Feature fusion quality assessment model for DASH video streaming

Hong Zhang1 | Fan Li1 | Zhisheng Yan2

1 School of Information and Communications Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, China
2 Department of Computer Science, Georgia State University, Atlanta, Georgia, USA

Correspondence
Fan Li, School of Information and Communications Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049, China. Email: lifan@mail.xjtu.edu.cn

Abstract
Dynamic Adaptive Streaming over HTTP (DASH) employs the flexible rate adaptation scheme to combat with time-varying channel conditions. In addition to compression impairment, DASH video streaming suffers from transmission impairment, such as rate switches and stalling events. Both of them severely degrade users’ Quality of Experience (QoE). Herein, an assessment model is established for DASH video streaming by directly jointing multiple QoE influential factors, which quantify impairments resulting from the compression and transmission. To demonstrate the influence of video content characteristics on users’ QoE, spatio-temporal content perceptual features are employed to represent the compression impairment. When reflecting temporal characteristics, a novel motion vector padding method is proposed to quantify the influence of intra macroblock on the human visual system. The proposed model is evaluated on a newly public Waterloo SQoE-III database, which is available for DASH video streaming. Experimental results demonstrate that the authors’ model outperforms the comparative models and owns the strong generalization ability to different video contents. Moreover, the proposed model is statistically superior to the existing models.

1 | INTRODUCTION

With the increasing popularity of video streaming, the expectation of user’s Quality of Experience (QoE) is also on the rise. Benefiting from the development of Dynamic Adaptive Streaming over HTTP (DASH), which is an international standard issued by the 3rd Generation Partnership Project (3GPP) [1], video quality improves from the technical context. However, because end users are the ultimate receivers of the video stream, it is a crucial issue to guarantee users’ QoE by assessing/predicting their experience or how the impaired degree of distorted video sequence affects human visual perception. It benefits the QoE-driven strategies of resource allocation and rate adaption for DASH video streaming. Therefore, it becomes a hot research topic to establish an objective QoE assessment model assessing/predicting the user’s QoE.

Objective QoE assessment model can automatically assess/predict users’ QoE for distorted video sequences. The critical point of establishing the QoE assessment model is to select QoE influential factors and searches out the relationship between them with users’ Mean Opinion Score (MOS). The process is termed as QoE modelling. However, the effectiveness of objective quality assessment model needs the verification of subjective tests regarded as the ground truth. In the following, we outline three aspects related to the QoE modelling for adaptive streaming, that is subjective assessment database, transmission-oriented objective assessment model, and objective assessment model considering impairments of compression and transmission.

Subjective tests directly obtain users’ QoE by viewers viewing the distortion video sequence under a specific test condition [2]. It is the most direct way to quantify users’ QoE, which has recommended in BT.500. Although subjective tests are the most reliable and straightforward method to assess users’ QoE, it is time-consuming and expensive. Moreover, due to the specific and restricted test environment, it cannot be widely applicable. Therefore, objective QoE assessment models are developed based on several subjective assessment databases for HTTP-based adaptive streaming [3–6]. However, test video sequences are all with the limited and hand-crafted distortion types in
these databases. For obtaining representative distorted video sequences to accommodate diverse application scenarios, a new publicly available database was established [7], in which test video sequences are generated by the most realistic kind of simulated test-bed.

One type of assessment models can not quantify the impact of video compressed impairment on the human visual perception, which is termed as a transmission-oriented objective assessment model. According to the process of obtaining the mapping function, these models can be classified into parametric assessment models [8–10] and learning-based assessment models [11]. The parametric assessment models own an explicit mapping form, which is determined before the fitting process with unknown parameters, between QoE influential factors and user’s QoE. Different exponent mapping functions, which accounted for the influence of stalling length and stalling number on human perception, were proposed [8, 9]. In [10], users’ QoE was assessed by a linear mapping function from QoE influential factors, such as initial rebuffering time, stalling frequency, and stalling length. Due to the complexity of HVS and the diversity of QoE influential factors, the mapping form cannot be well determined in advance, especially for multi-dimensional QoE influential factors. In this way, the learning-based assessment models can obtain the mapping relationship by mass data without determining the exact mapping form in advance. Two learning-based assessment models were established [11]. One is a multi-stage approach, in which the outputs of nine Hammerstein–Wiener (HW) models are employed to train another Wiener model for obtaining the mapping function. The other is a multi-learner approach, in which the Support Vector Regression (SVR) model is the alternative of the Wiener model. Nevertheless, Time-varying Video Subjective Quality (TVSQ) [12] model based on the Hammerstein–Wiener model was proposed to predict/assess the time-varying user’s QoE caused by rate variations and compression. It demonstrated that compression impairment is a key aspect to determine the user’s QoE. Moreover, it was validated by the subjective test [3].

Another type of assessment models measures users’ QoE by considering the influence of impairments resulting from compression and transmission, which is termed as an objective assessment model considering impairments of compression and transmission concurrently. Existing video quality assessment (VQA) metrics, such as PSNR, STRED [13], SSIM [14], SSIMplus [15], and VQM [16], are developed for the evaluation of streaming-based User Datagram Protocol (UDP) connection. They are used to quantify the quality of compression impairment. However, due to the flexibility and scalability, these metrics are only used in DASH streaming without rate switches and stalling events.

Therefore, researchers have focused on objective assessment models for DASH video streaming, such as parametric assessment models [7, 17–22], learning-based assessment models [23, 24], and hybrid assessment models [4, 25, 26]. In [17], the model quantified the impact of QoE influential factors, such as initial delay, stall length, level variation, and motion information. Linear regression models were established [18–21], which account for the influence of video quality, quality variation, stalling length, and stalling number. The QoE continuum model [22] was proposed based on an exponential smoothing factor. The linear function of the Quantization Parameter (QP) quantifies the video representation quality during normal playback in terms of the frame. Video ATLAS [23] was a learning-based model, which took VQA metric, rebuffering-aware features, memory-related features, and impairment duration feature as inputs to predict user’s QoE. In [24], the authors viewed video quality prediction as a time-series forecasting to establish three quality assessment models on LIVE-Netflix QoE database [6], such as open-loop NARX model, RNN model, and HW model. SVR-QoE [25] is composed of two sub-models, which are a learning-based prediction model during video playing normally and a parametric analytical model during stalling events.

Although many efforts have been made to develop objective QoE assessment models for DASH video streaming, there still exists some shortcomings:

1. **Indirectly Mapping the Influence of Compression and Transmission Impairments on Human Perception:** Existing models compute users’ QoE resulted from compression impairment in the unit of segment in advance, such as SQI [3], SVR-QoE [4], TVSQ [12], and Video ATLAS [23]. When stalling events occur, users’ QoE is obtained by mapping the previously computed users’ perception caused by compression and stalling QoE influential factor. However, due to the complexity, users cannot accurately distinguish the compression impairment and transmission impairment, nor can they respectively give the perceptual degree of artefacts. Actually, users’ perception is influenced by the objective property of video sequence caused by the compression and stalling QoE influential factors.

