Article

Video Quality of Experience Metric for Dynamic Adaptive Streaming Services Using DASH Standard and Deep Spatial-Temporal Representation of Video

Lina Du ^{1,2} , Li Zhuo ^{1,2,*}, Jiafeng Li ^{1,2}, Jing Zhang ^{1,2} , Xiaoguang Li ^{1,2} and Hui Zhang ^{1,2}

¹ Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China; Dulina@emails.bjut.edu.cn (L.D.); lijiafeng@bjut.edu.cn (J.L.); zhj@bjut.edu.cn (J.Z.); lxx@bjut.edu.cn (X.L.); huizhang@bjut.edu.cn (H.Z.)

² College of Microelectronics, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

* Correspondence: zhuoli@bjut.edu.cn; Tel.: +86-010-67392749

Received: 23 December 2019; Accepted: 2 March 2020; Published: 5 March 2020

Abstract: DASH (Dynamic Adaptive Streaming over HTTP (HyperText Transfer Protocol)) as a universal unified multimedia streaming standard selects the appropriate video bitrate to improve the user's Quality of Experience (QoE) according to network conditions, client status, etc. Considering that the quantitative expression of the user's QoE is also a difficult point in itself, this paper researched the distortion caused due to video compression, network transmission and other aspects, and then proposes a video QoE metric for dynamic adaptive streaming services. Three-Dimensional Convolutional Neural Networks (3D CNN) and Long Short-Term Memory (LSTM) are used together to extract the deep spatial-temporal features to represent the content characteristics of the video. While accounting for the fluctuation in the quality of a video caused by bitrate switching on the QoE, other factors such as video content characteristics, video quality and video fluency, are combined to form the input feature vector. The ridge regression method is adopted to establish a QoE metric that enables to dynamically describe the relationship between the input feature vector and the value of the Mean Opinion Score (MOS). The experimental results on different datasets demonstrate that the prediction accuracy of the proposed method can achieve superior performance over the state-of-the-art methods, which proves the proposed QoE model can effectively guide the client's bitrate selection in dynamic adaptive streaming media services.

Keywords: mobile video; quality of experience; metric; deep spatial-temporal representation; DASH

1. Introduction

In recent times, there has been a tremendous growth in the field of mobile communication technology and a tremendous growth has been observed in the mobile video services along with emerging new services. Mobile video service is a major contributor to mobile traffic, which has attracted attention for both industrial and academic research. Approximately 5 billion videos are viewed on YouTube [1] daily. As per the prediction of the Cisco Systems of the United States, the Content Delivery Network (CDN) traffic will account for more than 71% of the total network traffic, whereas the mobile video service traffic will occupy more than 75% of the total mobile internet traffic [2] by 2020. Mobile video services generally use HTTP adaptive streaming technology to improve the user's Quality of Experience (QoE). However, quantitative expression of the user's QoE is also a difficult point in itself. The accurate assessment of the QoE of the users has become a point of concern for the mobile video service providers and the network operators. It is also important to establish a metric that can accurately assess the QoE of the users.

QoE refers to a user's experience on the quality and the performance of the device, network, system, application and services [3], which reflects the satisfaction or the comfort of the users while using the service. In [4], QoE has been defined as "The degree of joy or annoyance experienced by a user while using an application or service." Compared to Quality of Service (QoS), QoE takes the "human" factor into account for focusing on the user's subjective perception of service quality.

There are various influencing factors that may affect QoE from all aspects of the communication system. Many researchers have proposed different QoE assessment models to study the factors influencing QoE for mobile video services for obtaining remarkable results. For exploring mobile video QoE assessment, it is essential to find a relationship between the influencing factors and the Mean Opinion Score (MOS) value. The modeling process is mathematically expressed as:

$$Y = f(X) \quad (1)$$

where Y refers to the user's QoE that is usually measured using the MOS value and X refers to the subjective and objective factors that influence the QoE. Hence, video quality assessment can be described as a multi-index assessment problem and the key to modeling is to find an optimal mapping relationship, i.e., $f(\cdot)$.

In order modelling the QoE, we firstly discuss the assessment of QoE for mobile video services, which includes video quality and the user's experience.

- Video quality

Video quality can be assessed in a subjective or objective manner. Subjective quality assessment methods allow the observers to make direct judgements on the quality of the video. The subjective assessment used in this study is the MOS. It involves classification of subjective feelings experienced by a person on watching a video into five different levels, in which each level represents a certain range of acceptance. The subjective assessment results are highly accurate and reflect a user's intuitive experience; however, they are difficult to implement on a large scale in practical applications.

The objective video quality assessment is evaluated by reconstructing the distortion degree of the video in comparison to the original reference video. Hence, the objective assessment method does not consider the user's intuitive experience, which cannot reflect a user's experience directly and accurately. The objective method is further divided into three types, namely, Full Reference (FR), Reduced Reference (RR) and No Reference (NR), based on the degree of dependence on the original video.

- User's Experiences

The criteria for measuring the QoE of a video service involves studying the user's engagement, such as the percentage of playback time during the duration of the video and the user's access times.

