Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Assessing the quality of experience in wireless networks for multimedia applications: A comprehensive analysis utilizing deep learning-based techniques

Xiaoliang Zhang *, Li Li

Information Engineering School Jiaozuo Normal College, Jiaozuo, 454000, China

ARTICLE INFO

Keywords:
Deep learning
Recurrent neural networks
QoE
Mesh networks
Multimedia

ABSTRACT

In the context of the burgeoning progression of wireless network technology and the corresponding escalation in the demand for mobile Internet-based multimedia transmission services, the task of preserving and augmenting user satisfaction has emerged as an imperative concern. This necessitates a sophisticated and accurate evaluation of multimedia service quality within the sphere of wireless networks. To systematically address the nuanced issue of user experience quality, the present study introduces a novel method for evaluating multimedia Quality of Experience (QoE) in wireless networks, employing an advanced deep learning model as the underlying analytical framework. Initially, the research undertakes the task of modeling the video session process, giving due consideration to the status of each temporal interval within the session's architecture. Subsequently, the challenge of QoE prediction is dissected and investigated through the lens of recurrent neural networks (RNNs), culminating in the proposition of an all-encompassing QoE prediction model that harmoniously integrates video information, Quality of Service (QoS) data, user behavior analytics, and facial expression analysis. The empirical segment of this research serves to validate the efficacy of the suggested video QoE evaluation method, engaging both quantitative and qualitative comparison metrics with contemporaneous state-of-the-art QoE models, employing the RTVCQoE dataset as the empirical foundation. The experimental findings illuminate that the QoE model elucidated in this study transcends competing models in performance metrics such as PLCC, SRCC, and KRCC. Consequently, this investigation stands as a seminal contribution to academic literature, furnishing an exacting and dependable QoE evaluation methodology. Such a contribution augments the user experience landscape in multimedia services within wireless networks, and instigates further scholarly exploration and technological innovation in the mobile Internet domain.

1. Introduction

Multimedia services over wireless networks are experiencing rapid growth in the era of mobile Internet [1]. These services encompass various forms of multimedia content, such as video streaming, online gaming, and real-time communication applications [2]. For video providers, it is crucial to attract and retain subscribers in order to improve profitability [3]. Achieving this goal relies on delivering an excellent quality video experience that not only includes the delivery of video content but also meets user needs [4].

* Corresponding author.
E-mail address: zxl970427@jzsz.edu.cn (X. Zhang).

https://doi.org/10.1016/j.heliyon.2024.e30351
Received 27 December 2023; Received in revised form 18 April 2024; Accepted 24 April 2024
Available online 25 April 2024

Currently, video experience quality is predominantly evaluated through two methods: subjective rating and objective rating. Subjective rating relies on users' personal evaluations, providing a direct reflection of their feelings and thus yielding more accurate results [5]. However, subjective scoring suffers from poor real-time performance and high survey costs. On the other hand, objective ratings focus on objective factors within a video session that often reflect the user's preference, such as viewing duration and user behaviors during playback. In recent years, objective scoring has become a prominent area of research for video Quality of Experience (QoE) in wireless networks [6]. In contrast to existing methods that heavily rely on subjective opinion data, our approach places a distinct emphasis on user behavior data. While subjective rating methods offer valuable insights, they suffer from poor real-time performance and high survey costs. In our revised methodology, we enhance the reliance on objective rating methods, particularly focusing on user behaviors during video viewing. By prioritizing user behavior data, our model gains a more immediate and cost-effective understanding of user experience, aligning with the contemporary need for timely and resource-efficient evaluation methods in the dynamic landscape of multimedia services over wireless networks.

Traditional QoE prediction models primarily focus on collecting Quality of Service (QoS) parameters, analyzing the relationship between QoE and QoS, and establishing mappings between them using various methods. Extracted QoS measures currently include throughput, starting buffer and delay, signal strength, network speed, and more [7–9]. Additionally, QoE has a certain impact on user behaviors, establishing a mutual influence relationship between the two. Moreover, the state of each time period within a video session cycle may also impact the quality of the video session and subsequent time periods.

In summary, developing new and more effective QoE prediction models is the primary focus of current research on video quality of experience in wireless networks. Under the existing wireless network conditions, breaking through the limitations of current modeling methods by jointly considering factors that affect video QoE and using the model for QoE prediction to offer guidance for video providers presents a challenging and prevalent problem in both industry and academia.

2. Related works

In wireless network multimedia services, the factors influencing Quality of Experience (QoE) can be categorized as subjective and objective. Subjective factors encompass the user's life experience, emotions, and background, while objective factors include video coding and decoding methods, bandwidth, latency, and more [10]. Currently, when quantitatively assessing wireless network multimedia QoE, the Mean Opinion Score (MOS) average evaluation method is often employed, classifying it into five levels: excellent, good, average, poor, and very poor [11,12]. Subjective evaluations of wireless network multimedia services require evaluators to provide subjective scores based on their own service experiences, with the MOS analysis method being a commonly used scoring method. As direct recipients of wireless network multimedia services, evaluations based on the subjective thoughts of users hold significant authority and accuracy, providing genuine and effective feedback. To this end, the telecom standardization department has established specifications for subjective evaluations of wireless network multimedia services, clearly outlining the corresponding evaluation methods and scoring standards.