2. **Indirectly Mapping the Influence of Compression and Transmission Impairments on Human Perception:** Existing models compute users’ QoE resulted from compression impairment in the unit of segment in advance, such as SQI [3], SVR-QoE [4], TVSQ [12], and Video ATLAS [23]. When stalling events occur, users’ QoE is obtained by mapping the previously computed users’ perception caused by compression and stalling QoE influential factor. However, due to the complexity, users cannot accurately distinguish the compression impairment and transmission impairment, nor can they respectively give the perceptual degree of artefacts. Actually, users’ perception is influenced by the objective property of video sequence caused by the compression and stalling QoE influential factors.

3. **Indirectly Mapping the Influence of Compression and Transmission Impairments on Human Perception:** Existing models compute users’ QoE resulted from compression impairment in the unit of segment in advance, such as SQI [3], SVR-QoE [4], TVSQ [12], and Video ATLAS [23]. When stalling events occur, users’ QoE is obtained by mapping the previously computed users’ perception caused by compression and stalling QoE influential factor. However, due to the complexity, users cannot accurately distinguish the compression impairment and transmission impairment, nor can they respectively give the perceptual degree of artefacts. Actually, users’ perception is influenced by the objective property of video sequence caused by the compression and stalling QoE influential factors.
establish a QoE assessment model from the holistic perspective (Section 2).

(2) We propose a novel MV padding approach, which employs the prediction direction and partition mode of block to pad the encoding. Based on the completed MV field, the motion vector scale and motion vector direction are computed to quantify the characteristics of temporal information for video sequence (Section 3).

The remainder of this article is organized as follows. We present the model architecture in Section 2. In Section 3, we detail our novel motion vector padding approach. The method for representing the characteristics of the compressed video sequence is demonstrated in Section 4. Model evaluation method is presented in Section 5. Extensive experimental results and analysis are reported in Section 6. In Section 7, we discuss the applications of our proposed model. Conclusions are drawn in Section 8.

2 | MODEL ARCHITECTURE

In this section, we instance our quality assessment model. When the video sequence is transmitted to the user from the video server, it will be impaired by compression and transmission, which can be quantified by multiple QoE influential factors. Because of the complexity of HVS, the user cannot completely distinguish the specific impairment type and accurately give the perceptual degree for one single impairment. Actually, users’ QoE is directly determined by the characteristics of a video sequence, which is caused by the impairment of compression and transmission. Due to the difficulty in determining the explicit mapping function for multiple QoE influential factors, we employ the learning-based approach to treat multiple QoE influential factors as a whole to derive the mapping function. The architecture of our proposed model is shown in Figure 1.

For the video sequence $i$, our model needs to learn the map function between input features $x_i = (C, B, T, MVS, MVD, I_t, P_r, C_r, C_b, M_b)$ and the subjective Mean Opinion Score (MOS) $y_i$. The dimension of the input features is 30 for every video sequence. Among them, five perceptual features account for the compression impairment and five quantify the transmission impairment. The former is the five-dimensional vector for every feature, such as Contrast ($C$), Blur ($B$), Texture ($T$), Motion Vector Scale ($MVS$) and Motion Vector Direction ($MVD$), for distinguishing the difference of five continuous segments to human visual perception. Meanwhile, these features can be classified into two categories: spatial perceptual features and temporal perceptual features. The latter includes initial buffering time ($I_t$), stalling percentage ($P_r$), stalling count ($C_r$), bitrate switch count ($C_b$), and average bitrate switch magnitude ($M_b$). It is a one-dimensional scalar for every feature.

A variety of kernel functions can be utilized to learn the mapping function between multiple QoE influential factors and the subjective MOS. In our work, we chose $\varepsilon-SVR$ to obtain the regression mapping function [27]. The advantage can be explained from two aspects: a) the limited size of the subjective database for DASH video streaming [7], while the strong generalization ability of SVR to combat with small sample data; and b) the non-linearity of $\varepsilon-SVR$ better representing the non-linearity of HVS by mapping features into a high-dimensional space.

3 | NOVEL MOTION VECTOR PADDING APPROACH

Temporal information plays an important role in determining the user’s QoE [17, 28], so the objective metric is needed to quantify it. In this section, we detail two components, that is padding based on the prediction direction and post-processing based on the partition mode, for our novel MV padding approach. And it is the preliminary for representing the compression impairment when computing the temporal perceptual features in our proposed model.

In our work, we just extract MV for P and B frames. The reasons can be stated as follows. In Waterloo SQoE-III database [7], the scene-cut is removed during encoding, and the GOP size is set as twice frame rate. Meanwhile, the length of segment is 2 s, for which the video sequence can be independently encoded
and decoded. Consequently, one segment just contains one I-frame. Another reason is that the computation of our perceptual features is in terms of a single segment. For all sub-blocks in I-frame, only the subsequent frame information can be employed to estimate MV. However, it already contributed to feature computation. Thus, we ignore the estimation of MV in I-frames.

Due to the reasonability of reflecting temporal information and the feature of easy extraction during encoding, MV is often employed to measure the temporal information of video sequences. However, because of the diverse video content characteristics and the flexible modes selection during encoding, there will be I-MBs in P-frames and B-frames. Although I-MBs have no MVs, the content contained in I-MBs likely owns motion to human visual perception. When the temporal correlation between the current encoding MB and the previously encoded frame is weak, it usually means that the new content appears. So, the most similar MB cannot be found in the previously encoded frame, and current encoding MB will be encoded as I-MB based on Rate Distortion Optimization (RDO). Moreover, if the new content gradually appears, the following MBs after current encoding I-MB are possibly encoded by I-MBa. It results from the strong spatial correlation between the current MB and its neighbours. Therefore, in this article, based on the prediction direction and partition mode, we develop our novel MV padding approach for I-MB in P-frames or B-frames.

3.1 Padding based on prediction direction

There are two partition modes for I-MB, such as 4×4 sub-block and 16×16 MB. The former is applied in the region of rich details, and the latter is employed in the flat region. Note that, H.264/AVC standard provides nine optional prediction modes for 4×4 sub-block partition pattern, in which there are eight-directional predictions, which are demonstrated in Figure 2, and four optional prediction modes for the 16×16 MB partition pattern. We show the padding approach for 4×4 sub-block and 16×16 MB below, respectively.

3.1.1 4×4 sub-block

For 4×4 sub-block prediction mode, the MB is divided into sixteen 4×4 sub-blocks firstly, and every 4×4 sub-block is regarded as an individual. The prediction decision of the current 4×4 sub-block is made by ever encoded neighbouring pixel values, such as A–M, respectively. Except for Mode 2, the mean of pixel values about A–D and I–L is employed, when the neighbouring pixels are all available. And the other eight-directional modes have their own prediction direction. The pixel prediction value is approximately equal to the neighbouring pixels of the already encoded pixel value in the same direction. Benefitting from the directional prediction, we assume the 4×4 sub-block along the same direction has the approximate motion information. Therefore, we propose a weight-sum-based approach to pad the MV for I-MB in P and B frames. There are two cases to pad the MV.