Currently, two methods are used to devise the QoE models, namely, the optimal mathematical modeling method and the machine learning-based method. The representative methods of these two categories are introduced below.

1.1. QoE Assessment Model Based on Optimal Mathematical Modeling Method

The steps involved in this method are as follows: Firstly, a number of parameters that characterize the influencing factors are chosen; secondly, the functional model is set up; and finally, the coefficients of the model are determined to obtain optimum prediction performance. The QoE assessment model that is devised using this method usually adopts an exponential or logarithmic form.

Reference [5] is one of the earliest works that incorporated building a QoE model for HAS applications. This model quantifies QoE as a linear model based on factors such as initial delay, re-buffering frequency and duration. The exponential model can be used to map a relationship between the influencing factors and the MOS; whereas, in reference [6], the re-buffering frequency and duration

are also considered. eMOS [7] considers re-buffering time, count and the video coding bitrates as major influencing factors and establishes an exponential model between the three influencing factors and the MOS, separately. The three models are linearly combined to obtain the final QoE prediction model. Reference [8] studies the effects of the three different influencing factors, initial delay, re-buffering time and video quality switching on QoE, and establishes an exponential model for each of the three influencing factors and the MOS. These are linearly combined to obtain the QoE prediction model. Reference [9] focuses on the impact of interruption in the video playback process on QoE of the streaming media services and establishes an exponential model between the MOS and the duration of interruption. Reference [10] also studied the effect of switching of the quality of a video QoE, and concluded that the switching frequency between different Video Quality Levels (VQL) interferes with the user's attention. It also affects a user's QoE, and by combining the exponential and logarithmic model the effect of video quality switching on the user's QoE can be quantified. It is observed that when the frequency is greater than 1/14 per second, the value of MOS drops significantly. Reference [11] uses a linear combination of multiple exponential models to establish a relationship between QoE and video compression, initial delay and re-buffering time, separately.

Hence, it can be summarized that only few influencing factors can be considered in this model owing to limitations in parameters that are taken into consideration. Thus, this method is unable to establish a relationship between the complex influencing factors and QoE; the accuracy of this model is typically low. These QoE models usually cannot effectively guide the client's bitrate selection in dynamic adaptive streaming services.

1.2. QoE Assessment Model Based on Machine Learning

The machine learning-based method uses many training data samples to establish a mapping relationship between the various influencing factors and the subjective assessment results. In comparison to the optimal mathematical modeling method, this method can account for various influencing factors that are complex in nature and affect the QoE. This model is found to be highly accurate. In recent years, many works [12,13] have used the QoE prediction model established by machine learning to guide the client's bitrate selection in dynamic adaptive streaming services.

Reference [14] summarizes the methods used for building the QoE assessment models based on machine learning. Reference [15] uses the Hammerstein–Wiener method to develop the model for analyzing a user's QoE at the time of re-buffering. Reference [16] accounts for three factors, namely, the video quality, re-buffering time and memory to devise the QoE prediction model using Support Vector Regression (SVR). The P.NATS [17] uses the random forest approach to set up a model that would enable us to analyze the effects of re-buffering location, re-buffering time, frame rate and video quality on a user's QoE. Reference [18] proposes a multi-layer neural network based on the SGD-BP algorithm to set up the QoE model. Reference [12] uses the deep reinforcement learning method to perform the client's bitrates selection. First, a chunk-wise QoE model is established using three influencing factors such as video bitrate, quality switching and initial re-buffering time. The QoE model is used as a reward function to affect the client's bit rate selection. Then, a linear function is used to reflect the relationship between the influencing factors and the QoE. Finally, SVR is used to determine the coefficients of each of the variables. The experimental results demonstrate that this QoE model is better than the model that has been used in the Pensieve method [19].

In References [20,21], the effects of spatial and temporal characteristics of a video on a user's QoE have been considered. Spatial Information (SI) and Temporal Information (TI) are considered to be major influencing factors while developing the QoE model. The experimental results indicate that the temporal and spatial characteristics of a video help to improve the accuracy of the QoE model.

In spite of successful results, there is scope for improvement in certain areas such as:

1. For the existing methods, SI and TI are usually used to explain the characteristics of the content of a video. However, owing to its complexity, SI and TI are unable to explain the characteristics.

-
2. The volatility of video quality caused due to bitrate switching has not been taken into account sufficiently.

This paper proposes to develop a QoE assessment model based on the Dynamic Adaptive Streaming over HTTP (DASH) standard which shall consider various factors such as characteristics of the video content, video quality, playback fluency and quality volatility, average bitrate, re-buffering time and count, initial buffering time and quality switching amount. The parameters that define these influencing factors are combined together to form a feature parameter vector, and the ridge regression method is adopted to map the relationship between the feature parameter vector and the MOS value. The experimental results on the two public datasets depicts that deep spatial-temporal features can effectively improve the accuracy of the QoE model. Moreover, the proposed QoE assessment model exhibits higher accuracy in comparison to the state-of-the-art QoE models. The main contributions of this paper are as follows:

1. The spatial and the short-term temporal features of a video are extracted using Three-Dimensional Convolutional Neural Networks (3D CNN). The Long Short-Term Memory (LSTM) is then used to extract the long-term temporal features of the video. 3D CNN and LSTM are used together to extract the deep spatial-temporal features to represent the content characteristics of the video.
2. While accounting for the fluctuation in the quality of a video caused due to bitrate switching on the QoE, other factors, such as video content characteristics, video quality and video fluency, are combined together to form the input feature vector. The ridge regression method is adopted to establish a QoE metric that enables to describe the relationship between the input feature vector and value of the MOS.

