QoE Models for Online Video Streaming

Rongqi Zhang[*], Xu Zhang[†], Peiyao Guo[*], Qilin Fan[‡§], Hao Yin[¶], Zhan Ma[*]

[*]*School of Electronic Science and Engineering, Nanjing University, China,*
[†]*College of Engineering, Mathematics & Physical Sciences, University of Exeter, UK,*
[‡]*School of Big Data and Software Engineering, Chongqing University, China,*
[§]*Chongqing Key Laboratory of Digital Cinema Art Theory and Technology, Chongqing University, China,*
[¶]*Beijing National Research Center for Information Science and Technology, Tsinghua University, China.*

Abstract

With the rising popularity of video streaming services, Quality of Experience (QoE) acts as the crucial role in improving user's experience. Distinguished from Quality of Service (QoS), which mainly weigh the video streaming quality with network transmission performance, QoE focuses on the subjective and user-oriented assessment when users views the videos online. It is a vital issue for online streaming to ensure a user-satisfied transmission under dynamic network. This highlights the need for an accurate and feasible QoE model to balance the user experience and available transmission bandwidth. This paper summaries and analyzes existing explorations on QoE models for online video streaming, especially with the advancement on emerged learning-based models, to promote the development of this research field.

I. Introduction

It is reported that people spend 6.8 hours per week watching various types of online videos in average, which performs a 59% increase on time since 2016 [1]. Besides, the popularity of subscription video on-demand (SVOD) services rapidly rise by 17% over 2018 as 70% online viewers subscribe to one or more SVOD services. The global video streaming market was worth 42.6 billion dollars in 2019 and expected to register a Compound Annual Growth Rate (CAGR) of 20.4% till 2027 [2]. When facing vast online consumer demand, it is essential to supervise and maintain the streaming quality for users' viewing satisfaction. Hence, how to quantify the performance of various online streaming services for viewers becomes a key problem to upgrade online services [3]–[5]. Existing researches on this topic have gone through four stages as below, from handcrafted rating to mathematical modeling.

A. Quality of Service (QoS) Monitoring

QoS is first defined in 1994 [6] by International Telecommunication Union (ITU), which measures the quality of video transmission considering various factors, e.g., service stability, service availability, service reliability, and service scalability. It describes the service quality via objective attributes while neglecting the experience perceived by the viewers. As a sequence, users can experience diversified extent of satisfaction even with the same level of QoS.

B. Subjective Test

For online streaming service, the viewing experience from the receiver client is crucial to assess service quality. And the most related factor of viewing experience is various to different video contents. For example, people expect a smaller initial delay when viewing short films ($\leq$1 min) or shifting from one to another; while watching long movies or tv shows ($\geq$ 0.5 hour), high-definition content and broadcast without rebuffering could provide a satisfying viewing experience.

Given that the importance of users' viewing feedback to service quality, [7] defines a new criterion named as QoE, which describes users' delight or annoyance levels when using an application or service. And the QoE was adopted by the ITU in 2016. Different from QoS, QoE aims to assess the streaming quality from users' perspective. The most intuitive way to obtain QoE is to conduct subjective tests under a controllable environment. Large-scale subjective tests provide a more common measurement of viewers' satisfaction but cost too much time and effort. It is tough to spread subjective test for QoE measurement of streaming services.

C. Objective Quality Model

To avoid time-consuming subjective ratings, Some researches [9]–[13], [15], [17]–[22], [24] propose objective QoE models to estimate views' experience in online streaming. These models are proposed via objective quality factor (e.g., the accumulation of physical errors) in a mathematical function which accord with a limited set of subjective data. Without the need for human resources, the whole process of QoE estimation can be completed on one end device by using objective quality model. As a result, in the context of video streaming, objective quality model is the more suitable choice rather than subjective test.

Based on Shannon's Information theory, a variety of models [10]–[15] directly correlate the streaming quality with the fidelity of transmitted content. These models leverage the objective video distortion to measure the subjective experience degradation. However, these methods ignore the viewer's perceptual experience. It becomes a hit that integrating content visual features into the quality model [16], [17]. Besides, quantified network descriptor are also taken into consideration [13], [18]–[21] to further emphasize the impact of network fluctuation on perceptual quality, which combines QoS and QoE models.

D. Data-driven Quality model

Since above methods emerge from scale-limited service feedbacks, the extracted models are commonly restricted to few quality distortion factors and hand-crafted feature representation. With the vast expansion of online streaming market, more service data are accessible for QoE research. Conventional quality representations hardly work with massive data in a general way. Besides, learning-based methods (e.g., support vector machine (SVM), convolution neural network (CNN)) show a more superior performance on content feature extraction. More researches exploit these architectures to handle QoE model in a data-driven way. Compared with previous researches, QoE study for online streaming is further categorized into two kinds: general QoE model which depicts overall video quality and continuous QoE that describes the instant perceived quality at any moment during video playback. Especially, for more accurate temporal feature analysis, Recurrent Neural Network (RNN) and its variants are widely used in continuous QoE metrics.

In this paper, we pay more attention the studies on Objective quality model and Data-driven quality model for online streaming QoE. The following sections are organized as follows. Section II introduces basic objective quality models and their application in online streaming QoE. In Section III, recent researches on data-driven quality model are analyzed. In Section IV, we conclude current QoE models and further summary the challenges and possible future direction in the field of QoE estimation for online streaming.