Recent studies have indicated that Quality of Service (QoS) has an impact on user behaviors during video viewing, and there exists a correlation between these behaviors and QoE [13,14]. By utilizing user behaviors during video viewing, the accuracy of QoE prediction can be improved. By collecting user behaviors and QoS parameters, a QoE model can be constructed and the relationship between QoS, user behaviors, and QoE can be validated [15,16]. With the increasing volume of video content, user personalization preferences have also become factors influencing QoE. This approach argues that QoE is influenced by technical-level QoS parameters, while user factors also need to be considered. User preferences or the state of a specific time period may have a greater impact on QoE than the technical level [17]. However, due to the diversity of users and the complexity of the technical level, obtaining the characteristics and data of user factors can be relatively challenging. Simultaneously, to ensure the revenue of video providers, the prediction of QoE is gradually shifting from separate subjective ratings to objective ratings, considering factors such as user engagement, viewing time, and forthcoming user behaviors.

After determining the parameters that affect QoE, the method has been enhanced by proposing a scheme that combines support vector machine and BP neural network algorithms for QoE prediction [18,19]. Additionally, some studies have employed linear regression methods to predict QoE values. This method offers interpretability and allows for the observation of the weights assigned to each parameter, making the model more practical. Based on these weights, video providers can enhance specific parameters to ensure the quality of user experience. In recent years, with the advancements in deep learning, various deep learning models have been applied to QoE prediction. For example, one article utilized deep belief networks to implement the mapping from QoS to QoE [20,21]. The method preprocesses the values and utilizes deep belief networks to fit MOS values. Experimental results demonstrate significant improvements in training efficiency and model convergence compared to traditional BP neural networks, achieving better streaming QoE results.

QoE modeling schemes based on numerical analysis, traditional machine learning, and deep learning can effectively map QoE impact parameters to QoE values and be employed for QoE prediction. However, current research faces certain issues: most of the collected data is based on the collection of QoE impact parameters at a single moment, disregarding the influence of different stages and states of video users on overall QoE throughout the entire viewing process. Existing QoE prediction models primarily focus on subjective MOS values for prediction, while objective QoE metrics (such as user engagement, viewing time, upcoming user behaviors, etc.) that are highly sought after by video providers have not been adequately explored. This limitation makes it challenging to provide video providers with a comprehensive reference solution. In future research, it is crucial to further explore methods that involve multi-moment QoE data collection combined with objective metrics. This approach will provide a more comprehensive and accurate QoE prediction model and serve as a more effective decision-making basis for video providers [22].

In conclusion, the field of QoE prediction in wireless network multimedia services requires advancements in considering both subjective and objective factors, exploring new data collection methods, and leveraging machine learning techniques. By addressing these challenges, researchers can provide valuable insights and tools to enhance the overall quality of user experience in multimedia services over wireless networks.

3. Methodology

3.1. Video session process modeling

Fig. 1 illustrates the complete video session process, where each fixed time step corresponds to the states within that time period, which can be represented by QoE data such as video buffering and user behaviors.

Throughout the video session, we meticulously capture sample data at regular intervals to ensure that we obtain a comprehensive and accurate overview of the changing states. This includes data such as QoE metrics, video buffering patterns, and user behavior patterns, all of which play a crucial role in shaping the user's overall experience. By doing so, we are able to identify any inconsistencies or areas for improvement in the video streaming process, as well as gain valuable insights into user preferences and habits. We implement a strategy to evenly sample label data at multiple time points throughout the video session. This approach will enable us to gain a deeper understanding of user behaviors and preferences at different moments, and thus, offer more tailored and engaging experiences. By evenly distributing the sampling of label data across various time points, we will minimize the risk of bias and ensure that our dataset is more representative of the entire user journey.

If there are more than N data records, N records are evenly selected as sample data from all records; if there are less than N records, the records are filled up to N records. At the end of the video session, we capture the label data to correspond it to the sample data. Thus, the N status records generated by each video session can be defined as one sample data, and the label data at the end of the video session is used as the label data of this sample data.

3.2. Analysis of QoE prediction problems based on RNNs

3.2.1. Overview of RNNs in QoE prediction

At a high level, our QoE prediction model employs RNNs to capture the dynamic nature of QoE data over time. Let's illustrate this with an example: consider a video session with fixed time steps, where each step represents states like video buffering and user behaviors. The RNNs process this sequential data, incorporating information from previous time steps for better predictions.

Fig. 2 illustrates the complete video session process, where each fixed time corresponds to a state within that particular time period. These states can be represented by QoE data, including video buffering, user behaviors, and more. The sample data collected by the QoE Player client at time step k ($k < T$) are correlated with the data prior to time step k, collectively influencing the user's quality of experience for the video session.

3.2.2. Layers and processing steps

RNNs possess a signal feedback structure, enabling them to retain the memory of the sample data prior to time step k through the dynamic properties of RNNs [23]. To better address the temporal dynamics of the entire viewing process, we will enhance our model architecture. While RNNs provide a valuable signal feedback structure, we will introduce additional layers or mechanisms to capture longer-term dependencies and variations in user experience. This involves incorporating attention mechanisms or extending the memory capabilities of our model. These modifications will enable our model to better reflect the evolving states of the video session

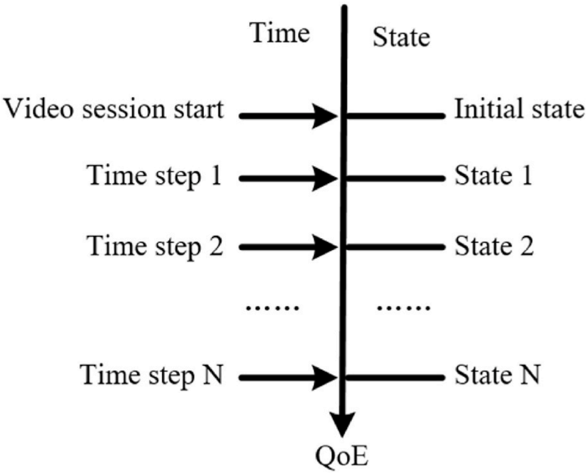

Fig. 1. Video session process.

