REGULAR PAPER

CrossMark

Modelling user quality of experience from objective and subjective data sets using fuzzy logic

Marko Matulin[1] · Štefica Mrvelj[1]

Abstract

One of the paramount research questions in the scientific community today is how to remotely assess user quality of experience (QoE) for a specific service. To this end, various user QoE assessment models have been developed; however, they are mostly based on the data gathered from controlled environment experimentation. The aim of this research was to model user QoE for the User Datagram Protocol-based video streaming service from the results of uncontrolled subjective tests. Specifically, using fuzzy logic, we have correlated the values of three objective network parameters (the packet loss rate and the number and duration of packet loss occurrences in one streaming session) with test subjects' subjective perception about perceived quality distortions. The dependencies between different values of the parameters and the subjects' perception of video quality were used to develop a no reference objective video quality assessment model for assessing user QoE. The key distinguishing feature of the developed model lays in the process of subjective evaluation, which was conducted with a panel of 602 test subjects who evaluated the quality of 1-h video in home environments. For this purpose, 72 different test sequences were prepared for rating. We showed that a strong positive linear relationship exists between the assessed QoE of the model and the Mean Opinion Scores of the subjects (a Pearson correlation coefficient equal to 0.8841).

Keywords Quality of experience · Home environment · Video quality assessment · Objective model · No reference

1 Introduction

The survival of telecommunication operators and service providers in today's competitive markets depends not only on well-known factors such as robust business plans, financial stability, wide service pallet, customer support, and technological innovations but also on the loyalty of indecisive users and their level of satisfaction, entertainment and enjoyment while using a specific service. Moreover, as user awareness of the provisioned service quality increases, the risk of user churn rises, which may lead to decreased income and market reputation, as discussed by Ahmad et al. in [1] and Nogueira et al. in [2].

Considering this context, it is clear that a number of quantitative and qualitative parameters affect user quality of experience (QoE). This introduces more complexity into the process of service quality evaluation, especially if a service is to be evaluated in uncontrolled environments. In such environments, measuring network performance and/or evaluating the service output (e.g., video or audio signal) using different metrics does not provide comprehensive insight into user QoE. To achieve a more complete understanding of user perception, different subjective parameters must be included in the evaluation process (for instance, level of user entertainment, stress and fatigue, past experience of service usage, and social context in which a service is used). This was previously discussed in [3], where the authors identified multiple QoE influential factors (IF) and grouped them into three main categories: human IF, system IF, and context IF.

In our past research presented in [4], we made an effort to discover to what extent specific human, system and context IF may impact user QoE for User Datagram Protocol (UDP)-based multimedia streaming services if the service is used in an uncontrolled environment, e.g., at home. We conducted analysis of user QoE by surveying 602 test

✉ Marko Matulin
mmatulin@fpz.hr

1 Department of Information and Communications Traffic, Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, 10000 Zagreb, Croatia

subjects who evaluated 72 different video sequences. The level of user QoE was analyzed against the following objective and subjective parameters: the packet loss rate, the number and duration of packet loss occurrences in one streaming session, subjects' level of annoyance, entertainment and fatigue, social context, and the existence/nonexistence of video subtitles. We also investigated the effect of human short-term memory and the recency effect.

Based on this research, in this contribution, we correlate two different data sets (objective and subjective) to develop a fuzzy-based no reference objective video quality assessment model for assessing user QoE. Since the developed model originates from a home environment experiment in uncontrolled conditions, it stands out in a group of similar models that are usually based on the results of laboratory testing in controlled environments. Conducting such uncontrolled subjective tests can be crucial for accurate QoE assessment, since it is shown in [5] that laboratory experiments lead to different results than the reality. This is more recently underlined in [6]. Primarily, the results of the uncontrolled experiments indicate that users are not as adversely affected by the occasional advent of quality degradations compared with the results obtained from the controlled experiments.

Our primary motivation in this research was to model human perception of video quality based on the results of the uncontrolled experiments. As will be shown in Sect. 3, only a few such video quality assessment models exist today. Since the results of the controlled and uncontrolled experiments are somewhat dissonant, we believe that these types of models have the ability to produce a more life-like assessment of user QoE. Second, we wanted to show how we have used fuzzy logic to bridge the gap between the objective and subjective data sets and calculate user QoE. Finally, we wanted to learn from the experience of developing such a model, which helped us define possible paths for future research.

After this brief introduction, Sect. 2 presents the background of this paper by discussing our past research, which is used to develop the video quality assessment model. In Sect. 3, the related work is discussed, while Sect. 4 presents the test results that are used to develop the inference system of the model. The first phase of the model development is described in Sect. 5, where the fuzzification process is discussed and fuzzy membership functions are presented. Section 6 explains how the output value (i.e., the assessed level of user QoE) of the model is calculated, while Sect. 7 discusses the correlation between the assessed QoE and the Mean Opinion Scores (MOS) of our test subjects. Apart from concluding remarks, Sect. 8 highlights the limitations of the developed model and the outlook of our research.

2 Background

This contribution is a follow-up of our past research, which was aimed at investigating how packet loss-related issues affect user perception of video quality. In that study, we tested user perception using 72 different test sequences, which were prepared in advance in an emulated network environment. The same type of content was used in all test sequences (1-h documentary film about the solar system); however, in each test sequence, packet loss rate (PLR), number of packet loss occurrences (PLOs), and duration of PLOs varied.

Details of the subjective test are discussed in the remainder of this chapter, but here, it is first necessary to recognize that we adopted the test methodology from [5]. The authors designed the experiment, which allows a researcher to prepare the test sequences in advance, thus retaining control over audiovisual quality. The sequences are then stored on, for instance, an optical drive or removable storage, and distributed to the subjects for rating in their home environments. Hence, subjective evaluation of the sequences is conducted in uncontrolled conditions, which is important for QoE evaluation, as emphasized in the Introduction.

Note that we have analyzed other approaches that include uncontrolled environments. For instance, in [7–9], the authors employed QoE crowdtesting, while in [10, 11], user network performances were remotely monitored during streaming sessions. These two approaches were not suitable for this study, because it would have been difficult to pursue the test subjects to download or stream 1-h video (i.e., several gigabytes of data) to their devices at home, which was necessary in [7–9] and [10, 11], respectively.

In [12], Staelens et al. prepared the test sequences in advance and stored them on tablet computers. The tablets were then distributed to the subjects, who watched the sequences in everyday conditions. The subjects rated the quality of the sequences directly on the tablets. The authors collected the rating data after the subjects returned the tablets. With this approach, video downloading or streaming is avoided, yet we decided not to follow it due for two reasons: (a) we did not have a sufficient number of tablet devices to conduct a large-scale study such as ours (602 test subjects) and (b) the player used for watching the test sequences contained a video quality rating scale; thus, the purpose of the test was revealed to the subjects. When the purpose of the test is known to the subjects, they are more focused on quality degradation during the test and less focused on content. This is unlike everyday service usage scenarios and also affects the user QoE rating. Finally, the group of authors in [13] implemented a QoE rating scale in the user interface of the VLC

Media Player. The player was installed on the subjects' devices, who used it for streaming multimedia content. This approach was also rejected since the visibility of the rating scale during playback reveals the purpose of the test to the subjects.

2.1 The experiment setup and creation of test sequences

In this experiment, test subjects evaluated the quality of 1-h video about the solar system in a home environment. We have used entertainment-oriented content selection [14], since we assumed that, while at home, the subjects usually watch video content that interests them. An additional reason for choosing the content that can entertain the subjects was the duration of the test, i.e., the video, and the need to capture the subjects' concentration for a full hour. We believe that if different content would have been used, that would have been perceived by the subjects as, for instance, boring, then their willingness to participate in our study would decline. Note that our test conditions were unlike controlled environment experimenting in laboratory, where the tests usually last 20–30 min, during which observers rate the quality of several short video clips. In such shorter tests, it is easier to retain the subjects' focus on the task at hand.

The experiment was performed using longer test sequences; thereby, we have acknowledged the findings presented in [15–17]; the authors in that study demonstrate how QoE evaluation requires an increase in the duration of the test sequences, because user perception cannot be entirely shaped when using shorter video clips. The original video used in the experiment was encoded with Advanced Video Coding (H.264/AVC) and Advanced Audio Coding (AAC). The video bitrate, audio bitrate and frame rate were 9.8 Mbps, 256 kbps and 50 fps, respectively. The video resolution was 1920×1080 pixels. Note that the video contained video subtitles.

To create test sequences, first, the original video was streamed six times between two computers, which were connected in a peer-to-peer connection (Fig. 1). The Network Emulator Client was installed on Computer 1. The client dropped the packets on the outgoing stream toward Computer 2. PLR varied between 0.05, 0.1, 0.5, 1, 1.5 and 2%, while the burst packet loss length was set to 1. VLC Media Player was used to stream the video between the computers. Each incoming (degraded) video signal was stored on Computer 2 in the same format as the original video. During streaming of the video between the computers, UDP was used on the transport layer. This differs from Hyper Text Transfer Protocol (HTTP)-based streaming, which uses TCP (Transport Control Protocol). Nowadays, HTTP-based adaptive streaming is the relevant scenario; however, UDP-based streaming is still used for delivering Internet Protocol Television (IPTV), in particular for those services that use set top boxes.

Second, we have imported the stored video signals into CyberLink Power Director and extracted 1, 4, 7 or 10 short video clips from a degraded video signal and inserted them into the original video signal. The duration of a single inserted clip, i.e., a single packet loss occurrence (PLO), varied between 1, 4 and 7 s. By varying the number of inserted PLOs and the duration of a single PLO, we were

Fig. 1 Test sequence creation process
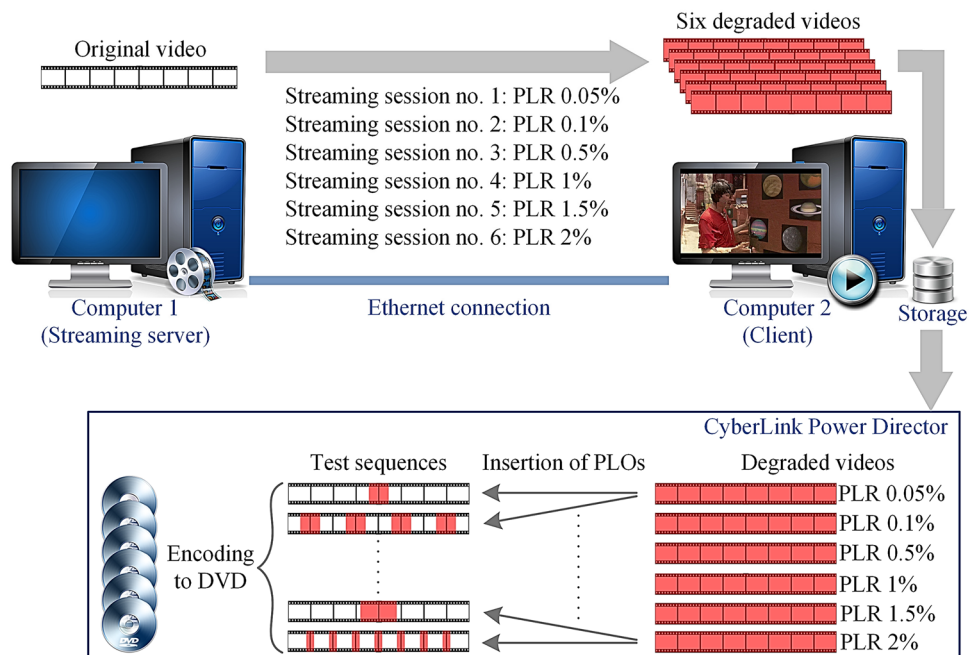
able to generate different total durations for all PLOs in a test sequence that equaled 1, 4, 7, 10, 16, 28, 40, 49 or 70 s. The total duration is derived by multiplying the number of inserted PLOs in a sequence (1, 4, 7 or 10) with the duration of a single PLO (1, 4 or 7 s). When selecting these particular values of the parameters, the objective was to generate a wide range of total duration of the distortions in the sequences. Additionally, we wanted to create sequences with equal total duration of distortion while containing a different number of inserted PLOs and different durations of a single PLO. For instance, two sequences can contain 28 s of total quality distortions, but one can have 7 PLOs each lasting 4 s, while the other has 4 PLOs each lasting 7 s. This enabled us to evaluate the impact of each parameter individually.

In [18], the authors revealed that when the distortions are grouped into the first few minutes of the video, the quality scores of the subjects are observed to increase. Conversely, if the distortions are grouped into the last few minutes, the scores are observed to decrease. This kind of subject reasoning is influenced by human short-term memory [19] and the psychological effect of recency [18, 20]. Thus, the PLOs were evenly distributed over the entire duration of all test sequences. However, in each test sequence, the first and last 7 min and 17 s were unaffected by the degradations, which allowed the test subjects to get involved with the content in the beginning of the session as well as critically think about the audiovisual quality toward the end.