II. Objective Quality Model

Most objective quality models take the signal receiving and processing behavior of Human Visual System (HVS) into consideration for better perceptual quality description. HVS performs sensitivity to local contrast, high frequency and temporal shift, which inspires objective quality models integrating visual features into the measured function to accord with subjective ratings.

There are three kinds of models which are categorized with the existence of reference content, i.e., Full-reference (FR), Reduced-reference (RR), No-reference (NR) as shown in Fig. 1. Since FR and RR model require the original video as inference, they need additional payload to network service when inferring QoE. In contrast, NR model only uses the transmitted content for measurement which is more efficient for conduct QoE assessment online, especially in self-adaptive system (e.g., a streaming framework that can adjust its network configuration according to QoE evaluation). And this also boost the usage of NR model in the context of edge computing. However, the absence of reference may lead to less accuracy in quality measurement. The choice among FR, RR and NR models is a trade-off between online computational complexity and estimation accuracy.

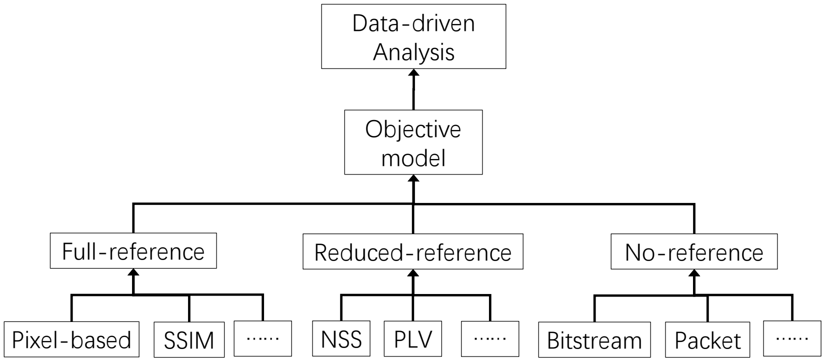

Fig. 1. Categorization of objective models.

A. Full-reference Model

Here, we introduce three classical full-reference objective quality models: pixel-wised model, Structural Similarity (SSIM) model and Natural Scene Statistics (NSS) model. The pros and cons of each model are specifically reviewed.

1) Pixel-wised Models: Mean Squared Error (MSE) and Peak Signal Noise Ratio (PSNR) [9] represent the pixel-wised distortion between the source and target image. They are commonly used as the benchmark due to the simple implementation.

MSE is denoted as,

$$MSE = \frac{1}{N}\sum_{i}(y_i - x_i)^2, \tag{1}$$

where x_i denotes the ith sample of the original signal, and y_i is the ith sample of the distorted signal. N is the length of the sequence.

To describe the signal with high dynamic range, PSNR is imposed in terms of the logarithmic decibel scale:

$$PSNR = 10\lg \frac{MAX}{MSE}, \tag{2}$$

where MAX is the maximum signal energy.

But pixel-wised models only quantify the objective distortion without any global or local feature analysis, which leads that high pixel-wised QoE index doesn't mean good perceptual quality.

2) SSIM Models: Considering HVS's characteristics, [16] proposed structural similarity (SSIM) to extract the structure representation (luminance, contrast, etc.) of the content from global view and later it is extended to multi-scale version [11]. [16] validates SSIM and it variants are strongly and positively correlated to the perceptual quality. The specific format of SSIM is shown as below,

$$\begin{aligned} \mathrm{SSIM}(\mathbf{x},\mathbf{y}) &= [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma}, \\ l(\mathbf{x},\mathbf{y}) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \\ c(\mathbf{x},\mathbf{y}) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \\ s(\mathbf{x},\mathbf{y}) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \end{aligned} \tag{3}$$

where $\mathbf{x}$ denotes the original video signals, and $\mathbf{y}$ is marked as the distorted ones. The luminance comparison function $l(\mathbf{x},\mathbf{y}), c(\mathbf{x},\mathbf{y}), s(\mathbf{x},\mathbf{y})$ respectively calculates the luminance, contrast and structure comparison between source and distorted signals. μ_* and σ_* are the mean and standard deviation of corresponding signals.

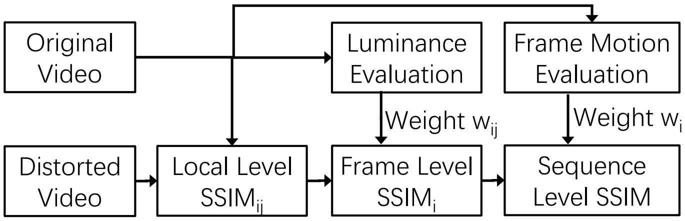

Fig. 2. video-SSIM model.

To further describe the SSIM of a sequence, video-SSIM [17] extend it with frame motion evaluation for a sequence-level QoE estimation, as shown in Fig. 2.

$SSIM_{ij}$ for local level is designed as a random-sampled 8x8 patch SSIM measurement according to Eq. (3). $SSIM_{ij}$ denotes the SSIM of jth window of ith frame:

$$SSIM_{ij} = W_Y SSIM_{ij}^{Y} + W_{Cb} SSIM_{ij}^{Cb} + W_{Cr} SSIM_{ij}^{Cr}, \tag{4}$$

where W_Y, W_{Cb} and W_{Cr} are weights of Y, Cb and Cr channels respectively.