The structure of the paper is as follows. Section 2 describes the QoE assessment model proposed in this paper. Section 3 describes the experimental results and analysis. Finally, Section 4 draws the conclusion.

2. The Proposed Video QoE Metric for Dynamic Adaptive Streaming Services

Figure 1 represents the framework of the QoE assessment model described in this paper. In the proposed model, firstly, 3D CNN and LSTM techniques are combined to extract deep spatial-temporal features to represent the content characteristics of a video. Next, factors such as video bitrate, video quality fluctuation, re-buffering and initial buffer, which distort the quality of a video, are extracted based on the DASH standard. Finally, the extracted features are combined to generate the input feature parameters vector, and the ridge regression method is used to map a model between the input feature parameters vector and the MOS value for predicting a user's QoE. The detailed process for developing the model is discussed below.

Figure 1. The framework of QoE assessment model. Deep spatial-temporal features and DASH based feature parameters are combined to generate the input feature parameters vector, and the ridge regression method is used to map the QoE model between the input feature parameters vector and MOS value, then the model is used to predict the user's QoE. QoE: Quality of Experience; DASH: Dynamic Adaptive Streaming over HTTP (HyperText Transfer Protocol); MOS: Mean Opinion Score.

2.1. Extraction of Deep Spatial-Temporal Features of Video

Many works have proved that 3D CNN and LSTM exhibit highly accurate results while studying the deep spatial-temporal features of the content of a video. The research results show that the 3D CNN can effectively extract the spatial and short-term temporal features of a video, and a $3 \times 3 \times 3$ convolution kernel can help in achieving optimum performance in all layers. However, the disadvantage of 3D CNN is that it is unable to extract the long-term temporal features of a video. This paper proposes to use 3D CNN to extract the spatial features and short-term temporal features of a video, and LSTM to extract the long-term features of a video to obtain the deep spatial-temporal features of the video.

Figure 2 shows the network structure which combines 3D CNN and LSTM together to extract the deep spatial-temporal features as mentioned above. The 3D CNN comprises of five convolution layers (Conv1–Conv5), five pooling layers (Pool1–Pool5) and two fully connected layers (F1 and F2). Softmax is used as a classifier whose output are the values between 0–1, indicating the probability of each category. The input video frames are normalized to be the size of 112×112 . Usually, low convolutional layer of the 3D CNN can extract low-level visual features such as edges and textures, whereas the high convolutional layer can extract more distinguishing high-level semantic features.

Figure 2. Network Structure combines 3D CNN and LSTM together. 3D CNN aims to extract low-level visual features and LSTM aims to model the long-term temporal dependency of a video.

LSTM aims to model the long-term temporal dependency of a video. As opposed to the 3D CNN, LSTM can simulate the dynamic evolution of the state of the video content through a series of memory units, which can represent the long-term temporal characteristics of a video.

In this paper, the second convolutional layer features of the 3D CNN is extracted and used as the input to LSTM network after being reshaped (denoted as ReshapedFeature1 in Figure 2). The output generated from the first layer of LSTM is extracted and used as the deep spatial-temporal feature of the video.

The 3D CNN network parameters are set as follows: the size of convolution kernel is set as $3 \times 3 \times 3$. The MaxPool3D's kernel size of the first layer is $2 \times 2 \times 1$ and other layer's kernel size is $2 \times 2 \times 2$.

The LSTM network parameters are set as follows: the size of the input tensor is reshaped as [1, 16, 200704] and the dimension of the output of the first layer is 50.

2.2. Study of Factors Influencing QoE Based on the DASH Standard

According to the DASH standard, a video is encoded at different bitrates, and each video bitstream is separated into chunks and a Media Presentation Description (MPD) is generated in binary extensible markup language format. The MPD file describes the information corresponding to a video, URL address and the segment format list of the video chunk, such as the encoding bitrate, resolution and length of a video, and is stored on the server. The client selects the most suitable video bitstream to download on the basis of current network condition, the processing capability of the hardware, and the cache status to improve a user's QoE. DASH-based video transmission may cause two kinds of distortion-video quality switching and stalling, leading to fluctuations in video quality and disfluency of the video, which seriously affect users' QoE.

The following parameters are used for quantifying various influencing factors of QoE based on the DASH standard.

- Video quality

(1) Average bitrate of video. The video encoding bitrate directly affects the reconstructed quality of a video. The higher the video bitrate, the better the quality of the video. In this paper, DASH standard method is used to extract the video bitrate, and the average value of bitrate is calculated to express the average level of video quality. Here, Avg_n represents the average bitrate of the n th chunk. For the $(n + 1)^{th}$ chunk, the average bitrate Avg_{n+1} is calculated using the following formula:

$$Avg_{n+1} = \frac{n * Avg_n + Bitrate_{n+1}}{n + 1} \quad (2)$$

(2) The frequency of each video quality level. This is used to represent the frequency of each video quality level appears in a video, which can be calculated as:

$$P_i = \frac{B_i}{N} \quad (3)$$

where B_i represents the number of i^{th} video quality level in a video, N is the number of segments of a video. Each segment is encoded with a bit rate, corresponding to a quality level.