Case 1: The adjacent reference coded 4×4 sub-blocks are inter-coded MBs, which means that the MVs of reference sub-blocks all exist.

The concrete padding approach for the current 4×4 sub-block is determined by its prediction direction, which is stated below shown in (1–9), and \((M_x', M_y')\) is the vertical and horizontal MV value for 4×4 sub-block. We employ the MV of neighbouring 4×4 sub-blocks, which are located in MB \(B_{00}–B_{33}\), to obtain the MV of the current 4×4 sub-block. The weights are determined by the proportion of reference encoded sub-block pixel value in the current 4×4 sub-block. The padding approach is the same for I-MBs in P-frames and B-frames. The only difference lies in that the prediction direction should be confirmed for B-MB firstly. If it is the forward prediction or backward prediction, the MV can be employed directly. If it is bi-direction prediction, the MV is the value difference of the forward MV and the backward MV.

\[
\begin{align*}
&M_{X}(i), M_{Y}(i) =
&M_{X}(i), M_{Y}(i) \\
&\text{b) Mode 1 (Horizontal):} \\
&(M_{X}(i), M_{Y}(i)) = (M_{X}(i), M_{Y}(i)) \\
&\text{c) Mode 2 (DC):} \\
&M_{X}(i), M_{Y}(i) =
&M_{X}(i), M_{Y}(i) \\
&\text{d) Mode 3 (Diagonal Down Left):} \\
&(M_{X}(i), M_{Y}(i)) = \frac{3}{8} (M_{X}(i), M_{Y}(i)) + \frac{5}{8} (M_{X}(i), M_{Y}(i)) \\
&\text{e) Mode 4 (Diagonal Down Right):} \\
&(M_{X}(i), M_{Y}(i)) = \frac{3}{8} (M_{X}(i), M_{Y}(i)) + \frac{3}{8} (M_{X}(i), M_{Y}(i))
\end{align*}
\]
f) Mode 5 (Vertical Right):

\[
\langle M_X(i), M_Y(i) \rangle = \frac{11}{16} \langle M_X(0), M_Y(0) \rangle + \frac{1}{8} \langle M_X(2), M_Y(2) \rangle + \frac{3}{16} \langle M_X(3), M_Y(3) \rangle
\]

(6)

g) Mode 6 (Horizontal Down):

\[
\langle M_X(i), M_Y(i) \rangle = \frac{1}{8} \langle M_X(0), M_Y(0) \rangle + \frac{11}{16} \langle M_X(2), M_Y(2) \rangle + \frac{3}{16} \langle M_X(3), M_Y(3) \rangle
\]

(7)

h) Mode 7 (Horizontal Left):

\[
\langle M_X(i), M_Y(i) \rangle = \frac{11}{16} \langle M_X(0), M_Y(0) \rangle + \frac{5}{16} \langle M_X(1), M_Y(1) \rangle
\]

(8)

i) Mode 8 (Horizontal Up):

\[
\langle M_X(i), M_Y(i) \rangle = \langle M_X(2), M_Y(2) \rangle
\]

(9)

Case 2: The adjacent reference coded 4×4 sub-blocks are intra MBs, which means that the MVs of reference 4×4 sub-blocks do not exist.

The padding approach employed in Case 1 is also applicable to Case 2. The reason is stated as follows. In the H.264/AVC standard, the encode types, partition modes, and prediction modes direction for MB are determined by RDO. This means that the distortion cost is minimal among all encode types and modes when the current MB is encoded by I-MB. Therefore, the most similar 4×4 sub-block for the current 4×4 sub-block exists in its neighbouring MB B_0–B_3. In this way, if the current 4×4 sub-block belongs to I-MB, its MV has been estimated before. The estimated MV of current 4×4 sub-block can be regarded as the existent MV when it is utilized as the reference 4×4 sub-block. This can be employed when there are continuous I-MBs one by one, which is shown in Figure 3. Note that the red draw is intra-MB, and the green is inter-MB.

However, when the reference encoded 4×4 sub-block for the current 4×4 sub-block is the member of inter-MB, it means that the current MB presents a new emerging content, which is shown in Figure 4. If all reference coded 4×4 sub-blocks own zero MVs, the padded MV for the current 4×4 sub-block would be zero MV. Although it is concerned by a human, its motion information will be zero. Therefore, we set a minimal MV intentionally, which is a just Noticeable Difference (JND) motion scale value [29], such as \( \langle M_X(i), M_Y(i) \rangle = (1, 1) \). From the statistical perspective, the rationality can be interpreted as follows:

- If the intra-MB is one by one, as shown in Figure 4, the error cost of estimation is very high, when estimating the MV for all the intra-MB. That is to say, the MB may contain motion information, but we regard it as no-motion to human perception.
- The setting MV is minimal. If it is a still object, the error cost for them is not very big compared to the whole content.

![](FIGURE_3.png)

FIGURE 3 The Continuous Intra-encoded MB in P-Frame

![](FIGURE_4.png)

FIGURE 4 The Single Intra-encoded MB in P-Frame

3.1.2 16×16 I-MB

There are four prediction modes. And for the first three prediction modes, the padding method is the same as the padding method for the 4×4 sub-block. Although Mode 3 employs lane function for the prediction, we also employ the mean value of MV of coded MB B_0 and B_3 and for padding all sixteen 4×4 sub-blocks in 16×16 I-MB. Note that, for 16×16 I-MB, sixteen 4×4 sub-blocks own the same padded MVs. Therefore, our novel MV padding approach is universal for both 4×4 sub-block and 16×16 MB.
3.2 Post-processing based on prediction partition

In the H.264/AVC standard, the smaller the prediction partition is, the more details or drastic change of the corresponding region contains. To employ this useful information, which the partition mode reflected for video content, the padded MV of 4×4 sub-block is post-processed by the partition mode. According to (10), when the prediction mode is small, the weight of MV for 4×4 sub-block will be big, which attracts more attention. Instead, when the partition mode is big, the MV will be scaled small, and the user pays less attention to it.

\[
(M'_{X(i)}, M'_{Y(i)}) = \left( \frac{M_{X(i)}}{W}, \frac{M_{Y(i)}}{H} \right)
\]

Where \(W\) and \(H\) are the width and height of the partition block before padding. \((M_{X(i)}, M_{Y(i)})\) is the vertical and horizontal value of padded MV. \((M'_{X(i)}, M'_{Y(i)})\) is the post-processed vertical and horizontal MV for the current 4×4 block.

4 PERCEPTUAL FEATURES FOR REPRESENTING COMPRESSION IMPAIRMENT

Feature selection is the critical issue for establishing a learning-based quality assessment model. We utilize the transmission impairment features, which have been defined by DASH industry forum [30]. Thus, the emphasis of our work just describes perceptual features caused by compression. The degradation caused by compression is related to video content characteristics [28, 31, 32], so the perceptual features should represent the video content characteristic. Either low-level features or high-level features can be employed to quantify the video content characteristic [33]. Low-level features, which include the texture, form, colour, and sharpness, are the basic element of video content characteristic. While the high-level features can be directly distinguished and recognized by human, such as objects, scene, and person. Note that high-level features are usually built upon low-level features [33]. Due to the size limitation of the used subjective database, we just employ low-level perceptual features to interpret the video content characteristics caused by compression, such as contrast, blur, texture, motion vector scale, and motion vector direction. In this section, we show the importance of video content characteristics to human visual perception. After that, we present selected perceptual features to represent.