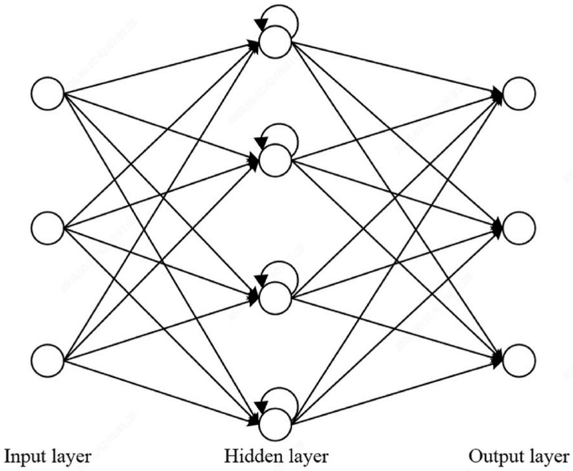

Fig. 2. Schematic diagram of RNN structure.

and enhance its predictive capabilities across diverse temporal intervals.

The input sequence goes through layers represented as: input layer (layer x), hidden layer (layer h), and output layer (layer o). The dynamic properties of RNNs, especially the nonlinear computations in the hidden layer, allow the model to retain memory and integrate historical information for improved future predictions. The typical RNNs structure is depicted in Fig. 2, where layer x denotes the input layer, layer h represents the hidden layer, and layer o signifies the output layer. Layer x receives input vector data, processes them using weights and biases, and transmits them to layer h. Finally, layer o outputs the predicted values, which constitute a set of data for future moments. The core of RNNs lie in the nonlinear computations performed by layer h, providing enhanced expressiveness to the overall model and integrating historical information from multiple moments to achieve better future predictions. When employing RNNs for mapping video information, QoS data, and user behaviors to QoE, the following problem can be defined: for a given video session, spanning from the start of video playback to the end of time step T, the QoE training sequence x(t) is obtained by collecting data at each fixed time step. After undergoing nonlinear operations in the h layer, the output vector sequence of the o layer is utilized to derive the prediction value vector sequence of QoE.

3.2.3. *Model modifications for QoE prediction*

Let's delve into the customizations for QoE prediction. For a given video session, we define a set of N status records as one sample data. Equations (1)–(4) detail the generation of sample data and the computations within the RNNs layers.

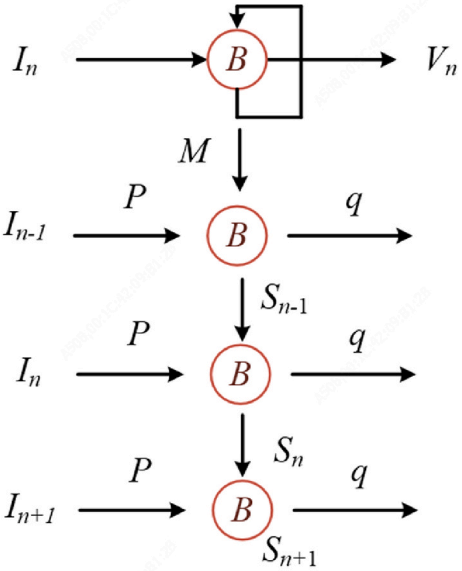

Fig. 3. RNN internal structure.

The sample data generated for each video session is defined as shown in Equation (1).

$$i = (i_1, i_2, \ldots, i_N) \tag{1}$$

Where $i_1, i_2, \ldots, i_N$ denote time steps 1, 2, …,N. The input first needs to be fed to the b-layer for operation to obtain the output of the b -layer. The output of the b -layer is defined as shown in Equation (2).

$$b^t = \left(b^t_1, b^t_2, \ldots, b^t_N\right) \tag{2}$$

Where b^t_N is jointly determined by the sample data input $i(n)$ (t) at time step n and the b-layer output $b^t(n-1)$ at one previous time step. The b-layer generates the output by nonlinear computation.

The o-layer is the output layer of the whole QoE prediction model. Considering the characteristics of the QoE prediction problem, there is only one node in the o layer, and the output of this node is the prediction value of QoE. Therefore, this paper adopts the "many-to-one" model based on RNNs for the interpretation and analysis of QoE prediction. The structure of RNN is shown in Fig. 3.

The output of each node in the b-layer is defined as shown in Equation (3), where $f(.)$ denotes the activation function of the b-layer, S_n denotes the state at moment n. P denotes the weight from the i-layer to the b-layer, and M denotes the weight from the input layer to the output layer.

$$S_n = f(Pi_n + MS_{n-1}) \tag{3}$$

The o layer retains only one node, so its output can be defined as

$$o_N = M_N f(Pi_N + MS_{N-1}) \tag{4}$$

3.3. QoE prediction model based on video information, QoS data, user behaviors and facial expressions

This section focuses on constructing the VSBC-E (Video information, QoS, Interbehavior, and Countenance-based QoE) model. We will provide a detailed description of the VSBC-E model.

Facial expression data of users is stored as images, and in the realm of deep learning, facial expression images are typically processed as multidimensional arrays. Convolutional neural networks (CNNs) have demonstrated significant advancements in tasks such as facial expression recognition. In this paper, for facial expression image data, convolutional operations are utilized to extract relevant features. By combining CNNs with gated recurrent units (GRU), the VSBC-E model is constructed as a QoE prediction model based on video information, QoS data, user behaviors, and facial expressions. The overall architecture of the VSBC-E model is depicted in Fig. 4.