- Fluency of a video

(1) Video frame rate. This refers to the frames per second and can be used to indicate the fluency of video playback.

(2) Re-buffering times and duration. When the network conditions change drastically, the bandwidth fluctuations, bit errors and other factors often lead to interruption in the process of video playback and re-buffering. This is one of the major factors that heavily affect the QoE. A research conducted by Amazon shows that frequent re-buffering reduces the user's interest in

watching videos by approximately 39% [22]. In this paper, the re-buffering time is counted to exhibit the impact of network conditions on QoE.

While watching a video online, the total duration of the video is divided into the following segments: The original video duration, initial buffering time and re-buffering time. The average re-buffering time $t_{stallingP}$ is denoted as:

$$t_{stallingP} = \frac{t_{stalling}}{t_{original} + t_{stalling} + t_{initial}} \quad (4)$$

where $t_{stalling}$, $t_{original}$ and $t_{initial}$ represent stalling time, original video time and initial buffering time, respectively.

(3) Initial buffering time. This indicates the time taken by a client for requesting a video for playback from the server. Initial buffering time is decided by video chunk size, network status and client attributes, separately. It is one of the most crucial factors that affect the QoE and is regarded as a key factor in this paper.

- Fluctuation of a video quality

(1) Switching times of video bitrate. The DASH standard adaptively selects the video bitrate on the basis of the network conditions and client buffering status to mitigate the impact of dynamic changes in network conditions on QoE. Frequent switching of bitrate will lead to fluctuation in the quality of a video. This paper counts the switching times of video bitrate to characterize the fluctuation of video quality, and reflects the impact of dynamic changes in network conditions on QoE.

(2) The variance in the proportion of each video bitrate. This indicates the overall level of volatility in video quality. The variance V can be defined as:

$$V = \frac{\sum_{i=1}^C (P_i - M)^2}{C} \quad (5)$$

where M represents mean value of P_i , $i = 1, \dots, C$, C is the number of video bitrates.

Based on the DASH standard, seven parameters are extracted and factors influencing QoE are grouped on the basis of video quality, video fluency and video volatility, as listed in Table 1.

Table 1. Parameters representing influencing factors of QoE extracted based on DASH standard.

Classification	Extracted Parameters
Video quality	Average bitrate Proportion of video bitrate Frame rate
Fluency	Average playback interruption length per second Average re-buffering times per second Initial buffering time
Volatility	Average video bitrate switching count per second The variance in the proportion of each video bitrate

2.3. Modeling Method

Due to a correlation between the feature parameters, such as the deep spatial-temporal features of video quality, fluency and volatility as well as due to small data amount, we chose the ridge regression method in this study, to establish the mapping relationship between the input feature parameter vector and MOS.

Ridge regression is a method of regularization of ill-posed problems; it is particularly useful to mitigate the problem of multi-collinearity in linear regression, which commonly occurs in models

with large number of parameters. In general, ridge regression improves the efficiency of parameter estimation, but at the same time it also leads to increased estimation bias.

In essence, ridge regression is an improved least square estimation method. By giving up the unbiasedness of the least square method, the regression coefficient obtained at the cost of losing part of the information and reducing accuracy is more practical and reliable. The fitting of ill conditioned data is better than the least square method.

The loss function of ridge regression can be expressed as:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (y_i - \sum_{j=0}^p \omega_j x_{ij})^2 + \lambda \sum_{j=0}^p \omega_j^2 \quad (6)$$

where y_i represents the score for i^{th} video, x_{ij} represents influencing factor, ω_j is regression coefficients and λ is regularization parameter. Equation (6) shows that the ridge regression expression is obtained by adding a regular term to the least square method. This regular term is the penalty for the coefficient to be obtained. Adding the regular term can avoid the occurrence of overfitting.

3. Experimental Results and Analysis

To evaluate the performance of the proposed model, we conducted various experiments using the Waterloo Streaming QoE Database-III [23] published by the University of Waterloo and the LIVE-NFLX-II [24] dataset published by the University of Texas at Austin. The experimental results are explained below.

3.1. Datasets

The Waterloo Streaming QoE Database-III dataset, published by University of Waterloo in 2018 [23], is one of the largest datasets containing 450 videos. Coding bitrates recommended by Netflix [25] and Apple [26] are used to encode the source video sequences of 20 different types of content in the dataset. The videos are encoded at 235–7000 Kbps, obtaining 11 coding bitrate levels, and then the video bitstreams are stored on the server. The client selects six representative ABR algorithms based on the DASH standard, simulates under 13 representative network conditions, and uses the ITU-R Absolute Category Rating (ACR) scale for subjective testing and scoring.