4.1 Video content characteristic on human visual perception

The video with dynamic content or static content has a significant discrepancy on users’ QoE. Meanwhile, due to the time-varying characteristic of the video sequence, the perception of one video sequence varies as the time elapsed. So for the constant bitrate encoding, the bitrate cannot demonstrate the video content difference. The snapshots for three different video sequences and the corresponding subjective MOSs are shown in Figures 5 and 6, respectively, which come from Waterloo SQoE-III database [7]. Three of them own the same impairments, such as the same encoding bitrate, rate switches, and stalling events, as shown in Table 1. Therefore, the influence
of different video content characteristics should be taken into account during measuring users’ QoE for DASH video streaming.

4.2 Representation of perceptual features

We briefly describe the extraction of content perceptual features for the video sequences. In the Waterloo SQoE-III database [7], there are eleven quality levels with different encoding bitrates and resolutions for one original video sequence. For every quality level video sequence, preliminary features are extracted in the unit of a single frame or two consecutive frames. After that, segment-level features are obtained by averaging the preliminary features in terms of 2 s, which is consistent with the segment encoding length [7]. Finally, patterns of rate switches are added according to every test video sequence, which are available in the Waterloo SQoE-III database [7]. In this way, we obtain the set of content perceptual features for every test video sequence.

4.2.1 Contrast

Contrast refers to the difference between the brightest and the darkest point in one frame, which is determined by the video acquisition and de-blocking filter during decoding in H.264/AVC. The former is the intrinsic property for the video sequence, and the latter is to improve the blocking effect caused by compression. Although the user is insensitive to the distortion in the brightest or darkest regions, an appropriate contrast plays a very important role in determining the user’s QoE. The contrast of one frame is usually defined as (11).

\[ C(n) = STD_{space}(F_n(i,j)) \] (11)

where \( F_n(i,j) \) is the Y component value at the coordinate of \((i,j)\) in the frame. \( STD_{space} \) stands for the standard deviation of all Y component values in the \( n_{th} \) frame.

4.2.2 Blur

The video acquisition or compression is the cause of the blur phenomenon, which can be easily perceived by a human. The acquisition will lead to the intrinsic blur phenomenon, while the compression leads to varying blur phenomenon depending on encoding bitrate, which occurs inside the block and on block boundary. For the inside blurring, although Discrete Cosine Transform (DCT) is lossless, the quantization process about DCT coefficients is relatively rough. So the inverse quantization cannot recover the original signal perfectly. In other words, DCT aggregates energy into different frequencies without information loss, while the no-continuous quantization discards partial magnitude of DCT coefficients. Due to the majority of video content concentrates on the low frequency, the quantization severely reduces the high-frequency information. While the high-frequency information usually lies in the edge region or contains objective details, the reverse quantization results in significant distortion for the reconstructed signals. Therefore, the quantization leads to no-continuous visual perception. Another reason is the in-loop de-blocking filtering in encoding, whose essence is a spatial low-pass filter. It is employed to reduce artifacts across block boundaries, so it also brings the blur.

On visual effects, the blur appears as a loss of sharpness about the video content, such as the high-frequency information and objective details. We measure the sharpness degree, which is the inverse value of blur and normalized by block numbers [34]. Because the metric is based on the probability summation model rather than the whole frame, we need to select the block size according to the viewing distance, whose importance is also explained in [35], and display resolution. The chosen block size is consistent with the foveal region of HVS, which is approximate \( 2^6 \) of the visual angle. The number of pixels contained in the region \( N \) can be computed as (12) [34]. Note that the spatial area, which is centered at the location \((n_1, n_2)\) and covers \( 2^6 \) of the visual angle, is denoted as \( F(n_1, n_2) \).

\[ N(F(n_1, n_2)) = \left( 2 \left( dr \tan \left( \frac{ff}{2} \right) \right)^2 \right) \approx \left( 2 [r] \right)^2 \] (12)

where \( r \) donates the visual resolution of the display (pixel/degree), the calculation is shown in (13).

\[ r = dr \tan \left( \frac{B}{180} \right) \approx \frac{drB}{180} \approx \frac{d \times \frac{v}{57.3}}{} \] (13)

where \( d \) is the display resolution in pixel/cm, and \( v \) is the viewing distance in cm.

The settings of subjective experiments in Waterloo SQoE-III database [7] are stated as follows. The viewing distance is 31.5 inches, the display resolution is 1920×1080, and the screen size is 23 inches. According to this information, the computed block size with (12) will be 102×102. To be consistent with the block-based technology of the modern video encoding standard, we still select the block size is 64×64.

4.2.3 Texture

Since the texture is an important perceptual feature for image content, the video sequence consists of many continuous
images in the timeline. Therefore, we still utilize the image metric to measure the texture of the video sequence. A rich texture region can hide more artifacts than a smooth region or edge region. This is determined by the masking characteristics of HVS. We employ the Log-Gabor filter to extract the amount of texture information in one video frame [36], shown in (16). The transform function of the Log-Gabor filter includes two components, which are shown in (14) and (15).

\[
F_m(\omega) = \exp \left( -\frac{\ln((f')_{x,m})^2}{2\ln(\omega_{x,m})^2} \right) \quad (14)
\]

\[
F_k(\theta) = \exp \left( -\frac{(\theta - (\omega_k))^2}{2\pi k^2} \right) \quad (15)
\]

where \( m = 1, 2, ..., M \) represents the transform scale, \( k = 1, 2, ..., K \) represents the transform direction, and \( \omega_{x,m} = \frac{2}{mh}, \sigma_\theta = \frac{\pi}{K}, \phi_k = \frac{(k-1)\pi}{K} \). \( G_{mk}(\omega, \theta) \) demonstrates the output of the original picture with different scales and different directions.

4.2.4 | Motion vector scale

Motion vector can reflect the actual motion trajectory among successive frames for video sequences. Non-zero MV is a good indicator of how fast the overall movement occurs. To fully utilize the impact of temporal video content on human perception, we pad the MV for I-MB in P and B frames, which is stated in Section 3. Based on the completed MV field for one frame, we employ the mean of magnitude for the non-zero MVs, which have been normalized by the pattern of block partition, to describe the video motion scale shown in (17).

\[
MV^*_S(n) = \frac{1}{N_{MB}} \sum_{i=1}^{N_{MB}} \left( \sqrt{(M'_{X(i)})^2 + (M'_{Y(i)})^2} \right) \quad (17)
\]

where \( N_{MB} \) is the number of the non-zero motion vector in the unit of 4x4 sub-block in one frame. \( (M'_{X(i)}, M'_{Y(i)}) \) is the post-processed vertical and horizontal component of MV for the 4x4 sub-block.