Based on local-level SSIM, the frame-level SSIM is weighted sum of local-level SSIM of all windows in a frame, whose weights are determined by the sampled windows' illuminance. That is,

$$SSIM_i = \frac{\sum_j w_{ij} SSIM_{ij}}{\sum_j w_{ij}}, \tag{5}$$

where w_{ij} denotes the weights of $SSIM_{ij}$, and are determined as follow:

$$w_{ij} = \begin{cases} 0, & \mu_x \leq 40, \\ (\mu_x - 40)/10, & 40 < \mu_x \leq 50, \\ 1, & \mu_x > 50, \end{cases} \tag{6}$$

where μ_x is the mean of the Y components.

Sequence-level SSIM is a weighted sum of the frame-level SSIM across all the frames. The weight of one frame is determined by the motion with respect to the next frame. A motion-related parameter is proposed: $M_i = \sum\limits_j m_{ij}/(16N_i)$, where m_{ij} denotes the motion vector of jth window of ith frame and N_i denotes the number of windows in ith frame. The sequence-level SSIM is defined as follows:

$$SSIM = \frac{\sum_{i} W_i SSIM_i}{\sum_{i} W_i}, \tag{7}$$

where weights W_i is determined as follow:

$$W_i = \begin{cases} \sum_{j} m_{ij}, & M_i \leq 0.8, \\ (3 - 2.5M_i) \sum_{j} m_{ij}, & 0.8 < M_i \leq 1.2, \\ 0, & M_i > 1.2, \end{cases} \tag{8}$$

3) NSS Models: Different from random signals, videos are natural scenes with their unique statistical information named as naturalness [14], indicating that model based only on Shannon information fidelity is not the best choice for QoE evaluation. When compressed, the video quality is degraded and the naturalness is destroyed by these artifacts. That is to say, if the distorted video's statistical information is more natural, the perceptual quality of the video is considered to be higher. NSS can also serve as the basic VQA for online QoE estimation models.

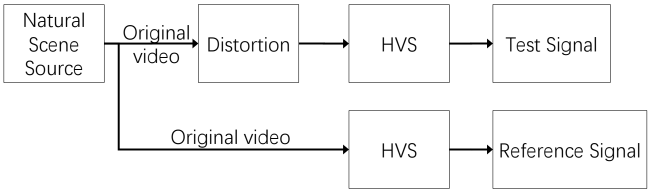

Fig. 3. Distortion model in VIF.

A representative example is Visual Information Fidelity (VIF) [15].

The distorted video signal $\mathcal{D}$ is defined as,

$$\mathcal{D} = a\mathcal{S} + \mathcal{B}, \tag{9}$$

where $\mathcal{S}$ denotes source video signals. a represents a deterministic scalar gain (attenuation) field and I represents all the spatiotemporal blocks of the video sequence. $\mathcal{B}$ is a stationary additive zero-mean Gaussian noise. Let $\mathcal{T}$ denotes test signals processed by HVS and $\mathcal{R}$ denotes reference signals processed by HVS as shown in Fig. 3. Human brain extract information from test signals and reference signals. $\mathcal{T}, \mathcal{R}$ are defined as follows:

$$\begin{aligned} \mathcal{T} &= \mathcal{D} + \mathcal{N}, \\ \mathcal{R} &= \mathcal{S} + \mathcal{N}', \end{aligned} \tag{10}$$

where $\mathcal{N}$ and $\mathcal{N}'$ are the noise from HVS channel. Information that can be extracted from on channel of reference signals and test signals are denoted as $I_{\mathcal{R}}$ and $I_{\mathcal{T}}$, which are defined as follows:

$$\begin{aligned} I_{\mathcal{R}} &= \frac{1}{2} \sum_{i \in I} \log_2(1 + \frac{s_i^2}{\delta_n^2}), \\ I_{\mathcal{T}} &= \frac{1}{2} \sum_{i \in I} \log_2(1 + \frac{a_i^2 s_i^2}{\delta_b^2 + \delta_n^2}), \end{aligned} \tag{11}$$

where δ_b and δ_n are the variances of $\mathcal{B}$ and $\mathcal{N}$ respectively. Thus, visual fidelity VIF is the ratio of total $I_{\mathcal{T}}$ to total $I_{\mathcal{R}}$:

$$VIF = \frac{\sum_{\text{allchannels}} I_{\mathcal{T}}}{\sum_{\text{allchannels}} I_{\mathcal{R}}}, \tag{12}$$

As shown in Eq. (12), VIF evaluates the visual fidelity of distorted videos.

B. Reduced-reference model

Compared to FR model, RR model only need partial features of original video to infer quality, which reducing extra payload in transmission. In this section, we present the Packet Loss Visibility- (PLV-) based and the pixel-based RR model.

TABLE I
PLV FACTORS CATEGORIZATION

Content-independent factors	content-dependent factors
temporal duration	Variance of motion
Initial spatial extent	Initial MSE
vertical position	residual energy

1) Packet Loss Visibility: In video streaming service, packet loss induces the degradation of video quality. And different absent packet cause various impairment levels. The packet loss at an I frame more severely damages the video quality than packet loss at a B/P frame since I frame is a complete image and the reference of B/P frame. The major problem in PLV based RR model is to find the position of missing packet for impairment definition. Commonly, the PLV model is extracted with supervision of manual perceptual labels and then it could directly estimate the visibility of packet loss in a classified or regressed way.