In this study, the methodology used for the subjective evaluation of the sequences was adopted from [5]. Thus, the sequences were distributed to the subjects on a DVD. This format was chosen due to the following reasons: (a) the test sequences were easily distributed to the subjects; (b) DVD players are more available compared to Blu-ray players (this was important since the survey was conducted among a student population); (c) DVD disks are cheaper compared to, e.g., memory sticks; and (d) according to [21], while evaluating other services, test subjects often use the quality of DVDs as a reference.

During the encoding of the test sequences to the DVD format, the PAL system, MPEG-2 video encoding format and variable bitrate encoding method were used. All video encoding settings were set to maintain the best possible video quality. Furthermore, all video enhancement features of the CyberLink Power Director software were turned off and the software did not use any error concealment methods. The resulting video bitrate, audio bitrate and frame rate were equal to 9.51 Mbps, 256 kbps and 25 fps, respectively.

Although we have used the settings which allowed the best possible video quality for the conversion to the DVD format, this process downgraded the quality of the sequences, since the original video (encoded with H.264/AVC) was re-encoded into the MPEG-2 format. In general, this decline in the video quality is hard to notice in low-motion scenes but can be noticed in high-motion scenes when moving objects, appearing in a scene, can become blocky. The difference in the quality becomes even more visible to the subjects if they are provided with the original sequence for the comparison (that was not the case in our study). Since we conducted the experiment using the documentary film, most of the scenes in the film were low-motion scenes (for instance, the presenter's monologue or dialogue with other persons appearing in the film). Thus, the quality degradation caused by the re-encoding process was not apparent. Furthermore, we need to highlight that in our test conditions, which were mimicking a lifelike viewing experience, the subjects were not focused on keeping track of the video quality, they were focused on the content instead. Thus, we believe that the decline in the video quality due to the re-encoding of the video did not impact the QoE of our test subjects.

2.2 Data collection

A fuzzy-based no reference objective video quality assessment model for assessing user QoE, developed in this research, correlates the values of the three objective parameters (discussed in Sect. 2.1) with the subjects' perception of video quality. Hence, the focus of this section is to report how the subjective data set was collected, needed for the development of the inference system of the model.

2.2.1 Design of the questionnaire

The questionnaire used in the survey had four pages. Page 1 contained a detailed description of the purpose of the test and instructions on how to fulfil the questionnaire. Questions related to the perceived video quality were printed on page 2. Page 3 was used to investigate the subjects' opinions about the video content as well as the subjects' environment and the equipment used to reproduce the video, the social context in which they watched the video, level of fatigue and other factors. Page 4 contained general questions used to collect subject demographic information and a blank space, where the subjects were able to leave comments. Pages 2 and 3 of the questionnaire can be found in Appendix 1.

The questionnaire contained multiple choice questions as well as 11-point numerical scales (designed by ITU-T in [22]) for questions related to the subjective perception of the video quality and subjects' level of annoyance caused by the degradations. The decision to use an 11-point scale, over the more commonly used discrete five-level scale, was taken, because the aim was to collect the continuous data and provide the subjects with a larger span of possible answers. The scales enabled capturing natural ambiguity and fuzziness of the subjects' opinions. The use of discrete rating scales or, for instance, questions with

two-alternative options (such as *was the audiovisual quality of the video acceptable* with possible answers *yes* or *no*) would cause the loss of valuable information about the impact of the objective parameters on the subjects' perception in different viewing conditions. This is further discussed in Sect. 4.

Furthermore, several questions were used to detect the subjects' abnormal rating. For instance, if the subject indicated noticing only one quality degradation in the entire 1-h video but rated that frequency as Annoyingly high frequency, this rating was considered abnormal, and the questionnaire was rejected. Furthermore, we have rejected all questionnaires in which the subjects' indicated noticing video artifacts, which were unrelated to our experiment, i.e., a specific test sequence. We considered that in those instances, the subjects' equipment may have been malfunctioning, which may have interfered with the audiovisual presentation and their rating. The questionnaires were also rejected if the subjects responded positively to the statement *When I watch DVDs as I usually do, their quality is often degraded* or *There is a possibility that my DVD player that I used to watch the video may be broken or malfunctioning*. Additionally, we have asked the subjects to evaluate the level of noise in their surroundings, while they were watching the video. This information was used to exclude those questionnaires in which the subjects indicated that they were unable to concentrate on the video due to noise. The questionnaires were also rejected if the subjects did not complete them immediately after the screening; thus, might have forgotten the quality distortions they experienced, potentially leading to false ratings.

Further details regarding the reasons for questionnaire rejection can be found in [4], as well as the number of rejected questionnaires per specific rejection criteria. As can be seen from this discussion, we have rejected the questionnaires by employing the methods discussed in [23]; i.e., the questionnaire contained consistency questions, and we have investigated the hardware environment and hidden influence factors.

The questionnaires were distributed to the subjects in sealed envelopes. Two questionnaires have been inserted into each envelope (if the subjects watched the video with a company, they were asked to pass the second questionnaire to one person in their company). Furthermore, we have printed an illustration on the envelopes explaining how to proceed with the test, indicating four main steps: (1) Take the envelope and the attached video to your home; (2) watch the video in everyday conditions; (3) open the envelope immediately after watching the video, read the instructions and complete the questionnaire; and (4) return the questionnaire. The test sequences were attached to the outer side of the envelopes, so they were accessible to the subjects without the need to open the envelope.

2.2.2 The test subjects

Since both authors of this paper are the employees of the University of Zagreb, it was decided that the survey would be conducted among the student population of the university; i.e., the convenience sampling method was used [24]. Another reason for targeting this particular population can be found in [25], where Datta et al. reveal that persons between the ages of 18 and 24 are common users of video streaming services. This corresponds with the age group of a typical student population.

The subjects were approached and asked to participate in the survey, while they were in classes at the university. At each occasion, only a few key points of the research were presented to them; namely, we made it clear that:

- the survey is anonymous;
- the participation in the survey is not mandatory;
- those who wish to participate will be asked to:

  - take one envelope and the attached video;
  - keep the envelope sealed and open it only after watching the video;
  - watch the video only once in the conditions they would normally watch television;
  - open the envelope immediately after watching the video and read the instructions;
  - complete the questionnaire;
  - pass the second questionnaire to one person in their company (if applicable and if that person also watched the video with them);
  - return the completed questionnaire(s);

- the video content is 1-h documentary film about the solar system;
- the questionnaire takes approximately 10 min to complete;
- the illustration printed on the envelopes reminds them about the steps of participation in the survey;
- the survey lasts 2 weeks.

During this brief presentation of what is expected from the test subjects, the purpose of the test and the content of the questionnaire were not revealed in any way. After a period of 2 weeks, the collected questionnaires were processed, and the QoE analysis was continued on a sample of 602 test subjects.

2.3 Discussion about the obtained results of the subjective evaluation

The results obtained from the survey were grouped into different categories. Specifically, we have conducted analysis of user QoE for each test sequence, tested and confirmed the IQX hypothesis [26], and examined the relationships between the level of user annoyance and PLR, number of PLOs and total duration of all PLOs in a sequence. These relationships will be later used in the fuzzification process (in Sect. 5). Furthermore, we have investigated the impact of human short-term memory and the recency effect, correlated user QoE with their level of entertainment and fatigue, and analyzed the impact of social context and video subtitles on user QoE.

The user QoE analysis indicated that the subjects' MOS remained reasonably high (always above 4, on a scale from 0, being bad quality, to 10, being excellent quality), even for those sequences that contained the most PLOs. This confirmed the findings presented in [5, 6, 27], where different authors underlined how the results of the uncontrolled experiments suggest that the subjects are not so negatively affected by the perceived quality degradations. This has direct implications on the inference system of the model and its output. Specifically, Sect. 7 will demonstrate how the model assesses a user QoE of 4.48, even for the most degraded test sequence. Detailed analysis of user QoE showed that when there is only one PLO in a 1-h video, the PLR and the duration of a single PLO do not affect user QoE. For PLRs $\geq 1\%$, a quality degradation that lasts ≥ 16 s can be negatively perceived by users. Furthermore, if the video contains 7 or more PLOs and PLR increases ($\geq 1.5\%$), the duration of a single PLO comes to the fore. The analysis also revealed that, for PLRs of $\geq 0.5\%$, an increase in the number of PLOs significantly influences user QoE.

After confirming the IQX hypothesis, we ranked the objective parameters by their order of importance in relation to their impact on user QoE as follows: (1) total duration of quality distortions in a video, i.e., total duration of PLOs; (2) number of PLOs; (3) PLR; and (4) duration of a single PLO.

The impact of human short-term memory [19] was tested by comparing the number of PLOs in a specific sequence with the number of perceived PLOs reported by the subjects. The analysis revealed that a considerable number of test subjects (408) failed to notice and/or memorize some or even all quality distortions inserted in the sequences. This can be related with three casualties. First, longer test sequences were used in the test. Presumably, after 1 h, some subjects forgot the degradations, which they may have noticed while watching the DVD. Second, experimenting in the home environment of the subjects encouraged them to watch the video in everyday conditions (at a known location, with or without company, at any time of day). In these lifelike viewing conditions, the subjects were not focused on noticing and memorizing the PLOs; they were focused on the content instead. Third, the subjects were uninformed about the purpose of the test. Thus, before watching and during the video, they were unaware of the degradations that would appear in a sequence. This inability to notice and/or memorize PLOs impacted the subjects, which was manifested as high MOS even for the most degraded test sequences (as previously discussed). We have also discovered how the PLOs that appeared in the middle of the video were more often unreported by the subjects compared to those PLOs that appeared toward the end of the video. This confirmed the impact of the recency effect [20] on our test subjects. Notwithstanding, since the test sequences used in this experiment lasted 1 h, we cannot neglect that these results may differ if shorter test sequences had been used (or other types of content).

It is worth mentioning that the results also disclosed how the overall user experience can be redeemed despite the perceived quality distortions if the content is entertaining to the viewer. Finally, a separate analysis was conducted to see if the video subtitles could draw viewers' attention to the bottom of the screen, thus making the PLOs harder to notice. It was observed that the subjects who watched the video with subtitles noticed fewer PLOs and achieved higher QoE compared with the subjects that watched the video without subtitles. However, we emphasized that further investigation of the impact of video subtitles on user QoE is needed.

2.4 Critical overview of the methodology

In terms of network-related parameters, which may impact user audiovisual perception, it can be observed that we have limited our research quest to the effect of packet loss-related issues on user QoE. The inclusion of more parameters, such as network delay, jitter, and throughput, increases the number of test sequences. We have created 72 test sequences just by combining different values of the three parameters (PLR, number and duration of PLOs in a test sequence). A larger number of test sequences would mean that we would have to reach more test subjects, which was considered unfeasible. Note that each of the 602 test subjects evaluated one test sequence by watching it once (as discussed in Sect. 2.2.2). If the subjects watched the same sequence more than once before completing the questionnaire or if they watched another sequence for the first time, they would have known the purpose of the test while watching the video. We underlined earlier that the goal was to avoid that since knowing the purpose of the test would make our test subjects more perceptive to the PLOs.

Apart from the sheer number of test subjects, it would be difficult to pinpoint the effect of packet loss on user QoE if other network-related parameters had been tested as well. Hence, it can be argued that not all the parameters were

tested against the subjects' perception, yet the impact of packet loss on a wide number of subjective parameters was tested meticulously.

When critically thinking about the test conducted in this study, it has to be taken into account that the methodology had to meet the following demands:

- the primary objective was to collect the rating data and use it to develop the QoE assessment model that would be able to produce more lifelike QoE assessments; thus, the data had to be collected from uncontrolled experiments (in a home environment);
- longer test sequences had to be used in the test, because short video clips are not adequate for user QoE evaluation (as reported in [15–17]);
- the test sequences had to be distributed to the subjects for rating in a manner that bypasses downloading or streaming of the video (due to its size);
- the accuracy of the model depends on the fuzzification and defuzzification processes, i.e., obliquely on the size of the data set used for the model development; that is, a sufficient number of test sequences of different properties had to be generated and evaluated by a sufficient number of test subjects;
- it was unfeasible to conduct the interviews with such a large number of test subjects (602); thus, the subjects' opinions were collected using hard copy questionnaires.

The methodology discussed in this chapter met all of the above demands and highly impacted the obtained results. While watching the video at their homes, in a familiar environment, possibly surrounded by known people, the test subjects were not focused on keeping track of the video quality fluctuations. We can also assume that the subjects were able to relax and were entertained by the video content (on a scale from 0, being least entertaining, boring, to 10, being very entertaining, the average level of entertainment was 7.62 with a margin of error of 0.15 and a confidence level of 95%). This home test environment made the subjects more forgiving to the perceived quality distortions, which was mirrored in the results. Hence, the test environment directly affects the inference system of the model.