The input of VSBC-E model includes two types of data. The first category is the one-dimensional data such as video information, QoS data and user behaviors, as shown in the figure $i_1, i_2, \ldots, i_N$ are shown. The other category is multidimensional data such as user facial expressions, as shown in *img1, img2, …, imgN*. These two types of input data are each subjected to feature extraction and learning through their own network structures, and then the two network branches are merged.

By incorporating user behaviors and facial expressions into our model, we aim to provide a more holistic and nuanced

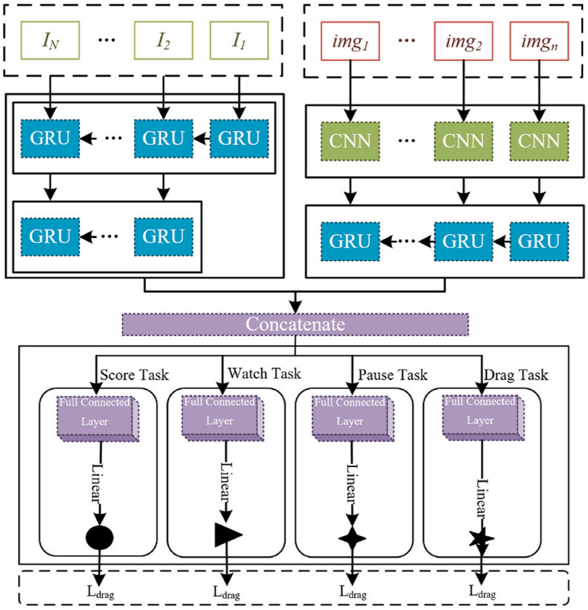

Fig. 4. Framework of QoE prediction model in this paper.

understanding of the factors influencing QoE during video streaming sessions. These additions contribute to the model's ability to capture user behaviors, ultimately improving the accuracy of QoE predictions.

3.3.1. Model input

The VSBC-E model is used to predict four QoE metrics, whose inputs include video information of each video session, QoS data, user behaviors and user facial expressions. According to the dimensionality of the data, the sample data can be divided into two categories: one-dimensional data and three-dimensional data. During each video session, these two types of sample data collected by QoePlayer correspond to four QoE values. In the training process, N one-dimensional sample data and N three-dimensional user expression data are input at a time. For the ith sample data, the number of progress bars dragged and the number of pauses can take values from 0 to any number, the user involvement can take values between 0 and 1, and the QoE subjective rating can take values between 0 and 10.

3.3.2. CNN-based user expression image feature extraction

In this paper, we mainly borrow the idea of VGG16, and for a single user facial expression image, the CNN performs feature extraction as shown in Fig. 5. Each image is processed by 5 convolutional blocks, and finally Flatten layer is added for flattening. The number of convolutional kernels from Convblock1 to Convblock5 are 32, 64, 128, 256, 512, respectively. the kernel size of the first convolutional operation is 7. After the processed user facial expression images undergoing a series of nonlinear transformations such as convolution and maximum pooling, the dimensionality of the feature map is finally converted to one dimension by the Flatten layer. At this point, the feature data extracted from the images can be merged with the first class of one-dimensional data after feature extraction and learning by GRU.

3.3.3. Model branch merging and multi-task result prediction

The VSBC-E model takes three types of 1D data—video information, QoS data, and user behaviors—as input, along with the user's facial expression picture. For the 1D data input, the VSBC-E model incorporates two GRU layers after the 1D data input layer to extract feature information from the 1D sample data. The features from the two branches are then merged to obtain a longer feature vector.

Following the merging of the branches, considering the complexity of the user expression features, the model is connected to a fully connected layer (FC layer) with 16 neurons to further extract features from the sample data. Subsequently, four tasks—drag count, pause count, user engagement, and subjective QoE score—are added as QoE predictors. All four tasks are directly connected to the upper layer with a linear activation function, outputting the prediction results for their respective tasks. This process defines the learned features by the VSBC-E model as:

$$\widehat{j}_x = f_2(i_x, img_x) \tag{5}$$

Where, the function f_2 represents the nonlinear mapping of the input video information, QoS data, user behaviors, and user facial expressions sample data through a multilayer network structure for multitask QoE prediction. The symbol x represents the one-dimensional sample data comprising video information, QoS data, and user behaviors, while the user facial expression data is denoted separately.

The prediction results of the VSBC-E model consist of dragging progress bar count, pause count, user engagement, and subjective QoE score. Among them, $\widehat{j}_x^{drag}$ represents the linear prediction result of the dragging progress bar count task, $\widehat{j}_x^{pause}$ represents the linear prediction result of the pause count task, $\widehat{j}_x^{watch}$ represents the linear prediction result of the user engagement task, and ,$\widehat{j}_x^{score}$ rep-

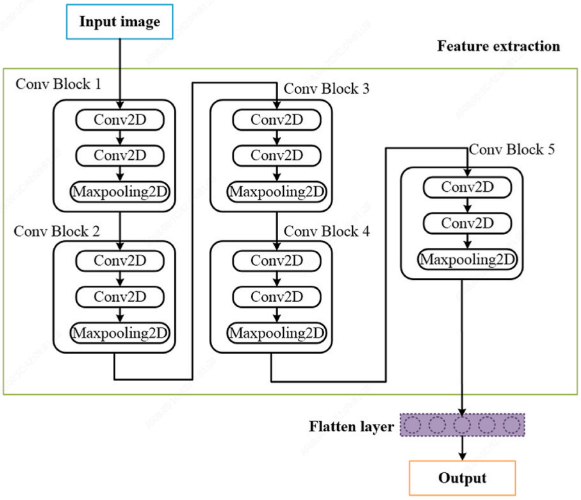

Fig. 5. Feature extraction process by CNN.

resents the linear prediction result of the subjective QoE score task. The VSBC-E model for QoE prediction primarily relies on GRU to handle sequential data in the samples and CNN to extract features from user expression pictures.