4.2.5 | Motion vector direction

Motion direction is also a key aspect to determine human visual perception. The human may pay attention to the large coherent motion and easily become aware of the change in consistent motion. Therefore, we define this metric as the standard deviation of MV direction in one frame, which is shown in (18).

\[
MV'_{D}(n) = \text{std}(\theta_{M'V}) \quad (18)
\]

where \( \theta_{M'V} = \arctan(M'_{Y(i)}/M'_{X(i)}), 0 \leq \theta_{M'V} \leq 2\pi \). \( (M'_{X(i)}, M'_{Y(i)}) \) is the vertical and horizontal size of motion vector of 4x4 sub-block.

5 | MODEL EVALUATION METHOD

The LIBSVM package was used to implement the \( \varepsilon - SV\mathcal{R} \) with Radial Basis Function (RBF) kernel [37]. When learning the regression model, parameter sets \((C, \gamma) \) need to be determined. \( C > 0 \) is a constant parameter to balance margin maximization and training error minimization. \( \gamma \) controls the kernel shape, which maps features onto a high-dimensional space. More details can be found on page sixteen [27]. We need a subjective assessment database to verify the performance of our proposed model.

To date, Waterloo SQoE-III database [7] is the latest public subjective database available for DASH video steaming, which simulated with six adaptive bitrate algorithms on 13 extensive and representative the network statuses through a test-bed. The distortion types are various according to the change of network conditions. Twenty original video sequences with diverse spatio-temporal content are carefully chosen. The brief information of the Waterloo SQoE-III database is shown in Table 2.

To avoid content dependencies for obtaining the effective performance, we randomly split the source video sequences into the training set and testing set randomly: 80% of sequences (16 source video sequences) were used for training, and 20% of sequences (4 source video sequences) were used for testing. There is no video content overlap in the training/testing set. For the convenience of performance comparison, we pre-generated the 1000 training/testing trials. The Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Order Correlation Coefficient (KROCC) were computed between objective predicted QoE scores and subjective MOSs for every testing set. The PLCC value presents the prediction accuracy, and the last two values measure the prediction monotonicity. Both values lie in \([-1, 1]\). The closer to 1 (or -1) of the values for PLCC and SROCC are, the better the performance can be. The value of KROCC is close to 1, and it means the agreement between the two rankings is perfect. Conversely, the disagreement between the two rankings is, if the value of KROCC is close to -1. Before computing the PLCC value, a five-parameter non-linear logistic regression will be utilized on the objective prediction scores [38]. The reason is that there exists non-linearity between subjective MOSs and objective predictive QoE scores due to the non-linear quality grade of human. Another reason is that it can transform the two type of values in the same range. The definition of five-parameter non-linear logistic regression is shown in (19).

\[
\text{Quality} = \beta_1 \left( 1 - \frac{1}{\exp(\beta_2(Q_{obj}-\beta_3))} \right) + \beta_4 Q_{obj} + \beta_5 \quad (19)
\]

where \( Q_{obj} \) is the objective predictive score obtaining from the objective QoE assessment model and presents the correspond-
ing value after logistic regression. $\beta_1, \beta_2, \beta_3, \beta_4$ and $\beta_5$ are the parameters of the logistic regression function.

To be convincing and stable, the median values of these three metrics are chosen as the final performance among 1000 random combinations of the train-test trail in our experiment. To obtain the optimal parameter settings for every training set and minimize the prediction error for the testing set, we perform grid-search and five-fold cross-validation. The optimal value of $(C, \gamma)$ is $(3.0314, 0.25)$ is for our model. Once the mapping function is learned, the user’s QoE can be obtained easily for DASH video streaming.

6 | EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we give the experimental results and analysis about our model from three aspects, model performance comparison, statistical significance, and the effectiveness of temporal content characteristic.

6.1 | Model performance comparison

To obtain a relative fair comparison with the state-of-the-art QoE assessment models, the values of PLCC, SROCC and KROCC are computed on our pre-generated 1000 testing trials for the comparative models. We also re-train the Video ATLAS on our pre-generated training set. We list the model description, including the used features, application scenario mapping function and so on, shown in Table 3. The median values of evaluation metrics across 1000 train-test trails are given in Table 4. Three major findings can be concluded from Table 4.

- The importance of compression impairment on human visual perception: For one thing, the performance of SSIMplus [15], and VQM [16], which just measure video quality caused by compression impairment, are all better than these models accounting for the transmission impairment, such as FTW [8], VsQM [9], and Mok’s [10]. It indicates that the compressed video characteristic is more important than transmission impairment on human visual perception.

For another thing, when representing the compression impairment, QP [22], bitrate [19, 20], and existing VQA metric [3, 23] are employed. The performance of these models is all lower than our proposed model. Specifically, the performance of QoE Continuum [22] model is worst. This is because QP is a rough video quality indicator, which is the direct cause of compression impairment, the performance [19, 20] is lower than our proposed model. This can be explained that the distinguishable ability for video content characteristic in the horizontal dimension, whose definition is presented in Section 6.3, is not fully utilized.

- The importance of how to joint compression impairment and transmission impairment: The performance of our proposed model is better than all the comparative models, especially SQI [3], Liu’s [17], and Video ATLAS [23]. The reason can be interpreted as follows. Firstly, as shown in Figure 7, the mapping form is different. These three models preferentially quantify the user’s QoE resulted from compression impairment. And then, when stalling events occur, mapping the user’s QoE and stalling QoE influential factors to obtain the final user perception. However, due to the complexity of HVS, the user cannot accurately distinguish the specific impairment type and respectively give the perceptual degree of artifacts when the video sequence is impaired by compression and transmission.

- The importance of motion information to human visual perception: For one thing, due to the use of QP as a quality indicator, which lacks motion information and doesn’t have the direct perceptual relation to human visual perception, the performance of QoE Continuum is poor [22]. For another thing, because the simple motion information representing method employed, which is based on the difference of the co-located pixel values to compute temporal content in two consecutive frames, Video ATLAS [23] performs better.

6.2 | Statistical significance and hypothesis testing

For all models, the spread of distributions in terms of SROCC and PLCC are shown in Figures 8 and 9. Different colours represent different QoE influential factors, which means that the focus on different impairments. The green indicates compression-oriented assessment models. The purple shows transmission-oriented models, and the red is assessment models accounting for impairments of both compression and transmission. These SROCC and PLCC values are obtained from

| Source videos | Resolution | Source video length | Test videos | Presentation quality | Stall length | Stall location |
|---------------|------------|---------------------|-------------|----------------------|--------------|---------------|
| 20            | 7 level    | 10(s)               | 450         | 11 level             | 0.1667–12.0667(s) | Simulated     |
TABLE 3  The comparison of existing assessment models and our proposed model