We are aware that the methodology has certain disadvantages. Namely, the success of such uncontrolled experiment largely depends on the honesty of the test subjects. Moreover, the test was conducted in environments, where a number of QoE influential factors may impact the subjects' rating (as discussed in [3]). We were not able to investigate all the factors on such a large target group. However, we invested effort in (a) clearly presenting what is expected from the test subjects in the study; (b) designing a questionnaire that returns enough information for the modelling; (c) discovering and rejecting outliers from the sample; and (d) removing

those questionnaires, where the subjects' answers indicated equipment malfunction or noisy environments that may have interfered with the viewing experience.

The statistical analysis of the collected data (conducted in [4]) and the obtained results that confirmed the findings of other authors allows us to argue that the objective of the test was achieved and its outcomes can be further used for the development of the model.

3 Related work

Revealing user perceptions and QoE for specific services can be a resource-consuming task. Thus, various objective quality assessment models are developed that can estimate subjective perception of quality based on the values of the network-related parameters. ITU-T in [28] categorized the objective video quality assessment models as full reference (FR), reduced reference (RR) and no reference (NR) models, depending on the availability of a reference (unprocessed) video signal for the assessment. To measure, assess or predict the quality of a video signal, the objective models employ the following statistical models [29]: (a) media-layer models; (b) packet-layer models; (c) bitstream-layer models; (d) hybrid models; and (e) planning models. This ITU-T classification provides a generally accepted and commonly referenced framework for classification of different metrics and quality assessment models.

In the following overview of the related work, the focus will be on the objective assessment models that combine objective and subjective data sets to deliver the output. Note that some authors call this type of model a hybrid model (for instance, [30]), but they must be distinguished from the aforementioned hybrid statistical models. The focus of this literature overview means that different metrics and models, such as MSE (mean square error), PSNR (peak signal-to-noise ratio), MPQM (moving picture quality metrics), MSAD (modified sum of absolute difference), SSIM (structural similarity) Index, VQM (video quality model) and others, will not be reviewed since their output is calculated without taking into account user perception, which is critical for QoE evaluation. In addition, this review includes the work of other researchers who use different techniques to make a correlation between objective and subjective data sets, regardless of the transport layer protocol that has been used in the test.

Different authors have attempted to improve standard video quality metrics. For instance, Chan et al. in [31] introduced three types of modifications to the PSNR metrics using subjective test results to derive more accurate MOS assessments. The subjective evaluation of different videos involved 21 test subjects, and the obtained results were used for the modifications. Another example of this type of

metrics modification (for MSE metrics and SSIM Index) is presented in [32, 33], respectively.

In [34], the authors showed how the random neural network can be trained with a subjective data set to assess the quality of a video stream on the receiver side. The subjective quality tests were conducted with a panel of 20 test subjects who evaluated the impact of stream bitrate, frame rate, packet loss rate, the burst packet loss length and the ratio of the encoded intra to inter macro-blocks on their perceived video quality. Similarly to [34], in [35], the authors also use the neural network for assessing the QoE of HTTP video streaming, taking into account the effects of pause position. Subjective evaluation was conducted among 54 test subjects (60 s video clips were used for testing). Another neural network application for QoE assessment for HTTP video streaming can be found in [36], while in [37], the authors use the network for QoE assessment of 3D video.

Mok et al. in [38] correlate user QoE with network QoS (quality of service) also for the HTTP video streaming service. Application layer QoS is expressed with three parameters (initial buffering time, mean re-buffering duration and re-buffering frequency), while the MOS are obtained by surveying a panel of test subjects who evaluated an 87-s video clip under different network conditions. Further information about the QoE of HTTP adaptive streaming can be found in the extensive literature overview presented in [39].

Apart from the abovementioned neural networks, other techniques can be used to correlate between user perception and objective parameters. A machine learning approach was employed by Menkovski et al. to develop the objective model, which can determine the extent of video quality degradations that may lead to unacceptable video quality as perceived by test subjects (see [40, 41]). The test subjects watched the sequences several times (the extent of quality distortions was increased each time) and indicated what they perceived as the acceptable limit of quality degradation in the sequence. Another example of testing the user acceptance of video quality can be found in [42]. The authors tested the user acceptance of mobile video against different objective parameters (re-buffering frequency, bitrate, frame rate, etc.) and developed a model for the acceptability of the quality of a mobile video session.

In [43, 44], the authors construct a k-dimensional Euclidian space, where k represents the number of network dependent and independent parameters that may affect user QoE. The space is then divided into N zones, and each zone is assigned with a QoE index; i.e., the authors define zones, where different values of various parameters lead to the same QoE rating. The QoE index is derived from subjective quality tests that were conducted among 77 test subjects who evaluated 18 test sequences. Later, Robalo and Velez in [45] use the results presented in [44] for mapping between the QoS and QoE.

Nguyen et al. in [46] apply a mixed effects model to predict user QoE for World Wide Web-based multimedia services. It is noteworthy that the authors include the state of mind parameter in their model (values of the parameter were: normal, bored, and stressed), proving how the QoE concept assumes a holistic approach to service evaluation. A further example of how network independent parameters may impact user QoE is presented in [47, 48], where different authors examine user perception of quality degradations while watching different types of video content. The obtained knowledge is used for development of the content-aware QoE assessment model.

In [49], the authors carried out subjective tests for the purpose of examining the perceptual experience of time-varying video quality. Based on the obtained results, the authors proposed an asymmetric adaptation model capable of mimicking human opinions when watching video with time-varying quality. Zhang et al. in [50] use fuzzy decision trees to predict user QoE from the log data collected from different Internet video service providers in China. After processing the raw data sets (the sets contained information about video ID, content and video type, client IP address and location, access device, join time, frame rate, bandwidth, buffering times and buffering ratio), the authors model user engagement as a key aspect of viewer behavior that, they claim, in some sense reflects users' QoE. Hameed et al. in [51] also use decision trees for the construction of a low-complexity video quality model that predicts user QoE. For this purpose, the authors prepared 288 test sequences, which were evaluated by 100 test subjects under controlled conditions.

The impact of viewing distance on user QoE was analyzed in [52], and no reference QoE assessment model is proposed. However, the authors explored the regularities between image QoE and viewing distance for different types of images but not for video. The QoE crowdtesting platform is used in [9] to collect subjects' opinions about the quality of adaptive media playout. The results were used to develop a nonlinear model that is able to describe user QoE relative to the audio/video distortions. Although the test is conducted in uncontrolled environments, the test sequence used in the study lasted only 51 s. Note that we have previously underlined that QoE evaluation requires using longer test sequences (based on the findings presented in [15–17]).

The models and metrics presented here are all based on subjective test results obtained in controlled laboratory environments, with two exceptions: [9, 50]. However, in [9], only short video clips were used for testing user perception, while in [50], the QoE is predicted from user engagement behavior and not from actual subjective quality tests. This lack of assessment models that originate from uncontrolled experiments was our primary motivation in this study.

4 Test results that are used to develop the inference system of the model

The model presented in this contribution assesses the level of user QoE using the values of the three objective parameters as inputs, namely, the PLR, the number and total duration of PLOs in a sequence. For the purpose of fuzzification of scalar values of these three parameters, we asked the subjects to evaluate the level of their annoyance in relation to: (a) the observed quality distortions, i.e., perceived video artifacts (in Fig. 2, we correlated these responses with the scalar values of PLRs); (b) the number of PLOs that they have noticed (Fig. 3); and (c) the total duration of all quality distortions in the video (Fig. 4). The subjects rated the level of their annoyance on an 11-point scale, designed by ITU-T in [22], which allows the linguistic meanings of different grades to be added as an aid during rating. The meanings that are used in this study are depicted on the secondary *y* -axes of Figs. 2, 3 and 4.

It can be observed that, for a given value on the *x*-axis of the figures, the test subjects' level of annoyance is sometimes spread over all the annoyance level categories. This is most noticeable for the data presented in Fig. 2 and can be explained by acknowledging that, for instance, a PLR of 2% was perceived as imperceptible quality distortion when the sequence contained only one PLO. However, the same PLR was perceived as very annoying quality distortion if the sequence contained 10 PLOs, with 70 s of quality degradations in total. Similar results are presented in [53], where the authors also discuss the ambiguity between the PLR and the MOS values.

In Sect. 2.2.1, we announced that this chapter will elaborate further the importance of using 11-point continuous scales for collecting the subjects' opinions. Thus, let us assume that, instead of 11-point scales, five-level discrete rating scales have been used in the survey. In that case, the results shown in Figs. 2, 3, and 4 would differ. For instance, all data points depicted in Fig. 2, for a given value of PLR,
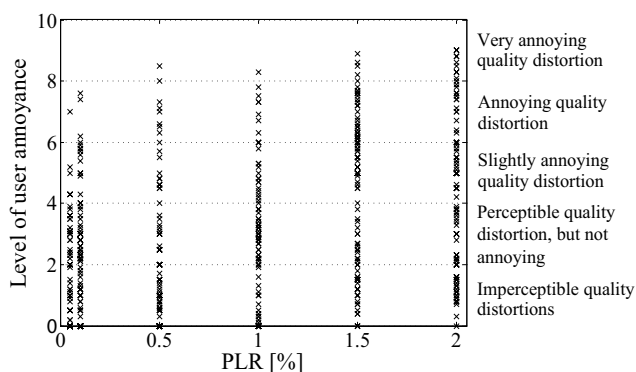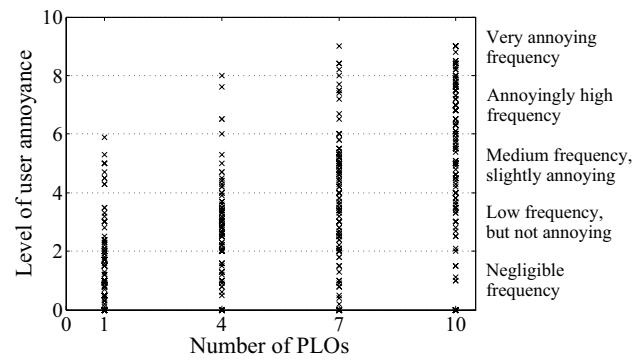

Fig. 3 Number of PLOs vs. level of user annoyance

would be grouped in no more than five discrete values on the *y*-axis. The results would appear less ambiguous and, due to the loss of the fuzziness of the data, we could conclude that the impact of PLR on the annoyance level of the subjects is clearer than it is. This also applies to the results shown in Figs. 3 and 4.

After correlating the values of the three objective parameters with the test subjects' subjective rating, the analysis of the user QoE for each of the 72 test sequences was conducted. The results are presented in Table 1. The subjects also rated the quality of the sequences on an 11-point quality scale that contained the following linguistic meanings: 0–2 bad quality; 2–4 poor quality; 4–6 fair quality; 6–8 good quality; and 8–10 excellent quality.

The obtained MOS varied on the interval [4.16, 8.96], meaning that even the most degraded test sequence with 70 s of quality distortions caused by PLR of 2% was still evaluated as a test sequence of fair quality (the average rating for the sequence number 72). In general, we can observe that the average QoE ratings, i.e., MOS, remain high despite the quality degradations of the specific sequences. For instance, from Fig. 4, it can be seen that almost all test subjects, who rated the sequences with 70 s of quality distortions, perceived that
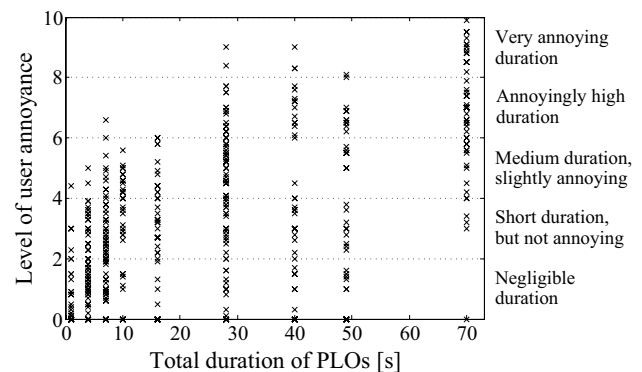

Fig. 2 Packet loss rate vs. level of user annoyance

Fig. 4 Total duration of PLOs vs. level of user annoyance

Table 1 Test sequence (TS) properties and their MOS (the properties are presented in brackets, where the numbers have the following meanings: PLR number of PLOs; duration of a single PLO; total duration of PLOs)