In our proposed methodology, the choice of the number of time steps is a critical aspect that requires a careful balancing act. This decision influences the granularity of our sample data, impacting both computational efficiency and the precision of our predictions. "

Increasing the number of time steps allows for a finer granularity in modeling user interactions and system responses. This heightened granularity is essential for a more nuanced understanding of how QoE-related factors evolve during a video session. It enables us to dissect the user experience into smaller, more manageable units, providing a detailed perspective on the interplay between various factors. Reshaping the sample data with multiple time steps contributes to the improved predictive capabilities of our model. The inclusion of a richer temporal context allows the model, particularly based on RNNs, to learn and leverage dependencies and patterns over time. This, in turn, enhances the model's ability to make more accurate predictions regarding future states and user experiences.

3.3.4. Real-time and cost-effectiveness discussion

Our approach stands out for its real-time-friendly nature, prioritizing the utilization of user behavior data, providing more timely insights compared to traditional subjective opinion data. Moreover, our method offers an economically viable alternative, as the reliance on user behavior data significantly reduces survey costs, making it a cost-effective solution for evaluating video QoE in wireless networks.

4. Experiment and analysis

In this paper, experiments were conducted using the constructed RTVCQoE dataset, which is divided into two parts based on application scenarios: the camera scenario and the screen sharing scenario. The training set and test set were divided in an 8:2 ratio. The effectiveness of the proposed video QoE evaluation method, which is based on modeling the nearest neighbor frame relationship, was verified through quantitative and qualitative evaluations of the algorithm's quality regression performance.

4.1. Dataset details

4.1.1. Data collection process

The data were collected using a systematic approach that involved capturing video sessions under different network conditions and user behaviors. We considered factors such as video buffering, user engagement, and other relevant QoE indicators during the data collection process.

4.1.2. Labeling process

To ensure the dataset's accuracy, labels were obtained through a combination of subjective evaluations and automated methods. Subjective evaluations were collected from users who provided feedback on their QoE during the video sessions. Automated methods involved leveraging algorithms to analyze specific QoE-related metrics.

4.1.3. Example of dataset usage

As an illustrative example, consider the camera scenario in our dataset. Video sessions were recorded using a diverse set of cameras, simulating real-world scenarios. Users interacted with the video content, and their behaviors were tracked. Subjective evaluations were obtained through user feedback surveys, and objective metrics were automatically recorded.

4.1.4. Limitations and mitigation strategies

Obtaining a sufficient quantity of diverse and representative data to effectively train our model is a significant challenge. Factors such as user preferences, network conditions, and content types contribute to the complexity of acquiring a comprehensive dataset. This limitation arises from the intrinsic diversity of user behavior and the dynamic nature of video content consumption.

To address the data collection challenge, we have implemented several mitigation strategies. These include leveraging synthetic data augmentation techniques to supplement our dataset, collaborating with diverse user groups to ensure a more comprehensive representation of preferences, and continuously updating our dataset to adapt to evolving video consumption patterns.

4.2. Experimental environment

The experiments were conducted on a hardware configuration consisting of an Intel Core i9-9900K CPU @ 3.60GHz × 16 platform and 2 NVIDIA TITAN Xp GPUs. The operating system used was Ubuntu 18.04, and the programming language employed was Python, along with the PyTorch framework, which leverages the advantages of its dynamic graph design.

4.3. Rating criteria

Following the recommendations of the Video QoE Expert Group (VQEG), three criteria were utilized in this paper to quantify the prediction accuracy and evaluate the prediction monotonicity of the model. The Pearson linear correlation coefficient (PLCC) was employed to assess prediction accuracy, while the Spearman ranking correlation coefficient (SRCC) and Kendall ranking correlation coefficient (KRCC) were used to evaluate prediction monotonicity. A superior objective QoE model should yield higher PLCC, SRCC, and KRCC values.

To assess the performance of our QoE prediction model, we rely on correlation metrics, including the Pearson linear correlation coefficient (PLCC), Spearman ranking correlation coefficient (SRCC), and Kendall ranking correlation coefficient (KRCC). These metrics are commonly used in video quality prediction studies to measure the correlation between predicted values and observed user experiences.

Our subjective opinion reports collected from users during video sessions serve as the de facto "ground truth" in our evaluation. These reports encompass diverse aspects of user perception, such as video quality and engagement. The choice of correlation metrics is deliberate, as they provide an objective and quantitative measure of how well our predictions align with users' actual experiences. This approach is consistent with established practices in the field of video quality prediction, where correlation metrics are widely accepted as reliable indicators of model performance.

Our evaluation is based on subjective opinion reports provided by users during video sessions. Although we do not explicitly mention "ground truth" values in the manuscript, these subjective reports are considered our reference standard for evaluation, representing the genuine feelings of users during their viewing experiences. Correlation metrics are employed to measure the consistency and correlation between model predictions and these subjective reports.