The LIVE-NFLX-II [24] subjective video QoE dataset is one of the most comprehensive datasets available [27]. The dataset consists of 15 source videos and 420 distorted videos. Seven mobile network traces and four client adaptation algorithms are adopted to generate the simulation data. Furthermore, the dataset includes both continuous and retrospective MOS. Due to using a dynamic optimizer to obtain the encoding bitrate, the uniform video quality cannot be achieved. To facilitate further processing, when the statistical video quality level switch, it is unified with six quality levels.

3.2. Experimental Parameter Settings

In this paper, 80% of the sample data, randomly selected from the dataset, was used as training data, and the remaining 20% was used to test the accuracy of the model.

3.2.1. Parameter Setting for Extraction of Deep Spatial-Temporal Features

While training the 3D CNN, the video frame is normalized to 112×112 pixels and every 16 frames are treated as a basic unit to be input to the 3D CNN. During the training process, the iteration count is set to 2000.

As for the parameters of the LSTM network, the dimension of the output of the first layer, the training batch for the overall network and the iteration count are set to 50, 30 and 500, respectively.

During the test phrase, every 16 frames is fed as an input to the 3D CNN. The size of feature maps at the second convolutional layer is 56×56 and the number of channels is set as 64 in the network.

The features extracted by the 3D CNN are reshaped to [1, 16, 200704], which serves as an input to the trained LSTM network. The features of the first layer of the LSTM network are extracted as the deep spatial-temporal features of a video.

The deep spatial-temporal features are combined with the parameters in Table 1 to form an input feature parameters vector. For Waterloo Streaming QoE Database-III dataset, the influencing factors include: 50-dimensional deep spatial-temporal features; 12-dimensional features represent video quality; four-dimensional features represent video's fluency and two-dimensional features represent video's volatility. Total dimension of the input feature parameters vector is 68.

For LIVE-NFLX-II dataset, the input features parameter vector has a total of 62 dimensions, including 50-dimensional deep spatial-temporal features; seven-dimensional features representing video quality (6 video quality levels); three-dimensional features representing video's fluency (no initial buffering time) and two-dimensional features representing video's volatility.

3.2.2. Mobile Video QoE Assessment Model Partial Parameter Setting

In the experiments, 80% of the sample data are selected randomly for training to obtain the QoE prediction model. The remaining 20% is used for testing. In this paper, Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank Order Correlation Coefficient (SROCC) are used as the metrics to evaluate the accuracy of the obtained QoE model. PLCC is to measure the linear correlation between two variables X and Y, while SROCC to measure the monotonic relationship between two variables, which can be expressed as:

$$PLCC = \frac{\sum_{i=1}^S (y_{pi} - \bar{y}_p)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^S (y_{pi} - \bar{y}_p)^2 (y_i - \bar{y})^2}} \quad (7)$$

$$SROCC = 1 - 6 \frac{\sum_{i=1}^S d_i^2}{S(S^2 - 1)} \quad (8)$$

where y_{pi} represents the predicted score for i th video, $\bar{y}_p$ represents the average of predicted objective scores, y_i is the subjective score for i th video and $\bar{y}$ is the mean value of all subjective scores; d_i represents the difference between ranks of subjective and objective scores for i th video, S is the number of scores.

3.3. Performance Comparisons of Different Modeling Methods

To verify the superiority of the modeling method selected in this paper, we use different methods to establish the mapping model between the feature parameters vector and the MOS value. The specific results are listed in Table 2, and the bold to show the superiority of the modeling method.

As listed in Table 2, compared with Bayesian regression, LASSO, SVR, decision tree, random forest, Xgboost method, LightGBM method and ElasticNet regression, the QoE model established by using the ridge regression method can obtain the excellent performance. This is because the data volume of the existing datasets is still small, and there is a correlation between the feature parameters such as video quality, fluency and volatility, which is more suitable for the ridge regression method. Therefore, the ridge regression method is adopted as the modeling method in this paper.

Table 2. Performance comparison of models established by multiple regression methods.

Method	Waterloo Streaming QoE Database-III		LIVE-NFLX-II	
	PLCC	SROCC	PLCC	SROCC
Bayesian regression	0.9274	0.9054	0.9543	0.9507
LASSO	0.8936	0.8737	0.7969	0.8057
SVR	0.6394	0.5981	0.6830	0.6668
Decision tree	0.8615	0.7858	0.8722	0.8581
Random forest	0.8268	0.8381	0.7659	0.8204
Xgboost method	0.9324	0.9133	0.8590	0.8465
LightGBM regression	0.8574	0.8462	0.8402	0.8462
ElasticNet regression	0.9357	0.9327	0.9484	0.9465
Ridge regression	0.9570	0.9465	0.9756	0.9718

3.4. Influence of Different Parameter Combinations on Model Accuracy

To directly address the role of each influencing factor such as video content, video quality, fluency and volatility on the accuracy of the model, this paper combines different parameters and establishes the QoE models, separately, by the ridge regression method. The specific parameter combination and the corresponding five QoE models established are listed in Tables 3 and 4. The comparison results of different models are listed.

Table 3. Influence of different parameters on model accuracy (Waterloo Streaming QoE Database-III).

