms of all performance criteria when multiple packet loss is presented. Therefore, the proposed frame quality estimation model has an excellent performance under different packet loss situations.

5.2. Performance Evaluation for the Video Quality Assessment Model. In this subsection, we will check the performance of the proposed video quality assessment model. A data set consisting of the *BasketballDrive*, *Crowdrun*, *Sunflower*, *ParkJoy*, *Optis*, *Cactus*, *PackC*, *Stockholm*, *Traffic*, and *Cyclists* sequences is used for performance validation. The resolution of each sequence is 640 × 360 (16 : 9). All the sequences are encoded by the ×264 codec with the frame rate 25 fps. The length of each video sequence is 8 s. The QP is set as 24, 32, 38, and 44, respectively, for all frames in a sequence. The GOP length is set as 30 with the structure of "IPPP," and the number of reference frame is set as 1. For the real-time video service, the RTP/UDP/IP protocol stacks are usually employed to transmit the encoded bit streams. In simulations, the compressed bit streams are packetized following the

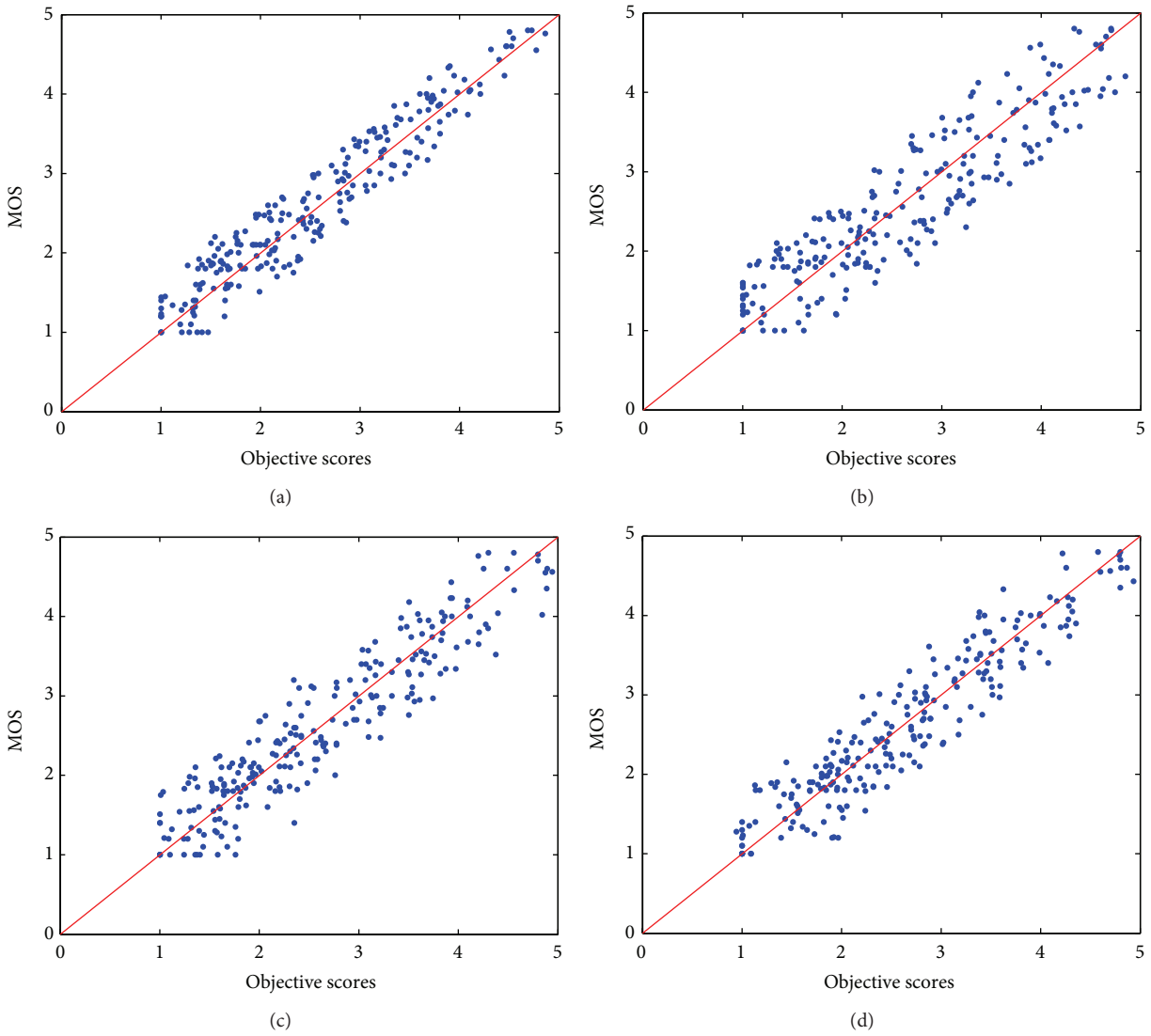

Figure 11: Scatter plots of the MOS versus the objective scores. (a) Proposed model, (b) P.1201.1 model, (c) NR QANV-PA model, and (d) baseline model.

format defined by "RTP Payload Format for H.264" [40], where the maximum size of the packet is set as 1500 bytes. Different packet loss rates are imposed on the test video stream to simulate lossy channel conditions. The packet loss rates used in the experiments are set as 0.5%, 1%, 2%, 3%, and 5%, where the 4-state Markov model is employed to model the packet loss distribution in IP networks [41]. In the experiment, the ZMEC error concealment strategy is selected as an option provided by the H.264 decoder FFmpeg 0.4.9 [42].