The calculation formula for PLCC is as follows:

$$PLCC = \frac{\sum_{x=1}^{T} (I_x - \overline{I})(J_x - \overline{J})}{\sqrt{\sum_{x=1}^{T} (I_x - \overline{I})^2}\sqrt{\sum_{x=1}^{T} (J_x - \overline{J})^2}} \tag{6}$$

Where, I_x represents the subjective evaluation value of the ith video, and J_x denotes the objective prediction value of the model. $\overline{I}$ and $\overline{J}$ represent the mean values of the subjective evaluation value and the objective prediction value of the model, respectively. T denotes the number of videos being tested. PLCC evaluates the prediction accuracy and takes values within the range of [0, 1], with values closer to 1 indicating higher accuracy of the model prediction and vice versa. The lower the accuracy, the further the value deviates from 1.

The SRCC calculation formula is as follows:

$$SRCC = 1 - \frac{6\sum_{x=1}^{T} d_x^2}{T\left(T^2 - 1\right)} \tag{7}$$

Where T denotes the number of videos to be tested, and di denotes the difference between the subjective evaluation value of the x-th distorted video and the objective prediction value. The monotonicity of the SRCC evaluation model prediction takes values between [-1,1], and the closer the absolute value to 1 indicates better monotonicity.

The KRCC calculation formula is as follows:

$$KRCC = \frac{U_c - U_d}{\frac{1}{2}T(T-1)} \tag{8}$$

Where, U_c denotes the logarithm of the subjective evaluation value that agrees with the objective evaluation value of the model, U_d denotes the logarithm of the subjective evaluation value that does not agree with the objective evaluation value of the model, and T denotes the number of videos to be tested. The value of KRCC ranges from [-1,1], and the closer the absolute value to 1 indicates better monotonicity.

4.4. Comparative experiments and analysis

To ensure a fair comparison, the dataset was randomly divided into training data (80 %) and test data (20 %) in a proportional manner, ensuring that the two sets do not overlap in content. Following the aforementioned principles, 10 replicate experiments were conducted on the comparison dataset, and the median of the three performance metrics was recorded for all comparison models.

Table 1, Table 2, and Table 3 present the PLCC, SRCC, and KRCC values, respectively, for all comparison methods on the RTVCQoE

Table 1
PLCC values of different QoE models.

QoE model	Camera	Screen Sharing	Average
VIDEO [24]	0.522	0.487	0.505
V-BLIINDS [25]	0.639	0.535	0.587
VSFA [26]	0.776	0.693	0.734
RIRNet [27]	0.843	0.819	0.831
Ours	0.954	0.935	0.945

dataset. For the purpose of comparison, we selected widely used and state-of-the-art QoE models, namely VIIDEO [24], V-BLIINDS [25], VSFA [26] and RIRNet [27].

It can be observed that the proposed method in this paper exhibits the best performance on the RTVCQoE dataset. In the camera scenario, the PLCC, SRCC, and KRCC metrics are 0.954, 0.947, and 0.91, respectively. Compared to 0.843, 0.828, and 0.814 for RIRNet, the improvement is 13.3 %, 14.5 %, and 11.9 %. In addition, the PLCC, SRCC and KRCC metrics of QoE in the screen sharing scenario are 0.935, 0.923 and 0.885, respectively, compared to 0.819, 0.81 and 0.797 of RIRNet, an improvement of 14.3 %, 14.1 % and 11.2 %.

4.5. Algorithm validation

To further validate the effectiveness of the proposed QoE method, this paper presents the confusion matrix for prediction using the second-best indicator. Fig. 6 illustrates the confusion matrix, where the darkness of the entries indicates the strength of correlation between the row and column values. Higher values (darker colors) suggest a stronger correlation. The numbers on the axes represent the range of QoE scores, with a value of 1 indicating that the interval from 1 to 2 is covered.

The analysis of the confusion matrix reveals that the algorithm introduced in this paper demonstrates fewer outliers, which further confirms the more stable performance of the QoE method proposed. This observation reinforces the validity and reliability of the proposed QoE method in achieving consistent and accurate predictions.

4.6. Computational efficiency comparison

Finally, to verify the computational efficiency of proposed QoE model, comparative experiments are conducted on the RTVCQoE dataset in this paper, and the results are shown in Table 4.

The experiment utilized the test time of the RIRNet model as the benchmark value and calculated the ratio of the test time of the other comparison models to the benchmark value. A smaller ratio indicates a lower computation time overhead for the model. As shown in Table 4, the RIRNet model had the longest test time, while the VSFA model, which also employed RNNs, had a test time close to 0.941 times that of RIRNet. In contrast, the test times of V-BLIINDS and VIIDEO, which used traditional methods, were significantly less, at 0.327 and 0.29 times that of RIRNet, respectively. The test time of the proposed QoE in this study was 0.365 times the benchmark value, slightly higher than the traditional methods V-BLIINDS and VIIDEO, but much lower than the deep learning-based RIRNet and VSFA. The experimental results demonstrate that the method proposed in this study effectively improves the computational efficiency of the model.

4.7. In-depth analysis of model performance

(1) Linear Relationship Analysis (PLCC):

The PLCC value of the model reflects the linear correlation between the predicted values and the actual observations. By comparing the model's predicted values with the actual observations, we can detect any potential biases or fitting issues. If a strong linear trend is observed, it indicates that the model may be overfitting the data during the prediction process, resulting in an overly simplistic relationship between predicted values and actual observations. Conversely, a weak or nonexistent linear relationship suggests bias in the model, indicating a failure to capture linear patterns.