Parameters	QoE Models				
	Model 1	Model 2	Model 3	Model 4	Model 5
Video content characteristics	√		√	√	√
Video quality	√	√		√	√
Fluency	√	√	√		√
Volatility	√	√	√	√	
Feature dimension	68	18	56	64	66
PLCC	0.9570	0.8664	0.3085	0.8002	0.8258
SROCC	0.9465	0.8243	0.3613	0.7265	0.7721

Table 4. Influence of different parameters on model accuracy (LIVE-NFLX-II).

Parameters	QoE Models				
	Model 1	Model 2	Model 3	Model 4	Model 5
Video content characteristics	√		√	√	√
Video quality	√	√		√	√
Fluency	√	√	√		√
Volatility	√	√	√	√	
Feature dimension	62	12	55	59	60
PLCC	0.9756	0.7083	0.4749	0.9384	0.9384
SROCC	0.9718	0.7054	0.8496	0.9112	0.9265

As can be seen from Tables 3 and 4, the proposed QoE model (Model 1) can obtain the highest PLCC and SROCC values. For the Waterloo Streaming QoE Database-III dataset, they can reach 0.9570 and 0.9465, respectively, and for the LIVE-NFLX-II dataset, 0.9756 and 0.9718, which can accurately evaluate and predict the user's QoE.

From the comparison results, the conclusions can be drawn that:

1. The parameters of video content characteristics have an extremely important impact on the prediction accuracy of QoE models.

Compared with Model 2, Model 1 takes the video content characteristics into account, and PLCC and SROCC increase by 9.06% and 12.22% on Waterloo Streaming QoE Database-III and by 26.73% and 26.64% on LIVE-NFLX-II dataset.

2. Video quality fluctuation is another essential factor affecting QoE.

Compared with Model 5, which does not consider video quality fluctuation, Model 1 increases PLCC by 13.12% and SROCC by 17.44% on the Waterloo Streaming QoE Database-III, and 3.72% and 4.53%, respectively, on the LIVE-NFLX-II dataset, which indicates that video quality volatility has a crucial influence on the prediction accuracy of the model.

3.5. The Influence of Deep Spatial-Temporal Features on the Accuracy of QoE Model

In addition, to verify the validity of the deep spatial-temporal features extracted in this paper, we compared it with SI and TI (Model 6). The obtained QoE model using SI and TI is labeled as Model 6. Table 5 shows that compared with SI and TI, the deep spatial-temporal features proposed in this paper can increase PLCC and SROCC by 7.51% and 7.66% on Waterloo Streaming QoE Database-III, and 6.03% and 6.26% on LIVE-NFLX-II dataset, which indicates that deep spatial-temporal features can more effectively characterize the spatial and temporal characteristic of a video. It also should be noted that the dimension of deep spatial-temporal features is much higher than SI and TI.

Table 5. The performance comparison using the proposed deep spatial-temporal features and SI and TI.

Parameters	Waterloo Streaming QoE Database-III		LIVE-NFLX-II	
	Model 1	Model 6	Model 1	Model 6
Video content characteristics	√	SI & TI	√	SI & TI
Video quality	√	√	√	√
Fluency	√	√	√	√
Volatility	√	√	√	√
Feature dimension	68	20	62	14
PLCC	0.9570	0.8819	0.9756	0.9153
SROCC	0.9465	0.8699	0.9718	0.9092

In summary, among the four aspects of influencing factors, namely, video content characteristic, video quality, fluency and volatility, video quality has the greatest impact on QoE and video fluency and volatility have an exceedingly important impact on QoE.

3.6. Performance Comparison with the State-of-the-Art Methods

To verify the superiority of the proposed QoE model, we compared it with the existing four QoE assessment models. The four assessment models include:

1. FTW model [5]: Consider re-buffering count and the re-buffering duration as the influencing factors and establish an exponential relationship between the influencing factors and MOS.
2. SQI model [11]: Use a linear combination of multiple exponential models to analyze the relationship between QoE and video compression, initial delay, re-buffering.
3. P.NATS model [17]: The random forest method is used to model the impact of re-buffering position, re-buffering duration, frame rate and video quality on the user's QoE.
4. Liu's model [28]: Take the effect of initial delay, re-buffering and quality fluctuation on QoE into account and establish the exponential and logarithmic models, separately. Then combine these two models as the QoE prediction model.

For a fair comparison, we test the above four models and the models proposed in this paper on the Waterloo Streaming QoE Database-III. The specific comparison results are listed in Table 6. The experimental results of the four models in the table are cited from the literature [23].

Table 6. Performance comparison of QoE models on the Waterloo Streaming QoE dataset.

Model	SROCC
FTW [5]	0.507
SQI [11]	0.7707
P.NATS [17]	0.8454
Liu's [28]	0.8039
Proposed Model	0.9465

Table 6 indicates that, compared with other models, the proposed QoE model can obtain much higher prediction accuracy with a large margin. This is because the proposed method not only considers video quality and fluency in parameter selection but also uses deep spatial-temporal features that can effectively characterize the content characteristics of the video, so it is obviously superior to other existing methods while predicting the user's QoE.