| QoE model | Related features | Application scenario | Mapping function | Used dataset |
|-----------|------------------|----------------------|-----------------|-------------|
|           | Compression      | Transmission         |                 |             |
| SSIMplus [15] | Mean, variance, covariance | — | √ | Statistical | Non-public |
| VQM [16] | si_loss, hv_loss, hv_gain, chroma_spread, i_gain, ct_ati_gain, chroma_extreme | — | √ | Polynomial | Non-public |
| FTW [8] | Stalling length, stalling number | — | √ | Exponential | Non-public |
| VQM [9] | Average stalling length per segment, number of stalling per segment, segment length | — | √ | Exponential | Non-public |
| Mok’s [10] | Initial buffering time, stalling length, stalling frequency | — | √ | Linear | Non-public |
| SQI+PSNR [3] | PSNR | Stalling number, stalling length | √ | √ | Exponential, Public |
| SQI+SSIMplus [3] | SSIMplus | | | polynomial (hand-crafted) |
| Liu’s [17] | Bitrate | Stalling length | √ | √ | Linear | Non-public |
| Yin’s [19] | Average bitrate, average quality variation, initial stalling time | Stalling length | √ | √ | Linear | Non-public |
| Bentaleb’s [20] | SSIMplus | Stalling number, stalling length | √ | √ | Linear | Non-public |
| Liu’s [21] | VQM [16] | Stalling length, stalling number, magnitude of MV | √ | √ | Polynomial | Non-public |
| ATLAS+PSNR [23] | PSNR | Rebuffering-aware features, | √ | √ | SVR | Public (hand-crafted) |
| ATLAS+SSIM [23] | SSIM | Memory-related features, | Impairment duration feature | | |
| ATLAS+STRRED [23] | STRRED | | | | |
| QoE Continuum [22] | QP | Stalling number, stalling length, frame size | | | |
| Proposed | Contrast, blur, texture, motion vector magnitude, motion vector direction | Initial buffering time, stalling percentage, stalling count, bitrate switch count, average bitrate switch magnitude | √ | √ | SVR | Public (simulated) |

1000 train-test trials for all considered comparative QoE models on the pre-generated 1000 test-train trials. The differences in the median value are obvious, and the ranking of performance can be seen easily. Meanwhile, the concentricity of SROCC and PLCC values presented by our proposed model. It means that our proposed model has strong robustness to different video content combinations on Waterloo SQoE-III database [7].

Although the differences between our proposed model and existing comparative models are obvious, which shown in Figures 8 and 9, the statistical significance cannot be guaranteed. So To further evaluate the statistical significance of our proposed model, we conduct the hypothesis test based on the t-test with SROCC values, which are obtained from 1000 train-test trials. Note that the independence and normality of data must be guaranteed before the t-test. Because every train-test trial is chosen independently, so the independency is easily guaranteed. Meanwhile, we also have verified that the kurtosis values of SROCC values are between 2 and 4 for all considered assessment models, which means that they all obey normal distribution [38]. For the t-test, the null hypothesis is that SROCC values of the two models are coming from one population, which own equal mean with 95% confidence. The test result
TABLE 4  Performance comparison between our proposed model and existing comparative models

| QoE model     | PLCC   | SROCC  | KROCC  | Model type |
|--------------|--------|--------|--------|------------|
| SSIMplus [15]| 0.6690 | 0.5884 | 0.4245 | P          |
| VQM [16]     | 0.7202 | 0.6056 | 0.4407 | P          |
| FTW [8]      | 0.3215 | 0.2771 | 0.2125 | P          |
| VsQM [9]     | 0.4221 | 0.4119 | 0.2868 | P          |
| Mok’s [10]   | 0.2593 | 0.2282 | 0.1301 | P          |
| SQI+PSNR [3] | 0.6974 | 0.7016 | 0.5154 | L          |
| SQI+SSIMplus [3] | 0.7949 | 0.7922 | 0.6028 | P          |
| Liu’s [17]   | 0.8053 | 0.7927 | 0.6044 | P          |
| Yin’s [19]   | 0.6447 | 0.6511 | 0.4734 | P          |
| Bentaleb’s [20] | 0.5820 | 0.5793 | 0.4141 | P          |
| Liu’s [21]   | 0.4723 | 0.5233 | 0.3762 | P          |
| ATLAS+PSNR [23] | 0.7516 | 0.6815 | 0.5089 | L          |
| ATLAS+SSIM [23] | 0.7497 | 0.7179 | 0.5358 | L          |
| ATLAS+STRRED [23] | 0.6253 | 0.6132 | 0.4385 | L          |
| QoE Continuum [22] | 0.4433 | 0.3997 | 0.2712 | P          |
| Proposed     | 0.8183 | 0.7982 | 0.6471 | L          |

*The fifth column: P, parametric assessment model; L, learning-based assessment model.

FIGURE 7  The difference of the mapping function for existing assessment models and our proposed model. (a) Existing assessment models, (b) our proposed model

FIGURE 8  The Box plot of SROCC values distribution over 1000 train-test trials about all considered QoE models

FIGURE 9  The Box plot of PLCC values distribution over 1000 train-test trials about all considered QoE models

FIGURE 10  Results of statistical significance based on t-test conducted on Waterloo SQoE-III database by using the SROCC values of all considered QoE models

is illustrated in Figure 10, where a symbol ‘1’ means that the row model is statistically superior to the column model, while a symbol ‘−1’ owns the opposite meaning. A symbol ‘0’ represents two models that come from the same population, which means that the performance for the row model and the column model is statistically indistinguishable. From the statistical result, which is shown in Figure 10, we can have three conclusions.

• Our proposed model statistically outperforms all comparative models.
• Transmission-oriented models are also relatively poor from the statistical perspective, which caused by the ignorance
of the importance of video content characteristic to human visual perception.

- The performance of SQI [3], Liu’s model [17], and Video ATLAS [23] statistically precede other models. Once again, it verifies that more attention should be paid to the impact of compression and transmission on the user’s QoE concurrently.

6.3 Distinguishable ability of perceptual features

Due to the flexibility and scalability, the difference of video content characteristics should be distinguishable on human visual perception for DASH video streaming. Due to the space limitation, we just demonstrate the distinguishable ability of partial content perceptual features, for which the impairment is caused by compression. In Figure 11, the effective distinguishable ability is obvious among different bitrate for the same video sequence, which refers to the vertical dimension distinguishing. In Figure 12, the effective distinguishable ability is also obvious for three video sequences with the same bitrate, which refers to the horizontal dimension distinguishing. Meanwhile, different segments of one video sequence also render discrepancy, shown in Figures 11 and 12. Although some feature owns weak discrimination in the vertical dimension in Figure 11, it just happens at high resolution and high bitrate. This can be interpreted as follows. With the increasing of video encoding bitrate, the user’s perception increases definitely. However, the increment trend does not present a positive linear relation, and the rate of increment slows actually. Moreover, the trend becomes steady after a certain bitrate according to different video content [28, 29, 32]. The reason can be interpreted that the relative loss of video content details could be small in encoding with high encode bitrate in the high resolution. Therefore, the video content is perfectly rendered to users. Recent researches have also concluded that the changing trend is unrelated to video encoding technologies, such as HEVC, H.264/AVC, or AV1 [31, 32].