The degree of impairment to video quality may vary depending on the type of packet loss (e.g.,a degraded still scene is more perceivable than a degraded moving scene). As result, the content-independent factors and content-dependent factors are proposed and adopted as the features that determine the visibility of packet loss as shown in Table I. Content-independent factors can be extracted from distorted videos. Content-dependent factors can be estimated with the help of reduced information of the original videos, inserting less load to the network. [12] imposed Classification And Regression Tree (CART) to classify the packet loss. However it's hard to distinguish packet loss visibility near the threshold or far from the threshold. In [13], General Linear Model (GLM) is used to predict the probability that the packet loss is visible, which is widely adopted in further research in this field.

2) Pixel-based Model: While PLV-based model only estimates the packet loss's visibility and couldn't give a quantified measurement of quality degradation, pixel-based model can quantify the degree of the degradation of video quality. Optical flow [23] measures motion vector of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene. [22] proposed a A optical flow-based QoE model to describe the degradation from the original video to that distorted one. Let U_k, V_k denote the kth frame's matrix of horizontal and vertical optical flow velocities respectively. R_k records the magnitudes of the kth optical flow velocities,

$$\mathbf{R}_k = \sqrt{\mathbf{U}_k^2 + \mathbf{V}_k^2}, \tag{13}$$

A reference descriptor of an optical flow map is generated from the iteration algorithm proposed. Initially, $T(\mathbf{R}_k,\ 0) = R_k$, $\delta(\mathbf{R}_k,\ 0) = \sum_{i=1}^{N-1} \sum_{j=0}^{M-1} |T(\mathbf{R}_0,\ t)|$, where t is the iteration index. During every iteration,

$$\begin{aligned} T(\mathbf{R}_k,\ t) &= \downarrow 2[T(\mathbf{R}_k,\ t-1) * \phi(\sigma)], \\ \delta(\mathbf{R}_k,\ t) &= \delta(\mathbf{R}_k,\ t-1) + \sum_{i=1}^{N-1} \sum_{j=0}^{M-1} |T(\mathbf{R}_k,\ t)|, \end{aligned} \tag{14}$$

where $\phi(\sigma)$ is a Gaussian kernel with standard deviation σ and $\downarrow 2[\cdot]$ is a downsampling operator. After all the iterations, a descriptor of an optical flow map $\delta(\mathbf{R}_k)$ is calculated:

$$\delta(\mathbf{R}_k) = \delta(\mathbf{R}_k,\ t), \tag{15}$$

Let $\delta(\mathbf{R}_{k-ref})$ and $\delta(\mathbf{R}_{k-test})$ denote the reference descriptor and the test descriptor respectively. Then a perceived quality descriptor P is defined based on the difference of $\delta(\mathbf{R}_{k-ref})$ and $\delta(\mathbf{R}_{k-test})$ as follows:

$$\begin{aligned} D_k &= |\delta(\mathbf{R}_{k-\mathrm{ref}}) - \delta(\mathbf{R}_{k-\mathrm{r-test}})|, \\ P &= \log\left[\underset{\forall k}{\mathbf{E}}[D_k]\right], \end{aligned} \tag{16}$$

where $\mathbf{E}$ denotes the expectation.

C. No-reference Model

NR model does not require the reference to original videos for assessment and could obtain real time QoE evaluation compared to RR and FR model.

Without having access to original videos, NR models pays more attention to measure QoE with known network parameters. These models could be categorized into three kinds based on the usage of network, i.e., bitstream layer, packet layer models and hybrid models that can be regarded as the mixture of other models.

1) Bitstream-layer Model: Bitstream-layer videos quality assessment is the common choice especially when videos are decoded with limited resources. In addition to packet layer information, bitstream-layer model utilizes further information about the encode bitstream. Compared to pixel-based model, bitstream-layer model has access to a lot more information, such as Quantization Parameter (QP), motion vectors, bit rate and frame rate, which helps to improve the correlation between the evaluated QoE and the quality perceived by human beings. However, more required information for a bitstream-layer model means the higher computational complexity. For the context of edge computing, a low-complexity solution is often preferred to evaluate QoE in a real-time manner. A classical method is exemplified as below.

QANV-PA [18] mainly focuses on estimating the QoE with Real-time Transport Protocol (RTP) or User Datagram Protocol (UDP). Unlike TCP with backoff and retransmission mechanisms, RTP/UDP based video services are far more likely to suffer from packet loss, which is exploited in [18].

With QANV-PA, the primary analysis begins after a simple decoding of the packet and the video frame header. The obtained information includes QP, frame type, bit-rate for each frame, lost packets, and the frame display duration. Considering the QP may fluctuate with macroblocks in a frame, the QP extracted from the header serves as a good approximation of the average QP of the frame, which makes it a great indicator of the video quality. Then the degradation caused purely by quantization is independently estimated by an objective quality model. Rest of the degradation that are caused by packet loss or error propagation is also additionally estimated. Together, the two parts of degradation are combined into the whole degradation of frame quality. A pooling function is employed to incorporate both spatial and temporal degradation. The quality of the video sequence is a weighted sum of the quality of each frame and the weights are determined by display time of each frame. The nth frame's quality that is determined by frame and coding distortion is:

$$Q_{F,C,n} \ = f(q_n) + (b_3 - f(q_n)) \left(\left(\frac{\delta_{S,n}}{a_1} \right)^{b_1} + \left(\frac{\delta_{T,n}}{a_2} \right)^{b_2} \right), \tag{17}$$

where $f(q_n)$ is a linear function of QP parameters q_n which is obtained from packet and video frame header. $Q_{F,C,n}$ denotes the packet-loss-free coded frame quality of nth frame, and the subscripts F and C denote frame and coding distortion. $\delta_{T,n}$ and $\delta_{S,n}$ denote temporal and spatial complexity respectively. a_1, a_2, b_1, b_2, b_3 are constants parameters.