(2) Rank Relationship Analysis (SRCC):

The SRCC metric examines the model's ability to capture rank relationships within the data. We will investigate the sensitivity of the model to variations in data ranks, especially under different data distributions. By comparing the relationship between actual ranks and predicted ranks, we can assess the model's performance in capturing rank relationships. Consistent rank relationships indicate the model's robust ability to capture the ordering of data across various scenarios.

(3) Sensitivity to Rank Changes (KRCC):

The KRCC metric further explores the model's sensitivity to changes in data ranks. We will analyze the model's performance when

Table 2
SRCC values of different QoE models.

QoE model	Camera	Screen Sharing	Average
VIDEO [24]	0.504	0.467	0.486
V-BLIINDS [25]	0.63	0.515	0.573
VSFA [26]	0.751	0.683	0.717
RIRNet [27]	0.828	0.81	0.819
Ours	0.947	0.923	0.935

Table 3
KRCC values of different QoE models.

QoE model	Camera	Screen Sharing	Average
VIDEO [24]	0.49	0.447	0.469
V-BLIINDS [25]	0.612	0.5	0.556
VSFA [26]	0.739	0.662	0.701
RIRNet [27]	0.814	0.797	0.806
Ours	0.91	0.885	0.898

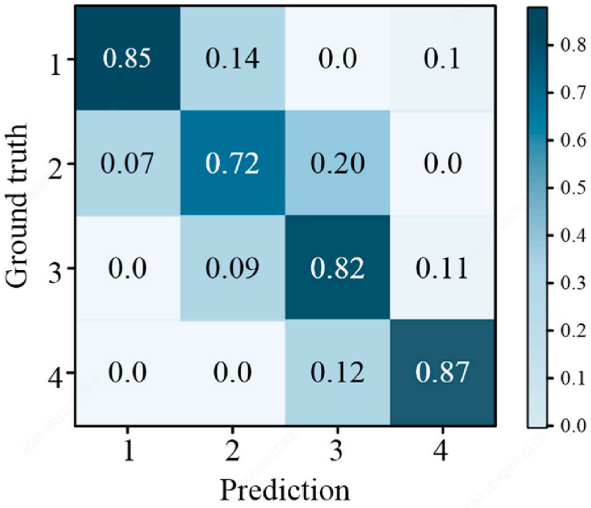

Fig. 6. Confusion matrix of the QoE model in this paper.

Table 4
Test time and performance comparison.

QoE model	Test time percentage	PLCC	SRCC	KRCC
RIRNet [27]	1	0.831	0.819	0.806
VSFA [26]	0.941	0.734	0.717	0.701
V-BLINDS [25]	0.327	0.587	0.573	0.556
VIDEO [24]	0.29	0.505	0.486	0.469
Ours	0.365	0.945	0.935	0.898

faced with variations in data ordering to determine whether the model is highly sensitive to changes in data ranks. This aids in evaluating the model's robustness in handling different sorting scenarios.

5. Conclusion

This study introduces a multimedia QoE evaluation method for wireless networks based on advanced deep learning models, with the goal of enhancing the quality of user experience. By modeling the video session process starting from the state of each time interval and analyzing the QoE prediction problem using RNNs, a QoE prediction model is proposed that incorporates video information, QoS data, user behaviors, and facial expressions. The experimental results indicate that the QoE model presented in this paper outperforms other models in terms of PLCC, SRCC, and KRCC performance metrics. This research provides new insights and methodologies for the study of video quality of experience in wireless networks and offers valuable references for both industry and academia. However, there are several research directions that warrant further exploration. Firstly, the structure and parameters of deep learning models can be further optimized to improve the accuracy and stability of QoE prediction. Secondly, additional influencing factors such as network congestion and device characteristics can be considered to enhance the overall performance of the QoE model. Furthermore, QoE evaluation methods specific to different application scenarios can be explored, allowing for customized model design and optimization for specific multimedia transmission services.

While our current experiments offer valuable insights into the performance of our model within controlled conditions, we acknowledge the importance of exploring real ambient scenarios. Adapting our methodology to on-demand scenarios requires a thorough understanding of diverse user preferences, content types, and interaction patterns. Future research avenues could involve collaborations with streaming platforms, conducting experiments in live streaming environments, or engaging in user studies across diverse geographical locations and network conditions.

Data availability statement

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

CRediT authorship contribution statement

Xiaoliang Zhang: Writing – review & editing, Writing – original draft, Supervision, Software. **Li Li:** Supervision, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xiaoliang Zhang reports article publishing charges was provided by Information Engineering School Jiaozuo Normal College. Xiaoliang Zhang reports a relationship with Information Engineering School Jiaozuo Normal College that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This article did not receive any financial support from institutions or individuals. The research was independently carried out by the author and contributing collaborators.

References

[1] T. Cao, C. Xu, J. Du, Y. Li, H. Xiao, C. Gong, D. Niyato, Reliable and efficient multimedia service optimization for edge computing-based 5G networks: game theoretic approaches, IEEE Transactions on Network and Service Management 17 (3) (2020) 1610–1625.

[2] P. Uthansakul, P. Anchuen, M. Uthansakul, A.A. Khan, Estimating and synthesizing QoE based on QoS measurement for improving multimedia services on cellular networks using ANN method, IEEE Transactions on Network and Service Management 17 (1) (2019) 389–402.

[3] S. Kesavan, E. Saravana Kumar, A. Kumar, K. Vengatesan, An investigation on adaptive HTTP media streaming Quality-of-Experience (QoE) and agility using cloud media services, Int. J. Comput. Appl. 43 (5) (2021) 431–444.