The subjective tests are carried out following the guidelines specified by VQEG. Twenty nonexpert viewers participate in this test, including 9 females and 11 males. A 5-point scale is used to obtain the MOS of reconstructed sequences. Different from the frame quality evaluation tests, the video quality subjective test is based on a single stimulus procedure according to absolute category rating (ACR) [29, 35]. The subjects will rate the video quality instead of the frame quality. All the subjective data constitute the validation set V3.

To better check the performance of the proposed model, the P.1201.1 model [7], the NR QANV-PA model [9], and the baseline model are simulated for comparison purposes. The P.1201.1 model uses the statistical parameters to evaluate the video quality, such as the average impairment rate of frame, impairment rate of video stream, and packet-loss event frequency. The NR QANV-PA model and the baseline model integrate the frame quality to video quality by temporal pooling scheme, using different parameters to evaluate the motion activity of the video content.

Figure 11 gives the scatter plots of the objective scores and the subjective scores. A linear relationship is clearly shown in Figure 11(a) between the prediction of proposed model and MOS values using data of V3 ($P < 0.05$ in statistical significance test). It is obvious in Figure 11 that the proposed model demonstrates a superior MOS prediction performance to that of the P.1201.1 model, the NR QANV-PA model, and the baseline model. Moreover, the detailed performance of each model is listed in Table 11. It can be concluded that the

Table 11: Performance comparison between objective data and subjective data.

Data set V3	PCC	RMSE	SCC	OR
Proposed model	0.959	0.292	0.943	0.046
P.1201.1 model	0.909	0.448	0.897	0.078
NR QANV-PA model	0.931	0.373	0.928	0.062
Baseline model	0.942	0.339	0.935	0.056

proposed model outperforms the other models under the condition of various packet loss rate.

6. Conclusions

In this paper, an effective method is proposed to evaluate the video quality suffering from packet loss. The principal contributions of this paper are concluded as follows. (1) The novel temporal complexity is proposed based on the character of HVS, which can distinguish the degree of motion activity effectively. (2) Considering the different effects of packet loss on a frame, the frames in a video sequence are classified into four categories, and the distortion of each kind of frame is accurately evaluated using different influential factors. The frame quality is then integrated to obtain the video quality by a two-level temporal pooling scheme, including the eye fixation short-term level and the long-term level in terms of human visual perception. Experimental results show that the performance of the proposed model is excellent in both frame quality evaluation and video quality evaluation.

The proposed video quality assessment method is well suited to the real-time mobile video applications, such as the videophone and mobile IPTV. The current model only concerns the motion activity of the video content but does not consider the influence of the spatial complexity to the video quality. Further work should include the development of a more accurate way to describe the characteristics of the video content, taking account of the influence of the moving direction and the spatial complexity. The extension of the proposed work is currently under investigation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (K5051301020, K5051301027) and the 111 Project (B08038).

References

[1] J. Maisonneuve, M. Deschanel, J. Heiles et al., "An overview of IPTV standards development," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 315–328, 2009.

[2] Q. Qi, Y. Cao, T. Li, X. Zhu, and J. Wang, "Soft handover mechanism based on RTP parallel transmission for mobile IPTV services," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2276–2281, 2010.

[3] O. Verscheure, P. Frossard, and M. Hamdi, "User-oriented QoS analysis in MPEG-2 video delivery," *Real-Time Imaging*, vol. 5, no. 5, pp. 305–314, 1999.

[4] J. M. Boyce and R. D. Gaglianello, "Packet loss effects on MPEG video sent over the public internet," in *Proceedings of the ACM International Conference on Multimedia*, pp. 181–190, ACM, Bristol, UK, 1998.

[5] M. Garcia and A. Raake, "Parametric packet-layer video quality model for IPTV," in *Proceedings of the 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA '10)*, pp. 349–352, 2010.

[6] ITU-T Recommendation G.1070, *Opinion Model for Video-Telephony Applications*, ITU, Geneva, Switzerland, 2007.

[7] ITU-T Recommendation P.1201.1, *Parametric Non-Intrusive Assessment of Audiovisual Media Streaming Quality-Lower Resolution Application Area*, ITU-T, Geneva, Switzerland, 2012.

[8] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal on Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, 2009.

[9] F. Z. Yang, S. Wan, Q. P. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1544–1554, 2010.

[10] F. Li, D. Zhang, and L. Wang, "Packet importance based scheduling strategy for H.264 video transmission in wireless networks," *Multimedia Tools and Applications*, 2014.

[11] F. Yang, J. Song, S. Wan, and H. R. Wu, "Content-adaptive packet-layer model for quality assessment of networked video services," *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 672–683, 2012.