TS no.	TS properties	MOS	Linguistic meaning
1	(0.05%; 1; 1 s; 1 s)	8.41	Excellent quality
2	(0.1%; 1; 1 s; 1 s)	8.08	Excellent quality
3	(0.5%; 1; 1 s; 1 s)	8.10	Excellent quality
4	(1%; 1; 1 s; 1 s)	8.81	Excellent quality
5	(1.5%; 1; 1 s; 1 s)	8.96	Excellent quality
6	(2%; 1; 1 s; 1 s)	8.40	Excellent quality
7	(0.05%; 1; 4 s; 4 s)	8.42	Excellent quality
8	(0.1%; 1; 4 s; 4 s)	7.89	Good quality
9	(0.5%; 1; 4 s; 4 s)	8.10	Excellent quality
10	(1%; 1; 4 s; 4 s)	8.75	Excellent quality
11	(1.5%; 1; 4 s; 4 s)	8.26	Excellent quality
12	(2%; 1; 4 s; 4 s)	8.17	Excellent quality
13	(0.05%; 1; 7 s; 7 s)	8.32	Excellent quality
14	(0.1%; 1; 7 s; 7 s)	7.90	Good quality
15	(0.5%; 1; 7 s; 7 s)	7.46	Good quality
16	(1%; 1; 7 s; 7 s)	8.63	Excellent quality
17	(1.5%; 1; 7 s; 7 s)	7.85	Good quality
18	(2%; 1; 7 s; 7 s)	8.19	Excellent quality
19	(0.05%; 4; 1 s; 4 s)	8.41	Excellent quality
20	(0.1%; 4; 1 s; 4 s)	7.84	Good quality
21	(0.5%; 4; 1 s; 4 s)	8.59	Excellent quality
22	(1%; 4; 1 s; 4 s)	8.23	Excellent quality
23	(1.5%; 4; 1 s; 4 s)	7.66	Good quality
24	(2%; 4; 1 s; 4 s)	7.67	Good quality
25	(0.05%; 4; 4 s; 16 s)	8.20	Excellent quality
26	(0.1%; 4; 4 s; 16 s)	7.73	Good quality
27	(0.5%; 4; 4 s; 16 s)	6.83	Good quality
28	(1%; 4; 4 s; 16 s)	7.56	Good quality
29	(1.5%; 4; 4 s; 16 s)	7.06	Good quality
30	(2%; 4; 4 s; 16 s)	6.70	Good quality
31	(0.05%; 4; 7 s; 28 s)	7.74	Good quality
32	(0.1%; 4; 7 s; 28 s)	7.60	Good quality
33	(0.5%; 4; 7 s; 28 s)	7.60	Good quality
34	(1%; 4; 7 s; 28 s)	6.54	Good quality
35	(1.5%; 4; 7 s; 28 s)	6.29	Good quality
36	(2%; 4; 7 s; 28 s)	5.88	Fair quality
37	(0.05%; 7; 1 s; 7 s)	8.07	Excellent quality
38	(0.1%; 7; 1 s; 7 s)	7.89	Good quality
39	(0.5%; 7; 1 s; 7 s)	7.03	Good quality
40	(1%; 7; 1 s; 7 s)	7.80	Good quality
41	(1.5%; 7; 1 s; 7 s)	6.84	Good quality
42	(2%; 7; 1 s; 7 s)	6.02	Good quality
43	(0.05%; 7; 4 s; 28 s)	8.20	Excellent quality
44	(0.1%; 7; 4 s; 28 s)	7.13	Good quality
45	(0.5%; 7; 4 s; 28 s)	6.84	Good quality
46	(1%; 7; 4 s; 28 s)	6.33	Good quality
47	(1.5%; 7; 4 s; 28 s)	5.88	Fair quality

Table 1 (continued)

TS no.	TS properties	MOS	Linguistic meaning
48	(2%; 7; 4 s; 28 s)	5.63	Fair quality
49	(0.05%; 7; 7 s; 49 s)	7.63	Good quality
50	(0.1%; 7; 7 s; 49 s)	6.75	Good quality
51	(0.5%; 7; 7 s; 49 s)	6.74	Good quality
52	(1%; 7; 7 s; 49 s)	6.23	Good quality
53	(1.5%; 7; 7 s; 49 s)	5.04	Fair quality
54	(2%; 7; 7 s; 49 s)	4.84	Fair quality
55	(0.05%; 10; 1 s; 10 s)	7.89	Good quality
56	(0.1%; 10; 1 s; 10 s)	7.87	Good quality
57	(0.5%; 10; 1 s; 10 s)	7.67	Good quality
58	(1%; 10; 1 s; 10 s)	6.89	Good quality
59	(1.5%; 10; 1 s; 10 s)	6.01	Good quality
60	(2%; 10; 1 s; 10 s)	6.03	Good quality
61	(0.05%; 10; 4 s; 40 s)	7.87	Good quality
62	(0.1%; 10; 4 s; 40 s)	7.18	Good quality
63	(0.5%; 10; 4 s; 40 s)	6.62	Good quality
64	(1%; 10; 4 s; 40 s)	6.41	Good quality
65	(1.5%; 10; 4 s; 40 s)	5.17	Fair quality
66	(2%; 10; 4 s; 40 s)	5.35	Fair quality
67	(0.05%; 10; 7 s; 70 s)	7.59	Good quality
68	(0.1%; 10; 7 s; 70 s)	7.10	Good quality
69	(0.5%; 10; 7 s; 70 s)	5.93	Fair quality
70	(1%; 10; 7 s; 70 s)	4.68	Fair quality
71	(1.5%; 10; 7 s; 70 s)	4.57	Fair quality
72	(2%; 10; 7 s; 70 s)	4.16	Fair quality

duration as annoying (ranging from slightly to very annoying); yet, this was not entirely reflected on their QoE level. As discussed in Sects. 2.3 and 2.4, the methodology for the subjective evaluation of the sequences highly impacted these results. Specifically, the subjects' unawareness of the purpose of the test, familiar home environment, longer test sequences, content that was mainly evaluated as mostly and very entertaining, affected the subjects, making them more forgiving to the occasional advent of video degradations.

We also emphasize that the video contained subtitles. In our previous research, we found reasons to believe that the sequences with subtitles can be rated higher by the subjects compared with the sequences without subtitles since the text can draw the viewer's attention away from the picture to the bottom of the screen. Hence, some quality distortions may remain unnoticed by the subjects.

5 Fuzzification of scalar values

The objective of this chapter is to discuss the first stage of the model development. Namely, the chapter will show how the scalar values of the input and the output parameters are

converted into the fuzzy variables which are then used by the inference system of the model. However, first, we want to justify the decision to use fuzzy logic for the modelling of user QoE.

As has been pointed out in Sect. 4, when the results presented in Figs. 2, 3, and 4 were discussed, the impact of the three objective parameters on the subjects' perception is not decisive. Yet, the ambiguity of the obtained results was expected mainly due to the following reasons.

- The test was conducted with a large number of test subjects, using a large number of test sequences of varying quality; thus, the natural ambiguity of human opinions surfaced.
- The subjects watched the sequences in real-life and uncontrolled test conditions.
- The design of the questionnaire and the use of 11-point numerical scales allowed collecting continuous data.
- The combined impact of all three objective parameters on test subjects' perception.

To cope with the uncertainty in the results, we have used fuzzy logic which is useful when trying to characterize concepts and phenomena with natural ambiguity [54]. Additionally, the logic allowed us to model the relationships between the objective and the subjective data sets which will be demonstrated in the remainder of this chapter.

5.1 Defining the clusters and fuzzy membership functions for the input parameters

In this step, the fuzzification of the crisp values shown in Figs. 2, 3, and 4 was conducted using the Fuzzy C-Means (FCM) clustering approach [55]. The objective of using this procedure was to group the data points presented on the figures into fuzzy clusters and to find centers of those clusters. Note that the FCM method allows the clusters to overlap. Therefore, the ith data point (x_i) can be a member of several clusters (j) with different degrees of membership (u_{ij}). According to [55], the method requires minimizing the objective function J_m:

$$J_m = \sum_{i=1}^{L} \sum_{j=1}^{C} u_{ij}{}^m \cdot \left\| x_i - c_j \right\|^2, \tag{1}$$

where c_j denotes the d-dimension center of the cluster, $||*||$ is any norm expressing the similarity between any measured data and the center, and m is any real number > 1. Note that in this study, $m = 2$. Using this method for grouping the data points from the figures into fuzzy clusters and finding the centers of those clusters assumed the iterative process, where the u_{ij} and the c_j are updated with Eqs. 2 and 3, respectively. The iterative process finishes when the stopping criteria (ε)

from Eq. 4 is met (k are the iteration steps). In our case, we used the value of ε that is set by default in MATLAB to 10^{-5}.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{2}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}{}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}{}^m} \tag{3}$$

$$\max_{ij} \left\{ \left| u_{ij}{}^{(k)} - u_{ij}{}^{(k-1)} \right| \right\} < \varepsilon \tag{4}$$

For each of the three input parameters, the subjects' responses are grouped into three fuzzy clusters. The procedure required 24, 69 and 33 iterations before the stopping criteria ε was met and the centers of the clusters were defined for the PLR, Number of PLOs and Total duration of PLOs, respectively. The results are presented in Figs. 5a, 6a, and 7a, respectively. Note that thickly printed plus, × and circle signs on the figures represent the cluster centers, and their coordinates can be found in Table 2.

The final step toward defining the boundaries of the clusters was taken when the values of the data points on the x axis were correlated with their degrees of membership u_{ij} to the specific cluster j (see the Appendix 2, Figs. 11, 12, 13). In his book about fuzzy logic and its engineering applications [54], Ross reports how the normal distribution corresponds better with the changes in human perception, because the transitions between different opinions and attitudes are usually happening gradually ("smoothly"). Thus, bell-shaped functions can mimic that behavior better compared with, for instance, triangular functions. For this reason, we used the normal distribution for the approximation of the membership functions. Note also that the "smooth" transitions in human perception were captured with the use of continuous 11-point numerical scales for subjective data collection. The "smoothness" would be lost if the five-level discrete scales would have been used or questions with two-alternative options (as discussed in Sects. 2.2.1 and 4).

The obtained functions are presented in Figs. 5b, 6b, and 7b. Since the fuzzy systems use linguistic variables to infer conclusions, each cluster presented on the figures is named. These linguistic variables will be used in Sect. 6.1 for defining a set of fuzzy rules of the model.

Properties of the fuzzy membership functions shown in Figs. 5b, 6b, and 7b can be found in Table 3. Note that the dashed and the dashed–dotted Gaussian functions of each input parameter had to be modified so that u_{ij} would be equal to 1 in cases when $x_i < \bar{x}$ (for the dashed functions) and $x_i > \bar{x}$ (for the dashed-dotted functions).
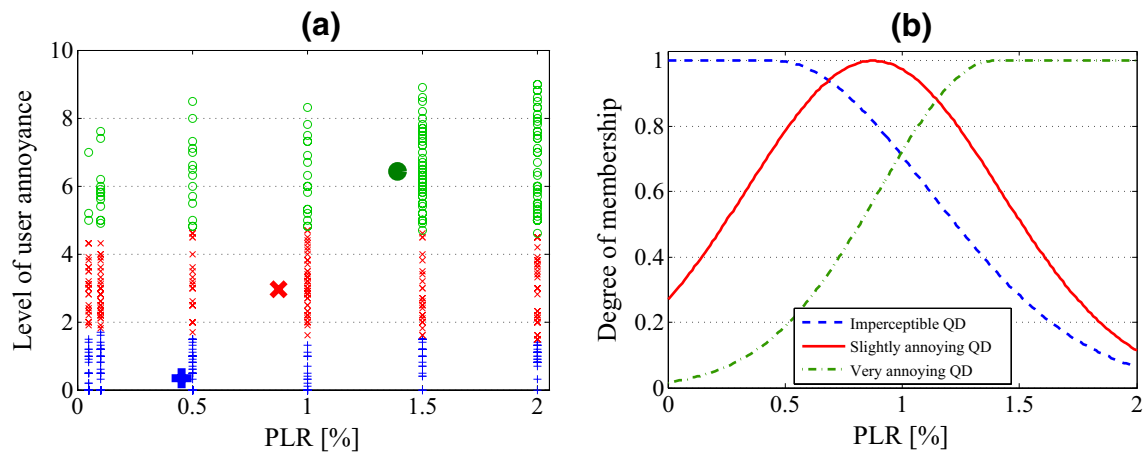
Fig. 5 Results of FCM clustering for the first input parameter: **a** three clusters and their respective centers and **b** membership functions of the three fuzzy clusters (QD stands for quality distortion)
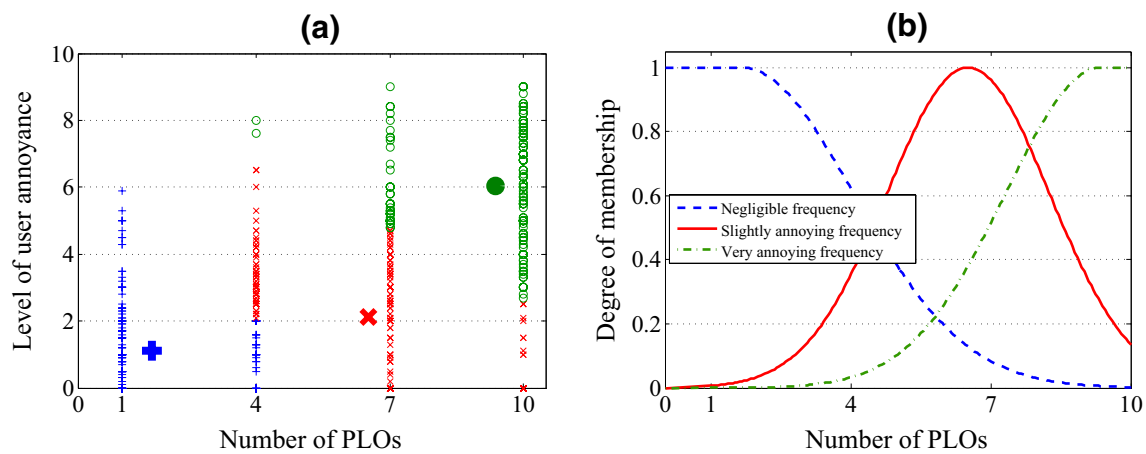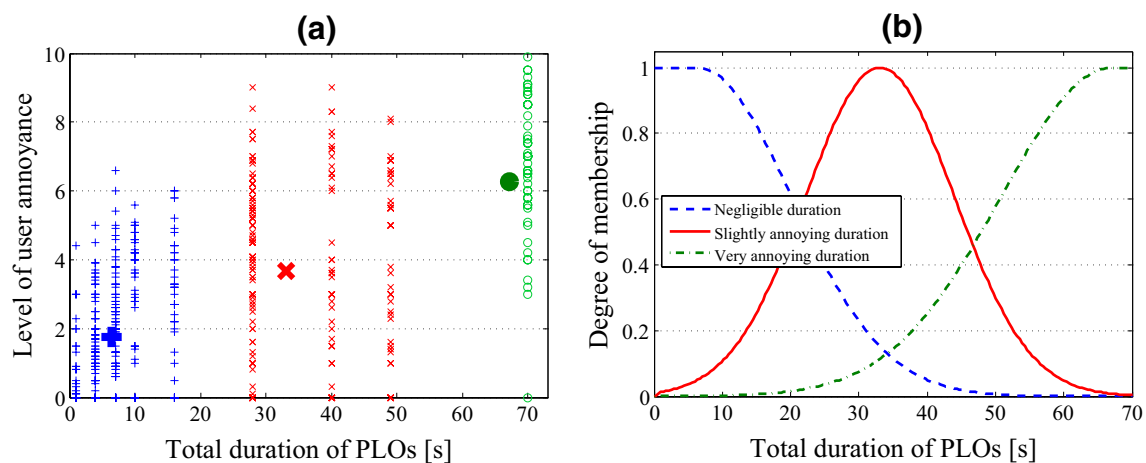