Let p_n denotes the degradation caused by the packet loss. Factors that determine p_n include the number of frames affected by the packet loss and the temporal complexity of nth frame $\delta_{T,n}$. Thus, the quality of nth frame $Q_{F,n}$ and the contribution of nth frame to the quality of sequence Q_{FtoS} are defined as follows:

$$Q_{F,n} = Q_{F,C,n} - p_n, \tag{18}$$

$$Q_{FtoS} = Q_{F,n} \left(a_4 + b_4 \delta'_{T,n} + c_4 \delta'_{T,n} log(T_n) \right), \tag{19}$$

where $\delta'_{T,n} = \delta_{T,n} / max(\delta_T)$ is the normalized temporal complexity.

As a result, the quality of the sequence is:

$$QANV - PA = \frac{\sum\limits_{n \in D} (Q_{FtoS} T_n)}{\sum\limits_{n \in D} T_n}, \tag{20}$$

where D denotes the set of successfully decoded frames, and T_n is the display time of nth frame.

2) Packet-layer Model: This type of models uses only the information from packet header for quality estimation, without exploiting the payload information. When payloads are encrypted, packet-layer model is more suitable comparing to the bitstream-layer model that requires the information of payloads. Since the packet-layer model does not depend on information from payload, its computational complexity is low, enabling lightweight measurement without resorting to media. Two typical models are introduced below.

V_q [19] estimates the quality affected by the packet loss. As the bitrate increased, V_q increased and saturated. This situation could be approximated by a logistic function. In order to calculate V_q, the first step is to estimate the quality with no Packet Loss (PL):

$$V_q|_{PL=0} = 1 + a_1 - \frac{a_1}{1 + (Br/a_2)^{a_3}}, \tag{21}$$

where Br denotes the bitrate and $a1, a2, a3$ are constant parameters. When $PL \neq 0$

$$V_q = 1 + (a_1 - \frac{a_1}{1 + (Br/a_2)^{a_3}}) \exp \left(- \frac{PL}{a_4} \right), \tag{22}$$

where a_4 is a constant parameter.

Content-Adaptive Packet Layer (CAPL) model proposed in [20] is another model on the basis of the above-mentioned bitstream-layer model, QANV-PA. Due to lack of payload information, the quality of nth frame affected by coding distortion $Q_{F,C,n}$ and temporal complexity $\delta_{T,n}$ have to be computed in different ways. $Q_{F,C,n}$ and $\delta_{T,n}$ are defined as follows:

$$\begin{aligned} Q_{F,C,n} &=, 1 + a_1 \left(1 - (\frac{R_n}{a_2 \delta_{T,n} + b_2})^{-b_1} \right), \\ \delta_{T,n} &=, |R_{P,n}/R_{I,n} - a_3|, \end{aligned} \tag{23}$$

where a_1, a_2, b_1, b_2 are constant parameters. R_n is the average number of bit allocation for a frame in Group of Pictures (GoP). $R_{P,n}$ and $R_{I,n}$ are the average bit allocation for the P frame and I frame in GoP respectively. The rest of flow of CAPL is similar to QANV-PA.

3) Hybrid Model: Measurement of online streaming QoE is related to packed loss, video codec, loss recovery technique, encoding bitrate and content characteristics. Measuring these features requires complex computation and is not suitable for large-scale online QoE monitoring. As a result, [21] proposed rPSNR, a lightweight NR model. rPSNR bypass the complex computation by measuring video quality against a quality benchmark that the network is expected to provide. Videos distortion D is measured through MSE, which is defined as follows:

$$D = P_e f(n) L D_1 \tag{24}$$

where P_e is the probability of packet loss event in video streaming. $f(n)$ is the average number of slices affected by pakcet loss; L is the number of packets used to transmit one frame; D_1 is the total average distortion caused by losing a slice and depends on the characteristics of an individual video, meaning that it may not be efficiently estimated when trying to fulfill the real-time quality monitoring of a huge amount of streamed videos. $f(n)$ varies with codec. Before H.264 if packet loss happens, the frame will be discarded. More sophisticated error concealment is adopted in H.264. All slices will be decoded, and those who are affected by the packet loss can be recovered by the help of the corresponding slices in previous slice and the motion information got frame other slices in the same frame. The estimation of D_1 depends on the error propagation caused by loss of one slice because of coding dependencies between frames. P_e and $f(n)$ are network-dependent factors. They are easy to obtain from network statistics. L is determined based on the configuraion. The best way to tackle this problem is to maintain a reference path whose QoE is already known. By comparing the quality of videos transmitted over a path with that transmitted over a reference path, we can tell if it's better or worse than the reference path. Relative PSNR is the difference between the monitored path and the reference path:

$$PSNR = 10 \lg \frac{255^2}{D} \tag{25}$$

$$rPSNR = PSNR - PSNR_{ref}, \tag{26}$$

As a consequence, rPSNR is not dependent on D_1, meaning that it is relatively easy to compute.