We also compared the proposed method with the latest QoE assessment model NAVE [29] on LIVE-NFLX-II dataset. NAVE is a kind of No-reference Auto-encoder VidEo (NAVE) quality metric, which uses deep Auto Encoder (AE) network to extract the deep features of a video to estimate the overall visual quality. NAVE used DIIIVINE to extract the NSS features of the video frame and then calculated the spatial-temporal indices. The total feature dimension is $90 \times N$ (N is the number of video frames), then the features are input into the automatic encoder for training.

For a fair comparison, the same 10-fold cross-validation experiments as in the NAVE method is conducted in this paper and the results are listed in Table 7. The experimental results show that the proposed model can achieve more accurate prediction because the NAVE method only uses the characteristics of the video, while other influencing factors, such as re-buffering and quality switching are neglected.

Table 7. Performance comparison of QoE models on LIVE-NFLX-II dataset.

Model	PLCC	SROCC
NAVE [29]	0.8225	0.8274
Proposed Model	0.9124	0.9170

3.7. Complexity Analysis of the Proposed Model

The complexity of the proposed model is mainly composed of two parts: deep spatial-temporal feature extraction and regression model establishment, where deep spatial-temporal feature extraction occupies the most parts of the complexity. The number of Float Operations (FLOPs) is usually used to evaluate the computational cost that an algorithm requires. Therefore, in this paper, we use FLOPs to evaluate the time complexity of the proposed model. The total number of parameters and the model size are adopted to the evaluated space complexity of the model. The complexity of the proposed model is shown in the Table 8.

Table 8. The complexity of the proposed model.

Model	Number of Parameters (Millions)	Model Size (MB)	FLOPs (Millions)
3D CNN Network	15.62	62.52	78.08
LSTM Network	40.15	160.62	80.28

It can be seen that the complexity of the model is high, which means the proposed model can achieve a high performance but at the cost of high complexity. Of course, GPU-acceleration can also be used for the proposed model to improve the processing speed of the model.

4. Conclusions

This paper presents a video quality assessment model that fully accounts for the influence of video quality fluctuation caused by video quality switching on QoE during network transmission. Moreover, various factors, such as video quality, fluency and volatility, are combined with deep spatial-temporal features of video, average bitrate, re-buffering duration and count, initial buffering time and video quality switching count to form the feature parameters vector, and use the ridge regression method to establish the mapping relationship between the feature parameters vector and MOS. The experimental results on the public datasets, namely, Waterloo Streaming QoE Database-III and LIVE-NFLX-II show that, compared with the existing video QoE assessment models, the proposed QoE model established in this paper can achieve higher prediction accuracy. The proposed model can be used to guide the client's bitrate selection to improve the user's QoE for DASH standard as well as other dynamic adaptive streaming technologies.

However, compared with the existing methods, to effectively express the content characteristics of the video, the dimension of the features selected in this paper is relatively high, and the established model is relatively complex. In further work, the complexity of the model will be optimized.

Finally, thanks to the support and help of the author of the Waterloo Streaming QoE Database-III Zhengfang Duanmu.

Author Contributions: Conceptualization, L.D. and L.Z.; methodology, L.Z. and L.D.; software, L.D.; validation, L.D. and L.Z.; formal analysis, L.D. and L.Z.; investigation, L.D. and L.Z.; resources, L.Z., J.Z., H.Z., J.L. and X.L.; data curation, L.D.; writing—original draft preparation, L.D.; writing—review and editing, L.Z., L.D. and J.Z.; visualization, L.D.; supervision, J.L. and L.Z.; project administration, L.Z. and J.Z.; funding acquisition, L.Z., J.L., J.Z., X.L. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (No. 61531006, No. 61971016).

Acknowledgments: The authors would like to thank all the colleagues who contributed to this research work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. YouTube. Available online: <https://cn.beyondsummits.com/sites/default/files/downloads/reports/Youtube%E5%B9%B3%E5%8F%B0%E7%A0%94%E7%A9%B6.pdf> (accessed on 11 November 2019).
2. Cisco Visual Network Index Forecasts and Methods, 2016–2021. Available online: https://www.cisco.com/c/m/zh_cn/express/case_center/vertical/wprsi001573.html?oid=wprsi001573&ccid=cc000069&dtid=odicdc000510.de1 (accessed on 11 November 2019).
3. Xiao, A.; Liu, J.; Li, Y.; Song, Q.; Ge, N. Two-stage rate adaptive strategy for improving real-time video QoE in mobile networks. *Communications* **2018**, *10*, 12–24.
4. Qualinet White Paper on Definitions of Quality of Experience, European Network on Quality of Experience in Multimedia Systems and Services. Available online: <https://hal.archives-ouvertes.fr/hal-00977812/document> (accessed on 11 November 2019).
5. Mok, P.K.; Chan, E.W.W.; Chang, R.K.C. Measuring the quality of experience of HTTP video streaming. In Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management, Dublin, Ireland, 23–27 May 2011.
6. Hoßfeld, T.; Schatz, R.; Biersack, E.; Plissonneau, L. Internet video delivery in youtube: From traffic measurements to quality of experience. *Lecture Notes Comput. Sci.* **2013**, *7754*, 264–301.
7. Claeys, M.; Latré, S.; Famaey, J.; Wu, T.; Van Leekwijck, W.; De Turck, F. Design and optimization of a (fa) q-learning-based http adaptive streaming client. *Connect. Sci.* **2014**, *26*, 25–43. [[CrossRef](#)]