Fig. 6 Results of FCM clustering for the second input parameter: **a** three clusters and their respective centers and **b** membership functions of the three fuzzy clusters

Fig. 7 Results of FCM clustering for the third input parameter: **a** three clusters and their respective centers and **b** membership functions of the three fuzzy clusters

Table 2 Centers of the clusters shown in Figs. 5a, 6a, and 7a

Figure	Cluster 1 (plus sign)		Cluster 2 (× sign)		Cluster 3 (circle sign)	
	x	y	x	y	x	y
Figure 5a	0.4545	0.3492	0.8758	2.9503	1.3937	6.4334
Figure 6a	1.6513	1.1024	6.5083	2.1201	9.3728	6.0449
Figure 7a	6.4254	1.7545	33.0713	3.6854	67.1134	6.2761

The mean values ($\bar{x}$) of the Gaussian functions depicted in Figs. 5b, 6b, and 7b correspond with the x-coordinates of the centers of the clusters presented in Table 2 and Figs. 5a, 6a, and 7a, respectively. This means that the data points have a higher degree of membership (u_{ij}) to a specific cluster j if they are located closer to the center of that cluster. Therefore, the shift in the sample population mean in this case is actually impacted by the number of clusters and the location of their respective centers.

We can observe how the fuzzy membership functions are overlapping due to the dispersion of the data points around the centers of the clusters. The overlapping of the membership functions is a key feature of the FCM procedure since it enables grouping a specific data point into more than one cluster. This is important, because it is not possible to unambiguously define, for instance, whether PLR of 1% is causing imperceptible, slightly or very annoying quality distortion from a user point of view. In fact, our results show that this particular PLR is a member of all three clusters (Fig. 5b) with different degrees of membership.

The most distinctive overlapping of the membership functions is present for the first input parameter (Fig. 5b). The stretch of these functions (by a factor of σ) shapes their slopes, which are not so steep compared with the functions depicted on the other two figures (Figs. 6b, 7b). This is due to the previously discussed fact that certain PLRs were evaluated differently by the subjects, depending on the number of PLOs in a test sequence and their total duration. These results give us solid arguments for inclusion of the three parameters in the inference system of the model since they clearly create an affiliated effect on user perception.
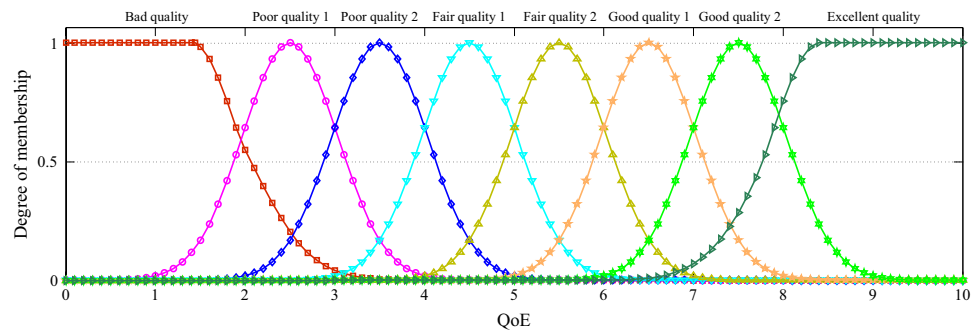
5.2 Development of fuzzy membership functions for the output parameter

We noted in Sect. 4, how the subjects evaluated the quality of the video and their viewing experience on an 11-point quality scale that also contained linguistic meanings of the specific ratings. These linguistic meanings were used to name the fuzzy clusters of the output parameter. However, to increase the accuracy of the model, instead of five clusters, the output parameter is modelled with eight clusters. While experimenting with different settings of the model, we have observed that it is better to increase the number of clusters of the output parameter, because the model became more responsive to the changes of the input values and assessed the QoE with more accuracy.

The membership functions of the eight fuzzy clusters of the output parameter are depicted in Fig. 8. Again, the Gaussian functions are used (their properties can be found in Table 4). For the clusters bad and excellent quality, the functions are modified so that u_{ij} would be equal to 1 in cases when $x_i < \bar{x}$ and $x_i > \bar{x}$, respectively. Note that the membership function of the excellent quality cluster has $\bar{x} = 8.44$, so that the ratings ≥ 8 would have a higher degree of membership to this cluster compared with the cluster good quality *2*.

6 Defuzzification to the output of the model

The previous chapter showed how the crisp values of the input and the output parameters are converted into the fuzzy variables needed for development of the inference system of the model. However, the model has to produce a quantifiable output, i.e., the assessed QoE rating in a numeric form. Therefore, the fuzzy values have to be defuzzified. This requires defining a set of fuzzy rules, choosing between

Table 3 Properties of the Gaussian functions shown in Figs. 5b, 6b, and 7b

Figure	First membership function (dashed line)		Second membership function (solid line)		Third membership function (dashed–dotted line)	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Figure 5b	0.4545	0.6574	0.8758	0.5398	1.3937	0.4887
Figure 6b	1.6513	2.4	6.5083	1.748	9.3728	2.061
Figure 7b	6.4254	13.73	33.0713	10.92	67.1134	16.33

Fig. 8 Membership functions of the output parameter

Table 4 Properties of the Gaussian functions shown in Fig. 8

Membership function of the cluster	$\bar{x}$	σ
Bad quality	1.42	0.648
Poor quality 1	2.5	0.5308
Poor quality 2	3.5	0.5308
Fair quality 1	4.5	0.5308
Fair quality 2	5.5	0.5308
Good quality 1	6.5	0.5308
Good quality 2	7.5	0.5308
Excellent quality	8.44	0.648

the conjunctive or disjunctive system of rules and employing a defuzzification method that calculates the model output (e.g., max membership principle, centroid method, weighted average method, center of sums or other).

6.1 A set of fuzzy rules of the model

The inference system of the model is based on a set of 24 fuzzy rules that are listed in Table 5. The linguistic values of the input and the output parameters are defined earlier, in

Table 5 Set of fuzzy rules of the model

	First input parameter: PLR		Second input parameter: number of PLOs		Third input parameter: total duration of PLOs		Output: QoE
IF	Imperceptible quality distortion	AND	Negligible frequency	AND	Negligible duration	THEN	Excellent quality
IF	Imperceptible quality distortion	AND	Slightly annoying frequency	AND	Negligible duration	THEN	Good quality 2
IF	Imperceptible quality distortion	AND	Very annoying frequency	AND	Negligible duration	THEN	Good quality 2
IF	Imperceptible quality distortion	AND	Negligible frequency	AND	Slightly annoying duration	THEN	Good quality 2
IF	Imperceptible quality distortion	AND	Slightly annoying frequency	AND	Slightly annoying duration	THEN	Good quality 2
IF	Imperceptible quality distortion	AND	Very annoying frequency	AND	Slightly annoying duration	THEN	Good quality 2
IF	Imperceptible quality distortion	AND	Slightly annoying frequency	AND	Very annoying duration	THEN	Good quality 2
IF	Imperceptible quality distortion	AND	Very annoying frequency	AND	Very annoying duration	THEN	Good quality 2
IF	Slightly annoying quality distortion	AND	Negligible frequency	AND	Negligible duration	THEN	Excellent quality
IF	Slightly annoying quality distortion	AND	Slightly annoying frequency	AND	Negligible duration	THEN	Excellent quality
IF	Slightly annoying quality distortion	AND	Very annoying frequency	AND	Negligible duration	THEN	Good quality 2
IF	Slightly annoying quality distortion	AND	Negligible frequency	AND	Slightly annoying duration	THEN	Good quality 1
IF	Slightly annoying quality distortion	AND	Slightly annoying frequency	AND	Slightly annoying duration	THEN	Good quality 1
IF	Slightly annoying quality distortion	AND	Very annoying frequency	AND	Slightly annoying duration	THEN	Good quality 1
IF	Slightly annoying quality distortion	AND	Slightly annoying frequency	AND	Very annoying duration	THEN	Good quality 1
IF	Slightly annoying quality distortion	AND	Very annoying frequency	AND	Very annoying duration	THEN	Fair quality 1
IF	Very annoying quality distortion	AND	Negligible frequency	AND	Negligible duration	THEN	Good quality 2
IF	Very annoying quality distortion	AND	Slightly annoying frequency	AND	Negligible duration	THEN	Good quality 1
IF	Very annoying quality distortion	AND	Very annoying frequency	AND	Negligible duration	THEN	Fair quality 2
IF	Very annoying quality distortion	AND	Negligible frequency	AND	Slightly annoying duration	THEN	Fair quality 2
IF	Very annoying quality distortion	AND	Slightly annoying frequency	AND	Slightly annoying duration	THEN	Fair quality 2
IF	Very annoying quality distortion	AND	Very annoying frequency	AND	Slightly annoying duration	THEN	Fair quality 1
IF	Very annoying quality distortion	AND	Slightly annoying frequency	AND	Very annoying duration	THEN	Poor quality 2
IF	Very annoying quality distortion	AND	Very annoying frequency	AND	Very annoying duration	THEN	Poor quality 2

Sects. 5.1 and 5.2, respectively. The values are linked using the IF, AND THEN logical operators.

The rules are defined after several iterations, where we experimented with different settings of the model. The number of fuzzy clusters for the input and the output parameters was changed multiple times, and different properties of the fuzzy membership functions were tested ($\bar{x}$ and σ). This was done for the purpose of making the model more responsive to the changes in the input parameters, thus achieving better correlation of the model output with the MOS of the users. The conclusions of each rule were also changed and tested multiple times for the same purpose.

While observing this set of fuzzy rules, it is worth remembering how the methodology used for the subjective evaluation of the sequences impacted the results discussed in Sect. 2.3. The test was taken by the subjects in an uncontrolled environment, longer test sequences were used, the subjects were unaware of the purpose of the test, and they were mostly entertained by the content as well as the video containing subtitles. The test conditions made the PLOs harder to notice/memorize and also raised the subjects' tolerance toward the distortions. Thus, the MOS remain reasonably high for all test sequences, and this is here reflected in the rules. For instance, when the last rule listed in Table 5 is activated (very annoying quality distortions of all three input parameters), this leads to the consequence QoE = poor quality 2 (the $\bar{x}$ for this set equals 3.5). Thus, even the most degraded sequences are not rated rigorously.