III. Data-driven quality model

With the prosperous development of video streaming services, lots of data have been generated, intensifying the urge for new understanding of video quality assessment and breakthrough in related research. Also by the assistance of booming development in the field of machine learning (ML), researchers can process a big amount of data in many different ways. Data-riven analysis is a learning-oriented framework for QoE estimation which is built on the basis of objective quality models (FR, RR, NR models). Since we still cannot accurately understand and explain the mechanism of the HVS, the adoption of ML can help researchers simulate HVS behaviors and design a more accurate model to measure QoE. These learning-oriented methods can be categorized into three groups based on their inputs: visual-feature based models, network-feature based models, and the hybrid models. In this section, we will introduce the three kinds of learning-oriented methods and some new challenges in recent scenarios.

A. Visual-feature based models

It's intuitive to adopt visual-feature as inputs of ML models designed to measure the quality of streaming videos, especially when the reference video is available. Since a video is consisted of a series of images, it's natural to use visual features as inputs for the deep learning models [25]–[29].

There are certain distortions that cause deterioration of video quality: ringing effects, which appear as bands or ”ghosts” near edges; blocking effects, which is caused by lossy compression algorithms; clipping, which is the truncation in the number of bits of the luminance or chrominance components of the image values, and contrast. Similar to Multi-Metrics Fusion (MMF) based models, the model presented in [27] takes four visual features as inputs: clipping, constrast, blocking4 (blocking effect with block size 4×4), blocking8 (blocking effect with block size 8×8), then let the inputs through a neural network (NN) with one middle hidden layer of 15 units. Since more features are adopted in this model, the performance is better than conventional models. The output of the NN is the general QoE of the video. Though the accuracy is attained in terms of goodness of fit $R = 0.8785$, the model's generality remains a drawback, indicating a worse performance when handling videos with other distortions.

Unlike the model in [27] using hand-crafted features, the model presented in [28] extracts frame features by a pretrained Convolutional Neural Network (CNN) which enables researchers to make use of a lot more information to measure the QoE. Then the $d \times N$ sequence data created by the CNN is fed into an Long-Short Term Memory (LSTM) network to predict video quality, where d stands for the length of the video sequence and N is the length of frame-level deep feature vector. The sequence-to-one regression is accomplished by a fully connected (FC) layer after the LSTM, in order to calculate the general QoE of the video. LSTM is an artificial RNN architecture with the capability of preserving long term and short term information, suitable for handling sequence data. CNN is proved to be efficient in extracting features. Various CNN architectures have been proposed to improve the performance, three of which are tested in [28]: AlextNet [30], Inception-V3 [31], Inception-ResNet-V2 [32]. Judging from the test result, Inception-V3's features perform slightly better than Inception-ResNet-V2's features while AlextNet's features give significantly poorer results than other's features.

Apart from dealing with spatial features and temporal features with CNN and RNN respectively, authors in [29] analyse spatial and temporal patterns together in order to get more complex spatio-temporal patterns. This can be accomplished by 3D convolutions, where the first convolutional axis is along the time direction, and the second and third axes are used for the spatial dimensions of the video frame. The model proposed in [29] is a FR one, meaning that reference video is available. The inputs of the model are the luminance channel of distorted frames and residual frames. The residual frames are the differences of distorted frames and reference frames. 2D CNN for spatial features extraction and 3D CNN for spatio-temporal features extraction are used in the network, followed by one global average pooling layer and two fully connected layers for sequence-to-one regression and calculation of the general QoE.

B. Network-feature based models

Apart from distortions introduced in videos during coding and decoding, transmission quality in the network is also a major factor of video QoE in the context of streaming. Hence, network-feature based models are proposed to measure or predict the QoE [33], [35], [36].

In [33], an artificial neural network (ANN) based architecture is proposed, with network delay, jitter, loss and Mean Opinion Score (MOS) as inputs. The output of the ANN is the general QoE of the video.

In contrast, the model proposed in [35] predicts the continuous QoE, which is much more useful in the context of QoE monitoring. The prediction of continuous QoE tends to be more difficult than general QoE, owing to relatively lacking information and the requirement of real-time QoE evaluation. Continuous QoE is mainly determined by playback interruption [18] and recency effect. the predictor in the model depends on non-linear autore- gressive with exogenous variables (NARX) which can preserve previous information and handle the recency effect.

C. Hybrid models

Since the more information would lead to more accurate result, most of the models tend to utilize both visual and network features. Though in the context of video streaming, the computation of perceptual quality of video frames often requires great amount of resources and probably availability of reference video, researchers adopt RR-VQA or NR-VQA to bypass the limitations [37]–[39].

Authors in [37] proposed a LSTM based model. The multi-layered multi-unit LSTM network in the model is a cascade of several LSTM units that are stacked up to constitute LSTM layers. Features selected as inputs can be categorized as Short Time Subjective Quality (STSQ), Playback Indicator (PI), Time Elapsed Since Last Rebuffering (T_R). To reduce the computational complexity, spatio-temporal RR entropic differences (STRRED) is selected as STSQ feature. These features help the model learn the spatio-temporal pattern more efficiently.

Though LSTM performs better than naive RNN when dealing with sequential input, it has more parameters and need longer time to train. As a result, temporal convolutional network (TCN) or CNN based models are proposed in [39] to improve the performance. TCN utilizes 1D convolution where the filters exhibit only one dimension (time) instead of two dimensions (width and height). Furthermore, 1D convolutions are well-suited for real-time tasks due to their low computational requirements. Also, casual convolutions, a convolution layer that ensure there is no information "leakage" from future into past, is introduced to TCN.