6.4 Effectiveness of temporal video content characteristic

To verify the effectiveness of our representing approach for motion information representing approach, we respectively utilize three methods to quantify motion information in our proposed model respectively, such as TI, no padding (using MV directly) and our padding approach. The performance is
FIGURE 12  The distinguishing ability of perceptual features for three video sequence with the same bitrate in Waterloo SQE-III database. (a) Contrast, (b) Texture, (c) Motion Vector Scale, (d) Motion Vector Direction

TABLE 5  Performance comparisons between different motion representing methods

| Motion representing methods | PLCC   | SROCC  | KROCC  |
|-----------------------------|--------|--------|--------|
| TI                          | 0.7135 | 0.6499 | 0.4808 |
| No padding                  | 0.7068 | 0.7707 | 0.5473 |
| Padding (Proposed)          | 0.8183 | 0.7982 | 0.6471 |

also trained and tested on the Waterloo SQoE-III Database [7]. The median values of PLCC, SROCC, and KROCC for 1000 train-test trials are shown in Table 5. It can be seen that our padding method owns higher performance benefiting from reflecting video content characteristics consistent with human visual perception. Although the improvement is not obvious compared to utilizing MV directly, it can be interpreted that I-MBs is not the main encoding type in P-frames and B-frames, which depends on the video content characteristic.

7  DISCUSSION AND FUTURE WORK

In this section, we revisit three important issues.

Large-scale Database: Although the performance of our proposed model is better than the comparative models, we just verify the performance on the Waterloo SQoE-III [7] with limited video sequences and limited distortion types. To comprehensively consider the diversity of video content in the real world, a future direction is to establish a large-scale database using a crowdsourcing platform. The user-specific information, such as age, preference, gender, and view time, can be available. Thus, the video service provider can stream the customized video content to satisfy users’ QoE by employing the assessment model quantifying this information.

Model Improvement: Due to the size limitation of the available database, features extracted in our work are handcrafted. However, the human visual system is complex, and we cannot adequately consider the influential factors manually. Thus, once the large-scale database is available, a future direction is to use the end-to-end deep-learning network to extract features and obtain the predicted users’ QoE directly. We have the chance to further improve users’ QoE based on a more accurate assessment model.

Model Application: Our work aims to establish a QoE assessment model. However, we need to make the model valuable by applying the model into the rate adaptation algorithm to improve users’ QoE. As a future direction, integrating the
proposed assessment model into the rate adaptation algorithms has two different scenarios, for example post-evaluation at client and prediction at server or base station. We provide the statement in Appendix.

8 CONCLUSION

We proposed an objective QoE assessment model by implementing a learning-based approach for DASH video streaming. The model directly accounts for the impairments introduced by compression and transmission. Experimental results show that our proposed model outperforms the state-of-the-art QoE assessment models in both accuracy and statistical significance. Meanwhile, our proposed model owns a strong generalization ability to different video contents. The features in our model can effectively demonstrate the characteristics of video content induced by compression impairment in different dimensions. The features can be easily extracted and embedded in the media presentation description (MPD) file. Thus, it can be employed by rate adaptation strategies of DASH video streaming. In addition, a novel MV padding method is proposed to reflect the impact of video temporal characteristics on human visual perception, which can be easily obtained based on the block prediction direction modes and partition modes in encoding.

ACKNOWLEDGEMENTS

This research work was supported in part by the National Science Foundation of China (61671365, U1903213), and in part by the Key Research and Development Program of Shaanxi Province (2020KW-009).

REFERENCES

1. 3GPP TS 26.247 V16.1.0, Progressive download and dynamic adaptive streaming over HTTP. https://www.3gpp.org/ftp/Specs/archive/26_series/26.247/26247-g10.zip (2019). Accessed 10 May 2019
2. ITU-Recommendation BT.500-13, Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland. https://www.itu.int/rec/R-REC-BT.500-13-201201-S/en (2012). Accessed 28 Apr., 2019
3. Duanmu, Z., et al.: A quality-of-experience index for streaming video. IEEE J. Sel. Top. Signal Process. 11(1), 154–166 (2017)
4. Esvara, N., et al.: A continuous QoE evaluation framework for video streaming over HTTP. IEEE Trans. Circuits Syst. Video Technol. 28(11), 3236–3250 (2018)
5. Ghadiyaram, D., et al.: A subjective and objective study of stalling events in mobile streaming videos. IEEE Trans. Circuits Syst. Video Technol. 29(1), 183–197 (2019)
6. Bampis, C.G., et al.: Study of temporal effects on subjective video quality of experience. IEEE Trans. Image Process. 26(11), 5217–5231 (2017)
7. Duanmu, Z., et al.: A quality-of-experience database for adaptive video streaming. IEEE Trans. Broadcast. 64(2), 474–487 (2018)
8. Hoffeld, T., et al.: Internet video delivery in YouTube: from traffic measurements to quality of experience. In: Data Traffic Monitoring and Analysis, pp. 264–301. Springer, Heidelberg (2013)
9. Rodriguez, D.Z., et al.: Quality metric to assess video streaming service over TCP considering temporal location of pauses. IEEE Trans. Consum. Electron. 58(3), 985–992 (2012)
10. Mok, R.K.P., et al.: Measuring the quality of experience of HTTP video streaming. In: Proceedings of IFIP/IEEE International Symposium on Integrated Network Management, Dublin, 23–27 May 2011.
11. Ghadiyaram, D., et al.: Learning a continuous-time streaming video QoE model. IEEE Trans. Image Process. 27(5), 2257–2271 (2018)
12. Chen, C., et al.: Modeling the time-varying subjective quality of HTTP video streams with rate adaptations. IEEE Trans. Image Process. 23(5), 2206–2221 (2014)
13. Soundararajan, R., Bovik, A.C.: Video quality assessment by reduced reference spatial-temporal entropic differencing. IEEE Trans. Circuits Syst. Video Technol. 23(4), 684–694 (2013)
14. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)
15. Rehman, A., et al.: Display device-adapted video quality-of-experience assessment. In: Proceedings of IS&T/SPIE Annual Symposium on Electronic Imaging, San Francisco, 9–12 Feb 2015
16. Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Trans. Broadcast. 50(3), 312–322 (2004)
17. Liu, Y., et al.: Deriving and validating user experience model for DASH video streaming. IEEE Trans. Broadcast. 61(4), 651–665 (2015)
18. Yang, J., et al.: Cross-layer design of resource allocation and segment adaptation based on CPIQ model for DASH clients over LTE networks. IET Commun. 13(10), 1405–1414 (2019)
19. Yin, X., et al.: A control-theoretic approach for dynamic adaptive video streaming over HTTP. ACM SIGCOMM Comp. Commun. Rev. 45(4), 325–338 (2015)
20. Bentaleb, A., et al.: SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking. In: MM ’16: Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, 15–19 Oct 2016
21. Liu, X., et al.: A case for a coordinated internet video control plane. ACM SIGCOMM Comp. Commun. Rev. 42(4), 359–370 (2012)
22. Xue, J., et al.: Assessing quality of experience for adaptive HTTP video streaming. In: Proceedings of 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, 14–18 July 2014
23. Bampis, C.G., Bovik, A.C.: Learning to predict streaming video QoE: distortions, rebuffing and memory. https://arxiv.org/pdf/1703.00633.pdf (2017). Accessed 1 May, 2019
24. Bampis, C.G., et al.: Recurrent and dynamic networks that predict streaming video quality of experience. IEEE Trans. Image Process. 27(7), 3316–3331 (2018)
25. Kumar, V., Mahapatra, S.: Quality of experience driven rate adaptation for adaptive HTTP streaming. IEEE Trans. Broadcast. 64(2), 602–620 (2018)
26. Robitza, W., et al.: A modular HTTP adaptive streaming QoE model—Candidate for ITU-T P1203 (“PNATS”). In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, 31 May–2 Jun 2017
27. Scholkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
28. Su, H., et al.: Content-based bitrate model for perceived compression distortion evaluation of mobile video services. IET Image Process. 11(11), 1027–1033 (2017)
29. Yang, K., et al.: Perceptual temporal quality metric for compressed video. IEEE Trans. Multimedia 9(7), 1528–1535 (2007)
30. DASH industry forum: DASH-IF position paper: Proposed QoE media metrics standardization for segmented media playback. https://dashif.org/docs/ProposedMediaMetricsforSegmentedMediaDelivery-r12.pdf (2016). Accessed 11 November 2018
31. Tan, T.K., et al.: Video quality evaluation methodology and verification testing of HEVC compression performance. IEEE Trans. Circuits Syst. Video Technol. 26(1), 76–90 (2016)
32. Huang, T., et al.: QARC: Video quality aware rate control for real-time video streaming based on deep reinforcement learning. In: Proceedings of 2018 ACM Multimedia Conference (MM ’18), Seoul, 22–26 Oct 2018
33. Gao, G., et al.: Optimizing quality of experience for adaptive bitrate streaming via viewer interest inference. IEEE Trans. Multimedia 20(12), 3399–3413 (2018)
In this section, we state two application scenarios of the proposed model in rate adaptation algorithm, for example post-evaluation at client and prediction at server or base station.