6.2 The model output

The output of the model, the QoE rating, is calculated in MATLAB using methods that are thoroughly discussed by Ross in [54]. The mathematical descriptions of these methods (Eqs. 5, 6, and 7) are based on [54] and can be found below.

a. As is seen from Table 5, the most commonly used Mamdani inference system was implemented in the model. Specifically, the output y^k is a set of r propositions:

IF x_1 is A_1^k AND x_2 is A_2^k AND x_3
   is A_3^k THEN y^k is B^k for $k = 1, 2, \ldots, r,$ (5)

where x_1, x_2 and x_3 are the inputs, A_1^k, A_2^k and A_3^k are the fuzzy sets representing the k-th input triplets and B^k is the fuzzy set representing the k-th output.

b. The model is based on the most commonly used disjunctive system of rules. This implies that the output y is expressed by the fuzzy union of all individual rule contributions y^i, where $i = 1, 2, \ldots r$ and r is the number of IF–THEN propositions, as:

$$y = y^1 \cup y^2 \cup \cdots \cup y^r \tag{6}$$

c. The centeroid method was used for defuzzification to the output of the model. This method returns the center of an area under the curve and can be described with

$$y^* = \frac{\int u(y) \cdot y \, dy}{\int u(y) \, dy}, \tag{7}$$

where y^* is the defuzzified value and $u(y)$ is the curve describing the fuzzy union derived from Eq. 6. Note that we experimented with different defuzzification methods, and the most accurate results were obtained with the centeroid method.

7 Results of the model

The subjects rated 72 different test sequences in a home environment. The properties of each test sequence were previously presented in Table 1, and they include the following objective parameters: (a) PLR; (b) the number of PLOs in a sequence; and (c) total duration of PLOs. The values of these three parameters are plotted in Fig. 9, where each marker represents one test sequence, while the shape and the color of the marker represents the value of the assessed QoE by the model (according to the attached legend). While previously discussing the results presented in Table 1, we have mentioned how the MOS of the test subjects varied on the interval [4.16, 8.96], staying within the fair quality set even for the most degraded test sequence. The assessed QoE rating of the model varies between [4.48, 8.74]. This motivates us to investigate in the future to what extent a 1-h video has to be degraded to evoke higher dissatisfaction from the subjects.

The output of the model was tested for each test sequence by comparing it with the corresponding MOS of the subjects. Results of this analysis are presented in Fig. 10. The Pearson's correlation coefficient for this set of data equals 0.8841, indicating a strong positive linear relationship between the two variables.

To allow comparison between our model and similar models, Table 6 tabulates the properties of the models developed by other authors and the achieved correlation coefficients. The comparison includes only related work in which the effect of packet loss-related issues on user perception is investigated. In addition, since we used UDP on the transport layer, the comparison excludes models developed for assessing the QoE of HTTP-based video streaming. Note that the models developed by the authors referenced in Table 6 are all based on the controlled environment experiments (using only short test sequences), where the objective was to eliminate different QoE influential factors, which were the essential part of our study.

It could be argued that the models, which originate from the uncontrolled tests, produce outputs that are better suited for remote evaluation of service quality if the service is used in everyday, lifelike scenarios. This is due to the already mentioned fact that the subjective evaluation of service quality, when conducted in the uncontrolled environment, produces different results if compared with the results of the controlled environment experimenting. Considering these differences, it is not unexpected to find that certain authors question the usability of laboratory testing in controlled conditions (e.g., see the work presented in [27, 56]). Thus, the objective assessment models that are based solely on the subjective test results obtained from a controlled environment experiment have this mismatch integrated in them.

8 Conclusions and outlook

In this study, we have used the UDP-based multimedia streaming service to demonstrate how the objective and the subjective data sets can be correlated and the obtained fuzzy dependencies used for development of a no reference objective video quality assessment model for assessing user QoE. The primary use of the developed model is for the assessment of QoE for UDP-based IPTV services.

We are aware of certain usability limitations of the model. Namely, a subjective evaluation was conducted among the student population; hence, we cannot claim that the model output would have strong correlation with MOS of a more versatile panel of test subjects. Furthermore, only one type of video content was used for the subjective evaluation. Testing with other types of content of various duration may evoke different levels of user dissatisfaction for the same values of the objective parameters which we have tested.
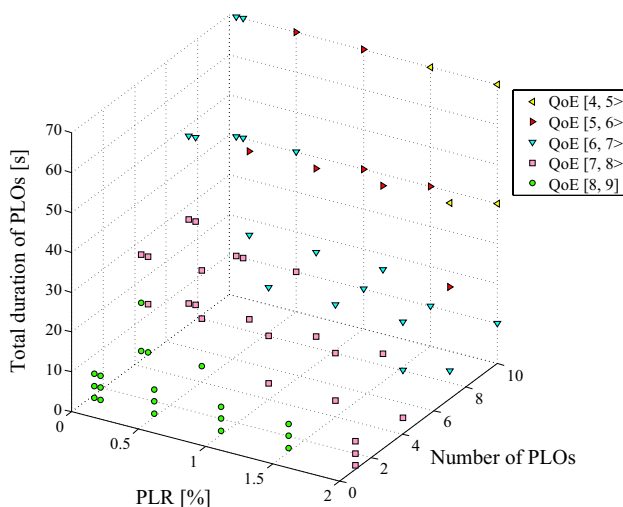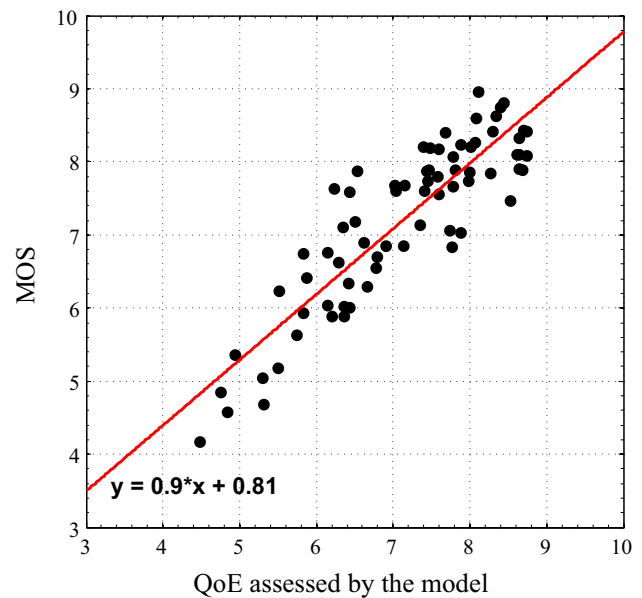
Fig. 10 Assessed QoE compared with the MOS of the subjects for each of the 72 test sequences

Hence, the inference system of our model would have to be modified with a purpose of making it applicable for the assessment of user QoE for other types of content of various duration. We also cannot ignore the impact of certain properties of the video, used in the study, on the test results. As emphasized, the video lasted 1 h and it contained subtitles. Thus, the recency effect affected the subjects' perception and some quality degradations became harder to notice by the subjects, respectively. This was reflected on the obtained results and the inference system of the model. Finally, the model cannot assess user QoE if PLR, number of PLOs and their total duration exceeds 2%, 10 and 70 s, respectively. These limitations of the model opened multiple paths for our future research and the opportunity to improve the model.

The results of the fuzzification process showed how a crisp value of one input parameter cannot be unambiguously related to the specific user rating; overlapping of the fuzzy clusters of the same input parameter exists as well as the fact that all three objective input parameters create an affiliated effect on user QoE. The latter proved that it was meaningful to include the three chosen objective parameters as inputs of the model.

It must be emphasized that the developed inference system of the model is predominantly influenced by the methodology used for the subjective evaluation of the sequences. In our study format, the subjects were unaware of the purpose of the test; they watched and evaluated 1-h video at their homes, in a familiar environment, with or without company at any time of the day (depending on their liking). These test conditions shifted the subjects' focus away from noticing and memorizing the video

Fig. 9 Assessed QoE rating by the model for each test sequence

Table 6 Comparison of different properties of other QoE assessment models

Reference	Number of test subjects	Number of test sequences	Number of input parameters	Mechanism of the inference system	FR, RR or NR	Pearson correlation coefficient
Chan et al. [31]	21	30 video clips	Two models were developed. The first model is based on averaged PSNR values of all distorted frames. The second model is based on the ratio of distorted frame rate and averaged PSNR of distorted frames	Linear regression	FR	0.8664 and 0.935 (for two types of the PSNR modification)
Leszczuk et al. [33]	24	59 video clips	Three parameters (SSIM index averaged over the worst second, number of separate losses, spatial activity)	Multiple linear regression analysis	FR	0.933
da Silva Cruz et al. [37]	35	2962 video clips, each lasting 10 s, used to train artificial neural network (first stage of the cascaded estimator) 20 video clips, each lasting 10 s, used for subjective evaluation (second stage of the cascaded estimator)	Three parameters (PLR, frame degradation ratio and information loss ratio)	Two-stage cascaded estimator consisted of the artificial neural network (ANN) used to estimate SSIM Index and Logistic model (ITU-R BT.500 − 13) used to estimate differential MOS	NR	0.894
Venkataraman et al. [43]; Venkataraman and Chatterjee [44]	77	18 video clips	Three parameters (PLR, delay and encoding bit rate)	Offline construction of a QoE (k-dimensional Euclidean) space	NR	0.9964
Robalo and Velez [45]	Uses the results from [44]	Uses the results from [44]	Uses the parameters from [44]	Sixth-degree polynomial equation	NR	0.9154
Anegekuh et al. [47]	97	6 video clips	Three parameters (content type metric combined with quantization parameter and PLR)	Combination of encoding quality prediction (depends on quantization parameter and content type in relation to MOS) and impact of PLR on the perceived quality Model coefficients are obtained through regression fittings	FR	0.92

Table 6 (continued)

Reference	Number of test subjects	Number of test sequences	Number of input parameters	Mechanism of the inference system	FR, RR or NR	Pearson correlation coefficient
Konuk et al. [48]	Not available. MOS values were taken from different video databases	Not reported	Two parameters (MOS for video and MOS for audio)	Linear combination of video and audio quality and their product. Model also includes video spatio-temporal characteristics	NR	0.8132 and 0.8443 for subjective and objective video MOS, respectively
Hameed et al. [51]	100	288 video clips	Seven parameters extracted from a video stream (the parameters describe video content characteristic, encoding distortions and packet loss effects)	A decision tree	NR	Not available. The accuracy of the model is calculated by comparing the model output (poor, bad, fair, good, very good or excellent quality) with the subjects rating which was categorized in the same groups. The model evaluated 128 out of 144 videos correctly (88.9%)

quality distortions to the actual content of the video. We believe that these test conditions provide more lifelike evaluation results of user QoE.

The subjective evaluation of user QoE that was conducted in uncontrolled test conditions represents the origin of one of the models most distinctive features; we showed that only a few such assessment models exist today. Usually, other authors develop similar models from the results of the controlled experiment. Due to the disparity between the results obtained in the controlled and uncontrolled test environments, it could be argued that, in the context of everyday service usage, the models based on the uncontrolled experiments assess user QoE with more accuracy compared to metrics and assessment models that are based solely on laboratory test results.

Our future research will strive toward developing a similar QoE assessment model that will be able to make assessments based on the video frame rate, bitrate and re-buffering frequency, possibly for 4K video streaming. Again, we plan to build the model from the results of uncontrolled experimenting. Thus, the new model will enable more lifelike assessments of user QoE compared to currently available models, which are mainly based on laboratory test results. However, considering the abovementioned limitations of the current model, this new subjective evaluation of video quality will test a more diverse group of test subjects (e.g., by employing QoE crowdtesting), who will evaluate different types of content (e.g., music videos or TV shows). Additionally, learning from the research of other authors presented in this paper (for instance, from those listed in Table 6), the new model could also combine different input parameters with the results obtained from known metrics for objective evaluation of video quality such as PSNR and SSIM. The model would then become an FR model.

We would also like to continue our investigation of how video subtitles affect user QoE. To the best of our knowledge, this particular field of research is still insufficiently explored. For this purpose, we will use eye-tracking glasses and inspect the eye movements of the subjects' when they watch the video with time-varying quality and subtitles. Then, we will be able to correlate the subjects' perception of video quality, their eye movements, the amount of text on the screen and specific quality distortions.

Appendix 1: the questionnaire used in the study

In this appendix, we present pages 2 and 3 of the questionnaire used in the study. As has been stated, page 1 contained only the instructions on how to complete the questionnaire, whereas page 4 contained several general questions regarding subject demographic information and a blank space, where the subjects could leave comments. Thus, pages 1 and 4 are not included in this appendix.

(A) THE PERCEIVED VIDEO QUALITY

A1. Mark on the scale your opinion of the audiovisual quality of the video that you have just finished watching:

0	1	2	3	4	5	6	7	8	9	10
Bad quality		Poor quality		Fair quality		Good quality		Excellent quality		

A2. While watching the video, I noticed that the video quality was degraded on one or multiple occasions.

Is this statement true?
a) Yes.   b) No.

A3. If you answered the previous question with **YES**, then please proceed to the next questions (**A3.1, A3.2, A3.3, A3.4 and A3.5**). If you answered the previous question with **NO**, then please skip to **Section B of the questionnaire**.