[12] H.-X. Rui, C.-R. Li, and S.-K. Qiu, "Evaluation of packet loss impairment on streaming video," *Journal of Zhejiang University: Science*, vol. 7, no. 1, pp. 131–136, 2006.

[13] O. Hadar, M. Huber, R. Huber, and R. Shmueli, "Quality measurements for compressed video transmitted over a lossy packet network," *Optical Engineering*, vol. 43, no. 2, pp. 506–520, 2004.

[14] H. Wu and K. Rao, *Digital Video Image Quality and Perceptual Coding*, CRC Press, Boca Raton, Fla, USA, 2006.

[15] H.-T. Quan and M. Ghanbari, "Temporal aspect of perceived quality in mobile video broadcasting," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 641–651, 2008.

[16] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[17] S. Winkler and P. Mohandas, "The evolution of video quality measurement: from PSNR to hybrid metrics," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.

[18] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 327–334, 2004.

[19] S. Wolf and M. Pinson, "Video quality measurement techniques," National Telecommunications and Information Administration (NTIA) Report 02-392, 2002.

[20] T. Liu, Y. Wang, J. M. Boyce, H. Yang, and Z. Wu, "A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts," *IEEE Journal on Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 280–293, 2009.

[21] ITU-T Recommendation J.246, *Perceptual Visual Quality Measurement Techniques for Multimedia Services over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference*, ITU-T, Geneva, Switzerland, 2008.

[22] T. Liu, H. Yang, A. Stein, and Y. Wang, "Perceptual quality measurement of video frames affected by both packet losses and coding artifacts," in *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEx '09)*, pp. 210–215, IEEE, San Diego, Calif, USA, July 2009.

[23] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Human Vision and Electronic Imaging X*, vol. 5666 of *Proceedings of SPIE*, pp. 149–159, March 2005.

[24] S. Kanumuri, S. G. Subramanian, P. C. Cosman, and A. R. Reibman, "Predicting H.264 packet loss visibility using a generalized linear model," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '06)*, pp. 2245–2248, Atlanta, Ga, USA, October 2006.

[25] D. Hands, O. V. Barriac, and F. Telecom, "Standardization activities in the ITU for a QoE assessment of IPTV," *IEEE Communications Magazine*, vol. 46, no. 2, pp. 78–84, 2008.

[26] S. Qiu, H. Rui, and L. Zhang, "No-reference perceptual quality assessment for streaming video based on simple end-to-end network measures," in *Proceedings of the International Conference on Networking and Services (ICNS '06)*, IEEE, Washington, DC, USA, July 2006.

[27] S. Argyropoulos, A. Raake, M.-N. Garcia, and P. List, "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *Proceedings of the 3rd International Workshop on Quality of Multimedia Experience (QoMEX '11)*, pp. 31–36, Mechelen, Belgium, September 2011.

[28] N. Staelens, G. van Wallendael, K. Crombecq et al., "No-reference bitstream-based visual quality impairment detection for high definition H.264/AVC encoded video sequences," *IEEE Transactions on Broadcasting*, vol. 58, no. 2, pp. 187–199, 2012.

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

[30] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 286–298, 2011.

[31] S. Wolf and M. Pinson, "A no reference (NR) and reduced reference (RR) metric for detecting dropped video frames," in *Proceedings of the 4th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2009.

[32] R. Feghali, F. Speranza, D. Wang, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 441–446, 2007.

[33] S. Wan, F. Yan, and Z. Xie, "Evaluation of video quality degradation due to packet loss," in *Proceedings of the 18th International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS '10)*, pp. 1–4, IEEE, Chengdu, China, December 2010.

[34] Y. Yang, X. Wen, W. Zheng, L. Yan, and A. Zhang, "A no-reference video quality metric by using inter-frame encoding characters," in *Proceedings of the 14th International Symposium on Wireless Personal Multimedia Communications (WPMC '11)*, pp. 1–5, Brest, France, October 2011.

[35] TU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," Geneva, Switzerland, 2006.

[36] F. Yang, S. Wan, Y. Chang, and Q. Xie, "Frame-rate adaptive temporal pooling for video quality assessment," in *Proceedings of the 4th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2009.

[37] A. Rohaly, J. Lu, N. Franzen, and M. Ravel, "Comparison of temporal pooling methods for estimating the quality of complex video sequences," in *Human Vision and Electronic Imaging IV*, vol. 3644 of *Proceedings of SPIE*, pp. 218–226, San Jose, Calif, USA, 1999.

[38] ITU-T Recommendation J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," Geneva, Switzerland, 2008.

[39] Multimedia Phase I Testplan Version 1.21, "Multimedia group test plan," 2008, http://www.its.bldrdoc.gov/vqeg/projects/multimedia-phase-i/multimedia-phase-i.aspx.

[40] IETF RFC3984, *RTP Payload Format for H.264 Video*, 2005.

[41] ITU, "Packet loss distributions and packet loss models," ITU Report COM12-D97-E, 2003.

[42] FFmpeg, 2013, http://sourceforge.net/projects/ffmpeg.