### A.1  
**Post-evaluation at client**

For the user side, the features needed in our proposed model can be conveniently obtained. And our model is utilized to assess users’ QoE after receiving five continuous segments for quality monitoring. It gives the feedback of the current users’ QoE to the server or the base station for the strategies of rate adaptation to guarantee the user’s QoE.

### A.2  
**Prediction at server or base station**

For the prediction at the server or base station, our model can be served as the utility function. In this situation, two types of feedback information are needed, which are users’ network status and buffer status. The information can be transmitted with specified time granularity, which is decided by the period of rate adaptation algorithm. Moreover, the features of the fifth segment are needed to estimate. The actual estimation process is provided as follows. The employed parameters for the feature estimating process of the fifth segment, are shown in Table A.1.

(1) Estimating the features about transmission impairment for downloading segment for presentation $p$

- **Downloading duration of segment** for presentation $p$,
  \[ \tau^p_j = \frac{\gamma^p_j}{C_j} \]  (A.1)

(2) Computing features related to transmission impairment for segment $l$

The letter with a symbol $\sim$ represents the real record for every feature, for which the transmission impairment

![Equation](https://doi.org/10.1049/jpr2.12012)
has already occurred. Therefore, the transmission-related features employed in our proposed model are computed by adding the estimated features, which uses the approach described in the first part of the Appendix. In the following, we present the computed features of segment \( l \) with different representations \( P \). These features are \( P \)-dimensional vectors except initial stalling delay. One component is employed when computing the features of segment \( l \) for presentation \( p \).

- **Initial stalling delay (\( I_l \))**

  When the number of viewing segments is more than five, there is no initial stalling delay available. However, the duration of stalling event corresponding to the initial stalling event is given a discount factor \( \frac{1}{3} \) [7]. Therefore, when applying our proposed model, the initial stalling delay can be regarded as a special stalling event and give the weight 3.

- **Stalling percentage (\( P_l \))**

  1
  \[
  \frac{1}{L_{\text{session}}} \left[ \bar{I}^{l-1}_{r} \times L^{l-1}_{\text{session}} + r^{l}_{r} \bar{I}^{l-1}_{r} \times L^{l-1}_{\text{session}} \right. \\
  + r^{2}_{r}, \ldots, r^{l}_{r} \times L^{l-1}_{\text{session}} + r^{l}_{r} \right]
  \] (A.6)

- **Stalling count (\( C_{lr} \))**

  \[
  [\bar{C}^{l-1}_{r} + a^{1}_{lr}, \bar{C}^{l-1}_{r} + a^{2}_{lr}, \ldots, \bar{C}^{l-1}_{r} + a^{l}_{lr}] = \bar{C}^{l-1}_{r} \\
  + [a^{1}_{lr}, a^{2}_{lr}, \ldots, a^{l}_{lr}] 
  \] (A.7)

- **Bitrate switch count (\( C^{l}_{b} \))**

  \[
  [\bar{C}^{l-1}_{b} + \bar{C}^{l-1}_{b}, \bar{C}^{l-1}_{b}, \ldots, \bar{C}^{l-1}_{b} + \bar{C}^{l-1}_{b}] = \bar{C}^{l-1}_{b} \\
  + [\bar{C}^{1}_{b}, \bar{C}^{2}_{b}, \ldots, \bar{C}^{l}_{b}] 
  \] (A.8)

- **Bitrate switch magnitude (\( M^{l}_{b} \))**

  \[
  \frac{1}{S^{l}_{b}} \times \left[ \bar{M}^{l-1}_{b} \times M^{l-1}_{b} + s^{1}_{l_{b}} \bar{M}^{l-1}_{b} \right. \\
  + s^{2}_{l_{b}}, \ldots, \bar{M}^{l-1}_{b} + s^{l}_{l_{b}} \right] 
  \] (A.9)

(3) **User’s QoE**

The features related to the compression impairment are available in MPD file. It corresponds to the encoding bitrate one-by-one, which means, for one specific bitrate \( B_{r}^{p} \), the corresponding features, \( C^{p}_{r}, B^{p}_{r}, T^{p}_{r}, MVS^{p}_{r}, MVD^{p}_{r} \), are all available. The predicted user’s QoE of downloading segment \( l \) for presentation \( p \) is shown in (A.10), where \( f \) is our proposed map function.

\[
QoE^{l}_{p} = f \left( C^{p}_{r}, B^{p}_{r}, T^{p}_{r}, MVS^{p}_{r}, MVD^{p}_{r}, \\
T_{l}, \frac{1}{L_{\text{session}}} \times P^{l-1}_{r} \times L^{l-1}_{\text{session}} + r^{p}_{r} \bar{C}^{l-1}_{r} + a^{p}_{r}, \\
\bar{C}^{l-1}_{b} + \bar{C}^{l-1}_{b}, \frac{1}{S^{l}_{b}} \times \left( \bar{M}^{l-1}_{b} \times M^{l-1}_{b} + s^{l}_{l_{b}} \right) \right) 
\] (A.10)