[4] G. Kougioumtzidis, V. Poulkov, Z.D. Zaharis, P.I. Lazaridis, A survey on multimedia services QoE assessment and machine learning-based prediction, IEEE Access 10 (2022) 19507–19538.

[5] P. Roy, S. Sarker, M.A. Razzaque, M.M. Hassan, S.A. AlQahtani, G. Aloi, G. Fortino, AI-enabled mobile multimedia service instance placement scheme in mobile edge computing, Comput. Network. 182 (2020) 107573.

[6] X. Min, G. Zhai, J. Zhou, M.C. Farias, A.C. Bovik, Study of subjective and objective quality assessment of audio-visual signals, IEEE Trans. Image Process. 29 (2020) 6054–6068.

[7] M. Morshedi, J. Noll, A survey on prediction of PQoS using machine learning on Wi-Fi networks, in: 2020 International Conference on Advanced Technologies for Communications (ATC), IEEE, Nha Trang, Vietnam, 2020, October, pp. 5–11.

[8] M. Hu, J. Chen, D. Wu, Y. Zhou, Y. Wang, H.N. Dai, TVG-streaming: learning user behaviors for QoE-optimized 360-degree video streaming, IEEE Trans. Circ. Syst. Video Technol. 31 (10) (2020) 4107–4120.

[9] M. Seufert, S. Wassermann, P. Casas, Considering user behavior in the quality of experience cycle: towards proactive QoE-aware traffic management, IEEE Commun. Lett. 23 (7) (2019) 1145–1148.

[10] J. Ruan, D. Xie, A survey on QoE-oriented VR video streaming: some research issues and challenges, Electronics 10 (17) (2021) 2155.

[11] M. Seufert, Fundamental advantages of considering quality of experience distributions over mean opinion scores, in: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), IEEE, Berlin, Germany, 2019, June, pp. 1–6.

[12] T. Hoßfeld, P.E. Heegaard, M. Varela, L. Skorin-Kapov, M. Fiedler, From QoS distributions to QoE distributions: a system's perspective, in: 2020 6th IEEE Conference on Network Softwarization (NetSoft), IEEE, Ghent, Belgium, 2020, June, pp. 51–56.

[13] F. Laiche, A. Ben Letaifa, T. Aguili, QoE-aware traffic monitoring based on user behavior in video streaming services, Concurrency Comput. Pract. Ex. (2021) e6678.

[14] F. Laiche, A. Ben Letaifa, I. Elloumi, T. Aguili, When machine learning algorithms meet user engagement parameters to predict video QoE, Wireless Pers. Commun. 116 (2021) 2723–2741.

[15] A.J. García, C. Gijón, M. Toril, S. Luna-Ramírez, Data-driven construction of user utility functions from radio connection traces in LTE, Electronics 10 (7) (2021) 829.

[16] K. Zhang, L. Chen, Y. An, P. Cui, A QoE test system for vehicular voice cloud services, Mobile Network. Appl. 26 (2021) 700–715.

[17] A.A. Laghari, H. He, K.A. Memon, R.A. Laghari, I.A. Halepoto, A. Khan, Quality of experience (QoE) in cloud gaming models: a review, Multiagent Grid Syst. 15 (3) (2019) 289–304.

[18] J. Zhao, S. Li, F. Hu, A QoE prediction model combining network parameters and video quality, in: 2022 International Conference on Culture-Oriented Science and Technology (CoST), IEEE, Lanzhou, China, 2022, August, pp. 31–35.

[19] X. Zhang, S. Li, F. Hu, Convolutional recurrent neural networks with attention mechanism for streaming QoE prediction, in: 2022 6th International Conference on Communication and Information Systems (ICCIS), IEEE, Chongqing, China, 2022, October, pp. 104–108.

[20] L.N. Onyejegbu, U.A. Okengwu, L.U. Oghenekaro, M.O. Musa, A.O. Ugbari, AI-based QOS/QOE framework for multimedia systems, in: Proceedings of the Future Technologies Conference (FTC), Springer International Publishing, Kohei Arai, 2022, October, pp. 248–259. Cham.

[21] M.B. Patil, R. Patil, Fractional squirrel–dolphin echolocation with deep belief network for network-controlled vertical handoff in disparate and heterogeneous wireless network, Int. J. Commun. Syst. 34 (12) (2021) e4893.

[22] K. Fizza, A. Banerjee, K. Mitra, P.P. Jayaraman, R. Ranjan, P. Patel, D. Georgakopoulos, QoE in IoT: a vision, survey and future directions, Discover Internet of Things 1 (2021) 1–14.

[23] W. Liu, N. Mehdipour, C. Belta, Recurrent neural network controllers for signal temporal logic specifications subject to safety constraints, IEEE Control Systems Letters 6 (2021) 91–96.

[24] W. Zhou, Z. Chen, Deep local and global spatiotemporal feature aggregation for blind video quality assessment, in: 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), IEEE, Macau, China, 2020, December, pp. 338–341.

[25] Z. Tu, C.J. Chen, L.H. Chen, N. Birkbeck, B. Adsumilli, A.C. Bovik, A comparative evaluation of temporal pooling methods for blind video quality assessment, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, Abu Dhabi, United Arab Emirates, 2020, October, pp. 141–145.

[26] D. Li, T. Jiang, M. Jiang, Quality assessment of in-the-wild videos, in: Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 2019, October, pp. 2351–2359.

[27] P. Chen, L. Li, L. Ma, J. Wu, G. Shi, RIRNet: recurrent-in-recurrent network for video quality assessment, in: Proceedings of the 28th ACM International Conference on Multimedia, Virtual Event, Seattle, WA), USA, 2020, October, pp. 834–842.