A3.1 What types of video quality degradation did you notice?

(mark your answer(s); multiple answers are possible)

a) The video appeared choppy (i.e., the reproduction was not "smooth").
b) The video was delayed in relation to the audio (synchronization issues).
c) The video image was incomplete (parts of the picture were not shown).
d) The video froze (the reproduction stopped).
e) Some parts of the video image appeared as if they were assembled from blocks.
f) The video image was split into several sections and it was clear that some sections were not a part of the current video image.

g) The video image appeared to be "broken" in some parts of the screen.
h) The video image contained colored blocks. It was clear that these blocks were not a part of the video image.
i) The audio was choppy.
j) The audio was delayed in relation to the video (synchronization issues).
k) The audio was incomplete (parts of the audio were not reproduced).
l) The audio reproduction stopped.
m) The reproduction of the entire content of the video stopped and then restarted after some amount of time.

If you experienced something that cannot be described by any of these answers, then please write what you experienced below:

A3.2 When you reflect back on the quality distortions that you experienced during the screening, you would say that they were:

0	1	2	3	4	5	6	7	8	9	10
Imperceptible quality distortions		Perceptible quality distortion, but not annoying		Slightly annoying quality distortion		Annoying quality distortion		Very annoying quality distortion		

A3.3 You noticed that distortions appeared in the video approximately _____ times *(write a number)*. You think that this was a:

0	1	2	3	4	5	6	7	8	9	10
Negligible frequency		Low frequency, but not annoying		Medium frequency, slightly annoying		Annoyingly high frequency		Very annoying frequency		

A3.4 If you were to quantify the total amount of time for which the quality distortions appeared on the screen, that time would be equal to _____ seconds *(write a number)*. You think that this was a:

0	1	2	3	4	5	6	7	8	9	10
Negligible duration		Short duration, but not annoying		Medium duration, slightly annoying		Annoyingly long duration		Very annoying duration		

A3.5 Considering the types of degradations that you noticed, their appearing frequency and total duration, how do you evaluate your experience of watching this video?

0	1	2	3	4	5	6	7	8	9	10
Bad experience		Poor experience		Fair experience		Good experience		Excellent experience		

| **(B)** | ABOUT THE CONTENT, USER ENVIRONMENT, SOCIAL CONTEXT AND OTHER |

B1. Mark on the scale how entertaining the video was to you.

0	1	2	3	4	5	6	7	8	9	10
Least entertaining, boring		Partially entertaining		Entertaining			Mostly entertaining		Very entertaining	

B2. Did you watch the entire video without interruptions?

a) I watched the entire video without interruptions.

b) I watched the entire video, but with interruptions. No. of interruptions: _____ .
(write a number)

c) I did not watch the entire video.

B3. How much noise was there in your surroundings while you were watching the video?

a) There was no noise whatsoever.

b) There was some noise, but not enough to distract me from watching.

c) It was a bit noisy, enough to disrupt my concentration for a short period of time.

d) There was a lot of noise, so much that I was unable to concentrate on the video for long time periods.

B4. Mark the statements that are applicable to you. If there aren't any or you are unsure, proceed to the next question.

(mark your answer(s); multiple answers are possible)

a) When I watch DVDs as I usually do, their quality is often degraded.

b) There is a possibility that my DVD player that I used to watch this video may be broken or malfunctioning.

c) I watched the video on a screen with a 4:3 aspect ratio.

d) I watched the video on a screen with a 16:9 aspect ratio.

e) My screen supports the HD format (Full HD or HD ready).

f) I watched the video on a CRT screen.

B5. What was the social context in which you watched the video?

a) I watched the video alone.

b) I watched the video in the company of _____ persons. (*write a number*)

B6. If you answered the previous question with b), did that person(s) suggest to you in any way that the quality of the video was degraded?

a) No, I noticed on my own that the quality was degraded.

b) Yes, without the person(s) in my company, I would not have noticed the quality degradations in the video.

c) No one noticed any quality degradations.

B7. What do you think about the video subtitles?

(mark your answer(s); multiple answers are possible)

a) Without the subtitles, I would not have understood the content.

b) They were useful, but I would be able to watch the video without them.

c) The subtitles were only distracting me.

d) The quality of the subtitles was good.

e) The quality of the subtitles was poor.

f) Instead of the subtitles, I would prefer a Croatian narrator.

B8. Do you see and hear well?

a) Yes, I see and hear well (either with or without visual and hearing aids).

b) I have impaired hearing.

c) I have impaired sight.

d) I have impaired sight and hearing.

B9. Where you tired while watching the video?

a) Yes, I was tired.

b) No, I was rested.

B10. Did you complete this questionnaire immediately after watching the video?

a) Yes.

b) No.

B11. Were you familiar with the topic of this research prior to watching the video?

a) Yes.

b) No.

Appendix 2: the fuzzy membership functions

The following figures depict the fuzzy membership functions of the three input parameters. For every parameter there are three subplots (one for each fuzzy cluster). The y axes depict the degree of membership, but the ticks are shown only in the first subplots (a). The degrees of membership of the subjects' responses are presented with gray bars (Figs. 11, 12, 13).
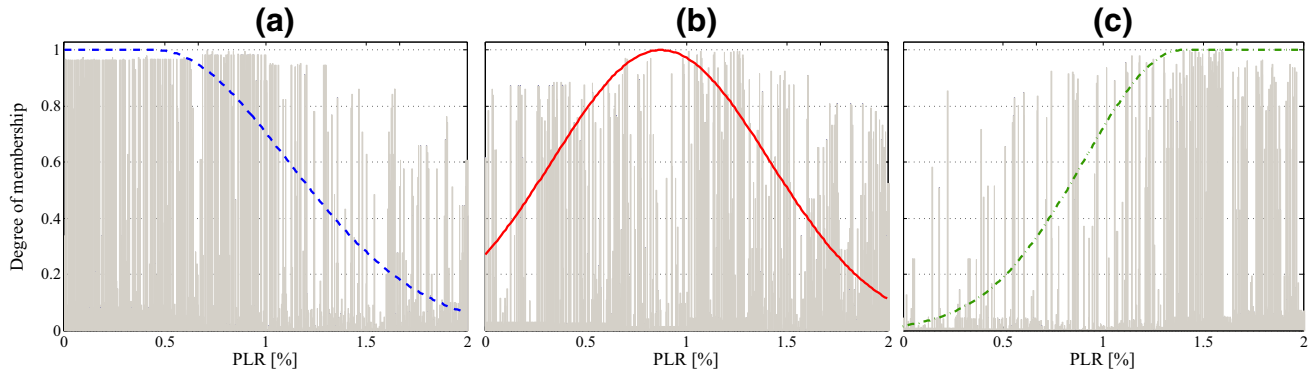

Fig. 11 Membership functions for the PLR: **a** cluster 1: imperceptible quality distortion, **b** cluster 2: slightly annoying quality distortion, and **c** cluster 3: very annoying quality distortion
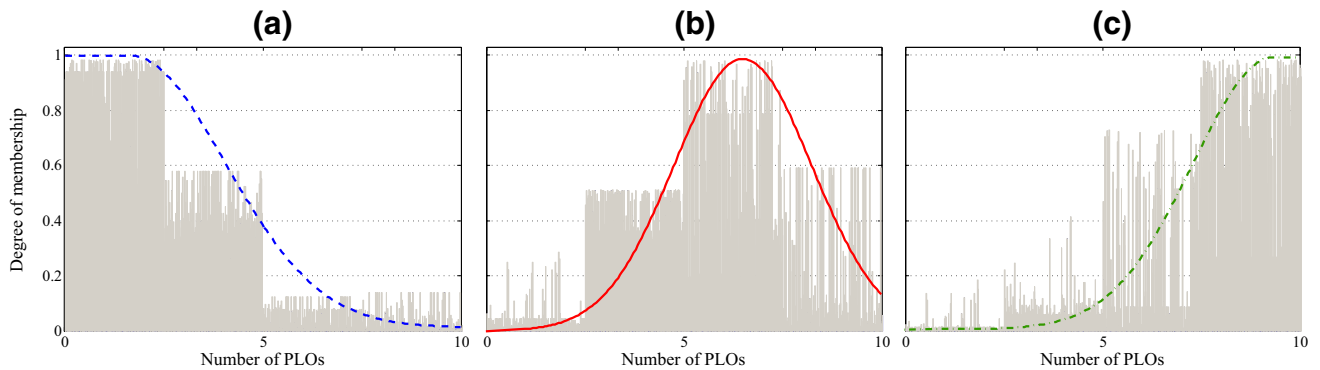

Fig. 12 Membership functions for the number of PLOs: **a** cluster 1: negligible frequency, **b** cluster 2: slightly annoying frequency, and **c** cluster 3: very annoying frequency
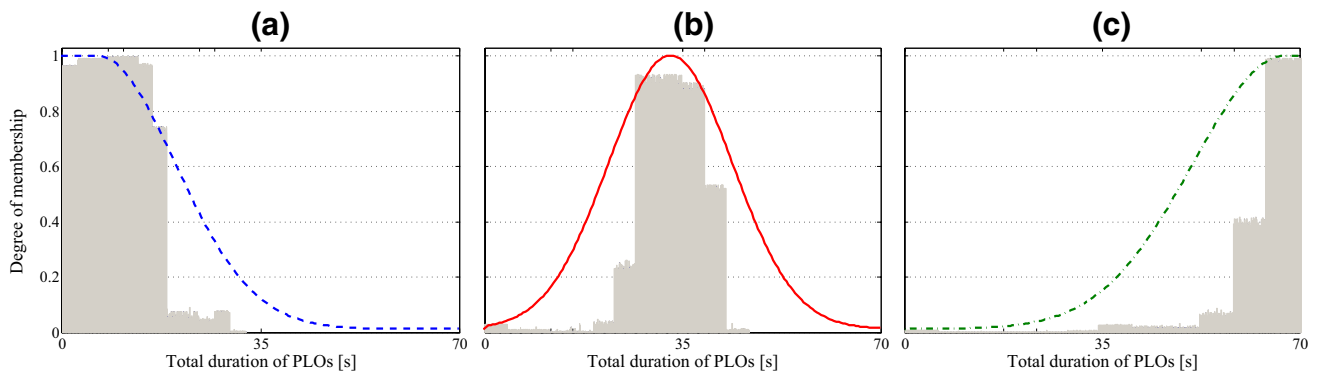

Fig. 13 Membership functions for the total duration of PLOs: **a** cluster 1: negligible duration, **b** cluster 2: slightly annoying duration, and **c** cluster 3: very annoying duration

References

1. Ahmad, A., Floris, A., Atzori, L.: QoE-centric service delivery: a collaborative approach among OTTs and ISPs. Comput. Netw. **110**, 168–179 (2016). https://doi.org/10.1016/j.comnet.2016.09.022

2. Nogueira, J., Guardalben, L., Cardoso, B., Sargento, S.: Catch-up TV analytics: statistical characterization and consumption patterns identification on a production service. Multimedia Syst. (2016). https://doi.org/10.1007/s00530-016-0516-7

3. Le Callet, P., Möller, S., Perkis, A.: Qualinet white paper on definitions of quality of experience. White paper COST Action IC 1003. http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf (2013). Accessed 8 Feb 2017

4. Mrvelj, Š, Matulin, M.: Impact of packet loss on the perceived quality of UDP-based multimedia streaming: a study of user quality of experience in real-life environments. Multimedia Syst. (2016). https://doi.org/10.1007/s00530-016-0531-8

5. Staelens, N., Moens, S., Van den Broeck, W., Mariën, I., Vermeulen, B., Lambert, P., Van de Walle, R., Demeester, P.: Assessing quality of experience of IPTV and video on demand services in real-life environments. IEEE Trans. Broadcast. **56**, 458–466 (2010). https://doi.org/10.1109/TBC.2010.2067710

6. Afshari, S., Movahhedinia, N.: QoE assessment of interactive applications in computer networks. Multimedia Tools Appl. **75**(2), 903–918 (2016). https://doi.org/10.1007/s11042-014-2331-5

7. Xu, Q., Huang, Q., Yao, Y.: Online crowdsourcing subjective image quality assessment. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 359–368. New York (2012). https://doi.org/10.1145/2393347.2393400

8. Gardlo, B., Ries, M., Hossfeld, T., Schatz, R.: Microworkers vs. Facebook: the impact of crowdsourcing platform choice on experimental results. In: Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 35–36. Yarra Valley, Australia (2012). https://doi.org/10.1109/QoMEX.2012.6263885

9. Rainer, B., Timmerer, C.: A quality of experience model for adaptive media playout. In: Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 177–182. Singapore (2014). https://doi.org/10.1109/QoMEX.2014.6982316

10. Ickin, S., Wac, K., Fiedler, M., Janowski, L., Hong, J., Dey, A.: Factors influencing quality of experience of commonly used mobile applications. IEEE Commun. Mag. **50**, 48–56 (2012). https://doi.org/10.1109/MCOM.2012.6178833