IV. Conclusion and future trend

This paper presents a survey on QoE estimation of online video streaming. Subjective test is the most direct way to measure the QoE, though with a lot of limitations. Objective quality model estimates QoE by exploiting the relation between QoE and QoS metrics (or HVS features), relying on subjective test. With the prosperous development of video streaming service, data mining techniques and machine learning, data-driven QoE model emerges. It is capable of coping with more features than other models. As a result, data-driven QoE model should be the frontier of future research.

However, most of the databases nowadays are not suitable for the research in continuous QoE. As a result, more comprehensive database should be built. Thanks to the prosperous market of video streaming and live streaming, more and more data would be generated. Only if all those data be categorized to form a database, scattered data is almost useless. As shown in [36], context parameters can help to estimate QoE. Different combinations of input features and machine learning approaches can lead to different performance. Combining the QoS parameters with features at media-level or other context parameters might lead to a better QoE model. As for RNN-based models that foucs on memory effects of continuous QoE, the development of RNN and its variants is promising. Adoption of better RNN variants such as SRNN could also improve the performance.

With the development of security and privacy preservation in the field of online streaming, less information is available for QoE estimation. There are lots of researches about QoE monitoring of encrypted video streaming. Most of them are designed for HTTPS based video streaming [40]–[45]. Majority of the models tackle the challenge of lacking metrics and information by adopting ML for inferring key performance indicators from observable traffic patterns and statistics. However, with the blossoming of blockchain, another way of improving security, more and more researches about video streaming via blockchain have been conducted. Yet, there are not much work done on QoE estimation in the context of blockchain.

ACKNOWLEDGMENT

This work is supported by the National Key Research and Development Program of China under Grant No. 2018YFB2100804, the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 898588, the National Natural Science Foundation of China under Grant No. 61902178, No. 62102053, No. 92067206 and No. 61972222, the Natural Science Foundation of Jiangsu under Grant BK20190295, the Leading Technology of Jiangsu Basic Research Plan under Grant BK20192003, and the Chongqing Key Laboratory of Digital Cinema Art Theory and Technology under Grant No. 2021KF01. This article reflects only the authors' view. The European Union Commission is not responsible for any use that may be made of the information it contains. X. Zhang is the corresponding author.

REFERENCES

[1] Limelight Networks, ”MARKET RESEARCH: THE STATE OF ONLINE VIDEO 2019,” Online: https://www.limelight.com/resources/white-paper/state-of-online-video-2019/

[2] Grand View Research, ”Video Streaming Market Size, Share & Trends Analysis Report By Streaming Type, By Solution, By Platform, By Service, By Revenue Model, By Deployment Type, By User, And Segment Forecasts, 2020 - 2027,” Online: https://www.grandviewresearch.com/industry-analysis/video-streaming-market

[3] X. Zhang, H. Chen, Y. Zhao, Z. Ma, H. Huang, Y. Xu, H. Yin and D. O. Wu. Improving Cloud Gaming Experience through Mobile Edge Computing. in IEEE Wireless Communications, vol. 26, no. 4, pp. 178-183, August 2019.

[4] H. Chen, X. Zhang, Y. Xu, J. Ren, J. Fan, Z. Ma, W. Zhang. T-Gaming: A Cost-Efficient Cloud Gaming System at Scale. in IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 12, pp. 2849-2865, 1 Dec. 2019.

[5] Z. Jiang, X. Zhang, W. Huang, H. Chen, Y. Xu, J.-N. Hwang, Z. Ma, J. Sun. A Hierarchical Buffer Management Approach to Rate Adaptation for 360-Degree Video Streaming. in IEEE Transactions on Vehicular Technology, vol. 69, no. 2, pp. 2157-2170, Feb. 2020.

[6] ITU, ”E.800: Terms and definitions related to quality of service and network performance including dependability,” ITU-T Recommendation. August 1994. Retrieved October 14, 2011. Updated September 2008 as Definitions of terms related to quality of service

[7] Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds., ”Qualinet White Paper on Definitions of Quality of Experience (2012),” European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Mar. 2013

[8] H. Nam, K. Kim and H. Schulzrinne, ”QoE matters more than QoS: Why people stop watching cat videos,” IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications, San Francisco, CA, 2016, pp. 1-9.

[9] Huynh-Thu, Quan, and Mohammed Ghanbari. ”Scope of validity of PSNR in image/video quality assessment.” Electronics letters 44.13 (2008): 800-801.

[10] Z. Wang and A. C. Bovik, “A universal image quality index,” IEEE Signal Process. Lett., vol. 9, no. 3, pp. 81–84, Mar. 2002.

[11] Z.Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[12] A. R. Reibman, S. Kanumuri, V. Vaishampayan, and P. C. Cosman, “Visibility of individual packet losses in MPEG-2 video,” in Proc. ICIP, 2004, vol. 1, pp. 171–174.

[13] S. Kanumuri, P. Cosman, and A. R. Reibman, “A generalized linear model for MPEG-2 packet-loss visibility,” in Proc. 14th Int. PV Workshop, 2004, pp. 1–9.

[14] Reinagel P, Zador A M. Natural scene statistics at the centre of gaze[J]. Network: Computation in Neural Systems, 1999, 10(4): 341.

[15] H. R. Sheikh and A. C. Bovik, “A visual information fidelity approach to video quality assessment,” in Proc. 1st Int. Workshop Video Process. Qual. Metrics Consum. Electron., 2005, pp. 23–25.

[16] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, ”Image quality assessment: from error visibility to structural similarity,” in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.