8. Seufert, M.; Wehner, N.; Casas, P. Studying the Impact of HAS QoE Factors on the Standardized QoE Model P.1203. In Proceedings of the International Conference on Distributed Computing Systems, Vienna, Austria, 2–5 July 2018.
9. Watanabe, K.; Okamoto, J.; Kurita, T. Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes. In Proceedings of the Image Quality and System Performance IV, San Jose, CA, USA, 30 January–1 February 2007.
10. Demóstenes, Z.R.; Zhou, W.; Renata, L.R.; Graça, B. The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over. *HTTP EURASIP J. Wirel. Commun. Netw.* **2014**, *216*, 1–15.
11. Duanmu, Z.F.; Zeng, K.; Ma, K.; Rehman, A.; Wang, Z. A Quality-of-Experience Index for Streaming Video. *IEEE J. Sel. Top. Signal Process.* **2016**, *11*, 154–166. [[CrossRef](#)]
12. Liu, J.; Tao, X.; Lu, J. QoE-oriented rate adaptation for DASH with enhanced deep Q-Learning. *IEEE Access* **2018**, *7*, 8454–8469. [[CrossRef](#)]
13. Huang, T.; Zhou, C.; Zhang, R.X. Comyco: Quality-aware adaptive video streaming via imitation learning. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019.
14. Aroussi, S.; Mellouk, A. Survey on machine learning-based QoE-QoS correlation models. In Proceedings of the 2014 International Conference on Computing, Management and Telecommunications, Da Nang, Vietnam, 27–29 April 2014.
15. Ghadiyaram, D.; Pan, J.; Bovik, A.C. A time-varying subjective quality model for mobile streaming videos with rebuffer events. In Proceedings of the Applications of Digital Image Processing XXXVIII, International Society for Optics and Photonics, San Diego, CA, USA, 11–15 August 2015.
16. Bampis, C.G.; Bovik, A.C. Feature-based prediction of streaming video QoE: Distortions, re-buffer and memory. *Signal Process. Image Commun.* **2018**, *68*, 218–228. [[CrossRef](#)]
17. Robitza, W.; Garcia, M.N.; Raake, A. A modular HTTP adaptive streaming QoE model-Candidate for ITU-T P1203 ('P. NATS'). In Proceedings of the 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX2017), Erfurt, Germany, 31 May–2 June 2017.
18. Lv, C.; Huang, R.; Zhuang, W. QoE prediction on imbalanced IPTV data based on multi-layer neural network. In Proceedings of the Wireless Communications and Mobile Computing Conference, Valencia, Spain, 26–30 June 2017.
19. Mao, H.Z.; Netravali, R.; Alizadeh, M. Neural adaptive video streaming with pensive. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGMM2017), Los Angeles, CA, USA, 21–25 August 2017.
20. Menkovski, V.; Member, S.; Liotta, A. Intelligent control for adaptive video streaming. In Proceedings of the IEEE International Conference on Consumer Electronics, Las Vegas, NV, USA, 11–14 January 2013.
21. ITU-T Recommendation, P.910; Subjective Video Quality Assessment Methods for Multimedia Applications; International Telecommunication Union: Geneva, Switzerland, 2008.
22. Amazon's Study. Available online: <https://mux.com/blog/buffering-reduces-video-watch-time-by-40-according-to-research/> (accessed on 11 November 2019).
23. Duanmu, Z.F.; Rehman, A.; Wang, Z. A Quality-of-Experience Database for Adaptive Video Streaming. *IEEE Trans. Broadcast.* **2018**, *2*, 474–487. [[CrossRef](#)]
24. Bampis, C.G.; Li, Z.; Katsavounidis, I. Towards perceptually optimized end-to-end adaptive video streaming. *arXiv* **2018**, arXiv:1808.03898.
25. Per-Title Encode Optimization; Netflix Inc.: Scotts Valley, CA, USA, 2015; Available online: <http://techblog.netflix.com/2015/12/per-titleencode-optimization.html> (accessed on 11 November 2019).
26. Best Practices for Creating and Deploying HTTP Live Streaming Media for Apple Devices. Available online: https://developer.apple.com/library/content/technotes/tn2224/_index.html (accessed on 11 November 2019).
27. Barman, N.; Martini, M.G. QoE Modeling for HTTP Adaptive Video Streaming—A Survey and Open Challenges. *IEEE Access* **2019**, *7*, 30831–30859. [[CrossRef](#)]

28. Liu, Y.; Dey, S.; Ulupinar, F.; Luby, M.; Mao, Y. Deriving and validating user experience model for DASH video streaming. *IEEE Trans. Broadcast.* **2015**, *61*, 651–665. [[CrossRef](#)]
29. Martinez, H.B.; Farias, M.C.; Hines, A. A No-Reference Auto-encoder Video Quality Metric. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP2019), Taipei, China, 22–25 September 2019.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).