11. Sladojevic, S., Culibrk, D., Mirkovic, M., Coll, D., Borba, G.: Logging real packet reception patterns for end-to-end quality of experience assessment in wireless multimedia transmission. In: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6. San Jose, California (2013). https://doi.org/10.1109/ICMEW.2013.6618453

12. Staelens, N., De Meulenaere, J., Claeys, M., Van Wallendael, G., Van den Broeck, W., De Cock, J., Van de Walle, R., Demeester, P., De Turck, F.: Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices. IEEE Trans. Broadcast. **60**, 707–714 (2014). https://doi.org/10.1109/TBC.2014.2359255

13. Ickin, S., Fiedler, M., Wac, K., Arlos, P., Temiz, C., Mkocha, K.: VLQoE: video QoE instrumentation on the smartphone. Multimedia Tools Appl. **74**(2), 381–411 (2015). https://doi.org/10.1007/s11042-014-1919-0

14. Pinson, M.H., Boyd, K.S., Hooker, J., Muntean, K.: How to choose video sequences for video quality assessment. In: Proceedings of the 7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics, pp. 79–85. Scottsdale, Arizona (2013). https://events.engineering.asu.edu/vpqm/vpqm13/Papers/Paper4.pdf. Accessed 27 Aug 2016

15. Frohlich, P., Egger, S., Schatz, R., Muhlegger, M., Masuch, K., Gardlo, B.: QoE in 10 s: are short video clip lengths sufficient for quality of experience assessment? In: Proceedings of the 4th International Workshop on Quality of Multimedia Experience, pp. 242–247. Yarra Valley, Australia (2012). https://doi.org/10.1109/QoMEX.2012.6263851

16. Li, W., Rehman, H., Kaya, D., Chignell, M., Leon-Garcia, A., Zucherman, L., Jiang, J.: Video quality of experience in the presence of accessibility and retainability failures. In: Proceedings of the 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, pp. 1–7. Rhodes, Greece (2014). https://doi.org/10.1109/QSHINE.2014.6928651

17. Tavakoli, S., Brunnström, K., Gutiérrez, J., García, N.: Quality of experience of adaptive video streaming: investigation in service parameters and subjective quality assessment methodology. Signal Process. Image Commun. **39**(Part B), 432–443 (2015). https://doi.org/10.1016/j.image.2015.05.001

18. Hands, D., Avons, S.: Recency and duration neglect in subjective assessment of television picture quality. Appl. Cogn. Psychol. **15**, 639–657 (2001). https://doi.org/10.1002/acp.731

19. Jelassi, S., Rubino, G., Melvin, H., Youssef, H., Pujolle, G.: Quality of experience of VoIP service: a survey of assessment approaches and open issues. IEEE Commun. Surv. Tutor. **14**(2), 491–513 (2012). https://doi.org/10.1109/SURV.2011.120811.00063

20. Shen, Y., Liu, Y., Liu, Q., Yang, D.: A method of qoe evaluation for adaptive streaming based on bitrate distribution. In: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops, pp. 551–556. San Jose, California: (2014). https://doi.org/10.1109/ICMEW.2013.6618453

21. Rahrer, T., Fiandra, R., Wright, S.: Triple-play services Quality of Experience (QoE) requirements. DSL Forum WT-126 v.0.5, Digital Subscriber Line Forum. https://www.broadband-forum.org/technical/download/TR-126.pdf (2006). Accessed 3 Mar 2017

22. International Telecommunication Union: Subjective video quality assessment methods for multimedia applications. International Telecommunication Union (ITU-T Rec. P.910). https://www.itu.int/rec/T-REC-P.910-200804-I/en (2008). Accessed 13 May 2016

23. Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans. Multimedia. **16**(2), 541–558 (2014). https://doi.org/10.1109/TMM.2013.2291663

24. Farrokhi, F., Mahmoudi-Hamidabad, A.: Rethinking convenience sampling: defining quality criteria. Theory Pract. Lang. Stud. **2**(4), 784–792 (2012). https://doi.org/10.4304/tpls.2.4.784-792

25. Datta, P., Izdebski, L., Kumar, N., Suh, K.: "It came to me in a stream… The upward arc of online video, driven by consumers. White paper, Cisco. http://www.cisco.com/c/dam/en_us/about/ac79/docs/sp/Online-Video-Consumption_Consumers.pdf (2012). Accessed 13 Nov 2016

26. Fiedler, M., Hoßfeld, T.: Quality of experience-related differential equations and provisioning-delivery hysteresis. In: Proceedings of the 21st ITC Specialist Seminar on Multimedia Applications-Traffic, Performance and QoE. Miyazaki, Japan (2010). https://www.diva-portal.org/smash/get/diva2:835271/FULLTEXT01.pdf. Accseesed 26 July 2016

27. Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T., Kankainen, A.: Usability testing of mobile applications: a comparison between laboratory and field testing. J. Usability Stud. 1(1), 4–16 (2005). http://uxpajournal.org/wp-content/uploads/pdf/mobile.pdf. Accseesed 8 July 2016

28. International Telecommunication Union: User requirements for objective perceptual video quality measurements in digital cable television. International Telecommunication Union (ITU-T Rec. J.143). http://www.itu.int/rec/T-REC-J.143-200005-I/ (2000). Accessed 13 May 2016

29. International Telecommunication Union: Reference guide to Quality of Experience assessment methodologies. International Telecommunication Union (ITU-T Rec. G.1011). https://www.itu.int/rec/T-REC-G.1011/en (2010). Accessed 13 May 2016

30. Alia, M., Lacoste, M.: An adaptive quality of experience framework for home network services. In: Proceedings of the 3rd International Conference on Communication Theory, Reliability, and Quality of Service, pp. 226–232. Athens, Greece (2010). https://doi.org/10.1109/CTRQ.2010.45

31. Chan, A., Zeng, K., Mohapatra, P., Sung-Ju, L., Banerjee, S.: Metrics for evaluating video streaming quality in lossy IEEE 802.11 wireless networks. In: Proceedings of the IEEE INFOCOM, pp. 1–9. San Diego, California (2010). https://doi.org/10.1109/INFOCOM.2010.5461979

32. Hu, S., Jin, L., Kuo, C.-C.: Compressed video quality assessment with modified MSE. In: Proceedings of the Signal and Information Processing Association Annual Summit and Conference, pp. 1–4. Siem Reap, Cambodia (2014). https://doi.org/10.1109/APSIPA.2014.7041643

33. Leszczuk, M., Janowski, L., Romaniak, P., Papir, Z.: Assessing quality of experience for high definition video streaming under diverse packet loss patterns. Signal Process. Image Commun. 28(8), 903–916 (2013). https://doi.org/10.1016/j.image.2012.09.006

34. Mohamed, S., Rubino, G.: A study of real-time packet video quality using random neural networks. IEEE Trans. Circuits Syst. Video Technol. 12(12), 1071–1083 (2002). https://doi.org/10.1109/TCSVT.2002.806808

35. Wang, R., Geng, Y., Ding, Y., Yang, Y., Li, W.: Assessing the quality of experience of HTTP video streaming considering the effects of pause position. In: Proceedings of the 16th Asia-Pacific Network Operations and Management Symposium, pp. 1–4. Hsinchu, China (2014). https://doi.org/10.1109/APNOMS.2014.6996568

36. Singh, K., Hadjadj-Aoul, Y., Rubino, G.: Quality of experience estimation for adaptive HTTP/TCP video streaming using H. 264/AVC. In: Proceedings of the IEEE Consumer Communications and Networking Conference, pp. 127–131. Las Vegas, Nevada (2012). https://doi.org/10.1109/CCNC.2012.6181070

37. da Silva Cruz, L.A., Cordina, M., Debono, C.J., Amado Assunção, P.A.: Quality monitor for 3-d video over hybrid broadcast networks. IEEE Trans. Broadcast. 62(4), 785–799 (2016). https://doi.org/10.1109/TBC.2016.2617278

38. Mok, R., Chan, E., Chang, R.: Measuring the Quality of Experience of HTTP video streaming. In: Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops, pp. 485–492. Dublin, Ireland (2011). https://doi.org/10.1109/INM.2011.5990550

39. Seufert, M., Egger, S., Slanina, M., Zinner, T., Hoßfeld, T., Tran-Gia, P.: A survey on quality of experience of HTTP adaptive streaming. IEEE Commun. Surv. Tutor. 17(1), 469–492 (2015). https://doi.org/10.1109/COMST.2014.2360940

40. Menkovski, V., Oredope, A., Liotta, A., Cuadra Sánchez, A.: Predicting quality of experience in multimedia streaming. In: Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia, pp. 52–59. New York (2009). https://doi.org/10.1145/1821748.1821766

41. Menkovski, V., Exarchakos, G., Liotta, A., Cuadra Sánchez, A.: Estimations and remedies for quality of experience in multimedia streaming. In: Proceedings of the 3rd International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services, pp. 11–15. Nice, Italy (2010). https://doi.org/10.1109/CENTRIC.2010.14

42. de Pessemier, T., de Moor, K., Joseph, W., de Marez, L., Martens, L.: Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching. IEEE Trans. Broadcast. 59(1), 47–61 (2013). https://doi.org/10.1109/TBC.2012.2220231

43. Venkataraman, M., Chatterjee, M., Chattopadhyay, S.: Evaluating quality of experience for streaming video in real time. In: Proceedings of the Global Telecommunications Conference, pp. 1–6. Honolulu, Hawaii (2009). https://doi.org/10.1109/GLOCOM.2009.5425806

44. Venkataraman, M., Chatterjee, M.: Inferring video QoE in real time. IEEE Netw. 25(1), 4–13 (2011). https://doi.org/10.1109/MNET.2011.5687947

45. Robalo, D., Velez, F.: A model for mapping between the Quality of Service and experience for wireless multimedia applications. In: Proceedings of the IEEE 79th Vehicular Technology Conference (VTC Spring), pp. 1550–2252. Seoul, South Korea (2014). https://doi.org/10.1109/VTCSpring.2014.7022781

46. Nguyen, L., Harris, R., Punchihewa, A., Jusak, J.: Application of a mixed effects model in predicting quality of experience in world wide web services. In: Proceedings of the 4th International Conference on Computational Intelligence, Modelling and Simulation, pp. 316–321. Kuantan, Malaysia (2012). https://doi.org/10.1109/CIMSim.2012.26

47. Anegekuh, L., Sun, L., Jammeh, E., Mkwawa, I., Ifeachor, E.: Content-based video quality prediction for HEVC encoded videos streamed over packet networks. IEEE Trans. Multimedia. 17(8), 1323–1334 (2015). https://doi.org/10.1109/TMM.2015.2444098

48. Konuk, B., Zerman, E., Nur, G., Bozdagi Akar, G.: Video content analysis method for audiovisual quality assessment. In: Proceedings of the Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6. Lisbon, Portugal (2016). https://doi.org/10.1109/QoMEX.2016.7498965

49. Rehman, A., Wang, Z.: Perceptual experience of time-varying video quality. In: Proceedings of the 5th International Workshop on Quality of Multimedia Experience, pp. 218–223. Klagenfurt, Austria (2013). https://doi.org/10.1109/QoMEX.2013.6603240

50. Zhang, Y., Yue, T., Wang, H., Wei, A.: Predicting the quality of experience for internet video with fuzzy decision tree. In: Proceedings of the IEEE 17th International Conference on Computational Science and Engineering, pp. 1181–1187. Chengdu, China (2014). https://doi.org/10.1109/CSE.2014.230

51. Hameed, A., Dai, R., Balas, B.: A decision-tree-based perceptual video quality prediction model and its application in FEC for wireless multimedia communications. IEEE Trans. Multimedia. 18(4), 764–774 (2016). https://doi.org/10.1109/TMM.2016.2525862

52. Fang, R., Wu, D., Shen, L.: Evaluation of image quality of experience in consideration of viewing distance. In: Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing, pp. 653–657. Chengdu, China (2015). https://doi.org/10.1109/ChinaSIP.2015.7230485

53. Hernando, D., López de Vergara, J., Madrigal, D., Mata, F.: Evaluating quality of experience in IPTV services using MPEG frame loss rate. In: Proceedings of the International Conference on Smart Communications in Network Technologies, pp. 1–5. Paris, France (2013). https://doi.org/10.1109/SaCoNeT.2013.6654554

54. Ross, T.: Fuzzy Logic with Engineering Applications. John Wiley, Chichester (2004). (ISBN: 978-0-470-86074-8)

55. Alata, M., Molhim, M., Ramini, A.: Optimizing of fuzzy C-means clustering algorithm using GA. World Acad. Sci. Eng. Technol. 2(3), 224–229 (2008). http://waset.org/publications/13398. Accessed 11 April 2017

56. Sun, X., May, A.: A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices. Adv. Hum. Comput. Interact. 2013, 1–10 (2013). https://doi.org/10.1155/2013/619767