[17] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” Signal Process. Image Commun., vol. 19, no. 2, pp. 121–132, Jan. 2004.

[18] F. Yang, S. Wan, Q. Xie, and H. R. Wu, “No-reference quality assessment for networked video via primary analysis of bit stream,” IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 11, pp. 1544–1554, Nov. 2010.

[19] K. Yamagishi and T. Hayashi, “Parametric packet-layer model for mon- itoring video quality of IPTV services,” in Proc. IEEE ICC, 2008, pp. 110–114.

[20] F. Yang, J. Song, S. Wan, and H. R. Wu, “Content-adaptive packet-layer model for quality assessment of networked video services,” IEEE J. Sel. Topics Signal Process., vol. 6, no. 6, pp. 672–683, Oct. 2012.

[21] S. Tao, J. Apostolopoulos, and R. Guérin, “Real-time monitoring of video quality in IP networks,” IEEE/ACM Trans. Netw., vol. 16, no. 5, pp. 1052–1065, Oct. 2008.

[22] M. A. Aabed and G. AlRegib, ”Reduced-reference perceptual quality assessment for video streaming,” 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, 2015, pp. 2394-2398.

[23] Horn B K P, Schunck B G. Determining optical flow[J]. Artificial intelligence, 1981, 17(1-3): 185-203.

[24] R. Soundararajan, A. C. Bovik, ”Video quality assessment by reduced reference spatio-temporal entropic differencing”, IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 4, pp. 684-694, Apr. 2013.

[25] S. Ahn and S. Lee, ”Deep Blind Video Quality Assessment Based on Temporal Human Perception,” 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 619-623, doi: 10.1109/ICIP.2018.8451450.

[26] Hou R, Zhao Y H, Hu Y, et al. No-reference video quality evaluation by a deep transfer CNN architecture[J]. Signal Processing: Image Communication, 2020, 83: 115782.

[27] Rohil, M.K., Gupta, N. and Yadav, P. An improved model for no-reference image quality assessment and a no-reference video quality assessment model based on frame analysis. SIViP 14, 205–213 (2020).

[28] Varga, D., Szirányi, T. No-reference video quality assessment via pretrained CNN and LSTM networks. SIViP 13, 1569–1576 (2019).

[29] M. Xu, J. Chen, H. Wang, S. Liu, G. Li and Z. Bai, ”C3DVQA: Full-Reference Video Quality Assessment with 3D Convolutional Neural Network,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 4447-4451, doi: 10.1109/ICASSP40776.2020.9053031.

[30] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

[31] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

[32] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

[33] T. Begluk, J. B. Husić and S. Baraković, ”Machine learning-based QoE prediction for video streaming over LTE network,” 2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH), 2018, pp. 1-5, doi: 10.1109/INFOTEH.2018.8345519.

[34] Tsungnan Lin, B. G. Horne, P. Tino and C. L. Giles, ”Learning long-term dependencies in NARX recurrent neural networks,” in IEEE Transactions on Neural Networks, vol. 7, no. 6, pp. 1329-1338, Nov. 1996.

[35] C. G. Bampis, Z. Li, I. Katsavounidis and A. C. Bovik, ”Recurrent and Dynamic Models for Predicting Streaming Video Quality of Experience,” in IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3316-3331, July 2018.

[36] V. Vasilev, J. Leguay, S. Paris, L. Maggi and M. Debbah, ”Predicting QoE Factors with Machine Learning,” 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, 2018, pp. 1-6.

[37] Eswara N, Ashique S, Panchbhai A, et al. Streaming video QoE modeling and prediction: A long short-term memory approach[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[38] Duc, Tho Nguyen, et al. ”Bidirectional LSTM for continuously predicting QoE in HTTP adaptive streaming.” Proceedings of the 2019 2nd International Conference on Information Science and Systems. 2019.

[39] T. N. Duc, C. T. Minh, T. P. Xuan and E. Kamioka, ”Convolutional Neural Networks for Continuous QoE Prediction in Video Streaming Services,” in IEEE Access, vol. 8, pp. 116268-116278, 2020, doi: 10.1109/ACCESS.2020.3004125.

[40] Dimopoulos G, Leontiadis I, Barlet-Ros P, et al. Measuring video QoE from encrypted traffic[C]//Proceedings of the 2016 Internet Measurement Conference. 2016: 513-526.

[41] Orsolic I, Pevec D, Suznjevic M, et al. A machine learning approach to classifying YouTube QoE based on encrypted network traffic[J]. Multimedia tools and applications, 2017, 76(21): 22267-22301.

[42] Gutterman C, Guo K, Arora S, et al. Requet: Real-time QoE detection for encrypted YouTube traffic[C]//Proceedings of the 10th ACM Multimedia Systems Conference. 2019: 48-59.

[43] Mangla T, Halepovic E, Ammar M, et al. emimic: Estimating http-based video qoe metrics from encrypted network traffic[C]//2018 Network Traffic Measurement and Analysis Conference (TMA). IEEE, 2018: 1-8.

[44] Seufert M, Casas P, Wehner N, et al. Stream-based machine learning for real-time QoE analysis of encrypted video streaming traffic[C]//2019 22nd Conference on innovation in clouds, internet and networks and workshops (ICIN). IEEE, 2019: 76-81.

[45] Orsolic I, Skorin-Kapov L. A framework for in-network qoe monitoring of encrypted video streaming[J]. IEEE Access, 2020, 8: 74691-74706.