Subjective Evaluation of Quality of Experience for Video Streaming Service

Marko Matulin, Štefica Mrvelj
Department of Information and Communications Traffic
Faculty of Transport and Traffic Sciences, University of Zagreb
Zagreb, Croatia
mmatulin@fpz.hr, smrvelj@fpz.hr

Abstract—**This paper presents the results of the subjective evaluation of Quality of Experience (QoE) for video streaming service. The evaluation was conducted among 602 test subjects who rated the quality of 72 videos by watching them at their homes. The quality of each video was uniquely distorted (packet loss rate, number of packet loss occurrences during video streaming and total duration of those occurrences varied in each video). The results revealed that, in real-life environment, test subjects failed to notice significant number of video segments whose quality was degraded, they could not accurately quantify their total duration and they usually did not maintain negative attitude about perceived quality distortions.**

Keywords-Quality of Experience, video streaming, subjective evaluation, real-life, packet loss rate

I. INTRODUCTION

The Quality of Experience (QoE) concept introduced profound changes in subjective and objective methods used for evaluation of service quality. These changes originate from the fact that analysis of network performances may indicate service quality on the network level, but it cannot answer questions related to end user perceived service quality. This is accentuated by Bai et al. in [1] who have found that end users are not interested in how a network performs. Their only interest is the application level service quality. Since the user perception is tempered by many technical and non technical parameters, the QoE concept was developed to cope with this new and broader context of service quality evaluation.

The user gains the experience while using a service, thus it is meaningful to conduct the subjective tests of QoE in real-life environments [2]. One of the first real-life subjective tests of QoE was conducted by Reichl et al. in [3]. The service in the focus was the mobile multimedia streaming. The authors installed two cameras on a woman's hat. The woman was wearing the hat during her everyday routine, as well as when she was using the aforementioned application. The cameras were recording her facial expressions. Later, the stored video could be analyzed in order to determine the degree of the woman's enjoyment, frustration, boredom etc.

The same type of application was analyzed by Jumisko-Pyykkö et al. in [4] among two groups of test subjects. The subjects in the first group used the application on a train station, in a bus and in a coffee shop. The second group of test subjects viewed the same video clips under the same network conditions, but in controlled environment. During multimedia streaming sessions the quality of 60-second video clips was degraded by frame error rates of 1.7, 6.9, 13.8 and 20.7%. The length of the time interval when the errors occurred varied between 1, 4, 8 and 12 seconds. The results showed that the first group of users did not notice so many impairments as the second group.

Staelens et al. in [5] and [6] report on the results of the evaluation of QoE of full-length movies, whose quality was degraded in several segments by introducing packet loss and using video coded with lower bitrates. The test subjects were asked just to watch the movie, distributed to them on a DVD, as they would normally do (e.g. in the comfort of their home) and to evaluate its quality after watching by completing the questionnaire. Users were uninformed about the topic of the questionnaire prior to watching. Concurrently, the tests were conducted in the controlled environment with the second group of test subjects. The results indicated that the first group evaluated the quality of a DVD movie higher compared with the second group.

Although the network performances are not the only factor influencing user QoE, the quality of video content presentation highly depends on them. Hence, to understand better user ratings and to improve reasoning behind the obtained results, it is important to know how a network performed during streaming sessions. The video streaming service falls into the group of real-time services, so the real-life subjective testing of it imposes one important challenge: how to record network performances if test subjects are using the service e.g. from their homes?

Owing to that challenge, we streamed one hour video between two computers six times in emulated network environment, each time with different packet loss rate (PLR). Various sizes of PLR caused versatile video artifacts which degraded the quality of incoming video. The incoming videos were stored on a computer. In the next step several short video clips were taken from the stored video signal and inserted into the original video signal. This simulated temporal drop of network performances during one streaming session.

The diversity of quality distortions caused by different PLR, along with different number and total duration of PLR occurrences allowed the creation of 72 video signals, i.e. test sequences. These videos were distributed on a DVD to test subjects for evaluation. The subjects were asked to watch the video in the environment where they usually watch e.g. TV programme and to evaluate its quality by completing the

questionnaire. This paper reports on the obtained results of that research.

The paper is structured as follows: Chapter 2 briefly describes the process of creation of degraded video signals, as well as the method used for subjective evaluation; results of the evaluation are presented in Chapter 3 and discussed in Chapter 4; closing conclusions and future work are presented in Chapter 5.

II. RESEARCH METHOD

A. *Generating video signals of degraded quality*

Video content used in this research was one hour documentary film about the solar system. We avoided using short video clips, because Frohlich et al. in [7] found that, when using short video clips, the evaluation of QoE does not quite match the real-life quality perception, i.e. in real-life it is necessary to increase the duration of test sequences. The video was encoded using Advanced Video Coding (H.264/AVC) and Advanced Audio Coding (AAC). The video was coded at a bitrate of 9.8 Mbps and a frame rate of 50 fps. Resolution of the video was 1920x1080 pixels. The audio was coded at a bitrate of 256 kbps.

The video was streamed in emulated network environment between two computers. During streaming sessions a PLR of 0.05, 0.1, 0.5, 1, 1.5 and 2% was introduced using the emulator client (Network Emulator for Windows Toolkit). Six incoming video signals, each completely affected by different PLR and containing video artifacts, were stored in the same format as the original video. In the next phase 1, 4, 7 or 10 short video clips from one degraded video signal were taken and inserted into the original video. The duration of a single inserted clip varied between 1, 4 and 7 seconds. Different variations of these 3 parameters allowed us to create 72 different combinations of video signal whose quality had to be subjectively evaluated by test subjects. Total duration of all inserted video clips varied between 1, 4, 7, 10, 16, 28, 40, 49 and 70 seconds, depending on their number and a single duration.

The inserted clips of degraded video were evenly distributed over the entire duration of the test sequence as much as possible, because when the clips are grouped in the last few minutes of the screening the quality scores are lower. Conversely, if the clips are grouped in the first few minutes the quality scores are higher [8]. Even distribution of inserted clips allowed test subjects to immerse into the film in the beginning of the screening and to contemplate about what they have experienced in the end.

B. *Test subjects and response rate*

In this research test subjects were students of the Faculty of Transport and Traffic Sciences, University of Zagreb. Two main reasons why this population was targeted are: a) according to [9] video streaming service is mostly used by users between the ages of 18 and 24 which corresponds with the age group of a student population and b) simplicity of conducting such a survey (this population was easy accessible, i.e. convenience sampling method was used).

Without revealing the purpose of the research, 864 students received a sealed envelope (containing the questionnaire) and one DVD with the degraded video, i.e. test sequence whose quality had to be evaluated. The students were instructed not to open the envelope before the end of the screening and to fulfil the questionnaire immediately after the screening ends. Each student watched only one video once. The questionnaire contained multiple choice questions. However, for those questions related to the subjective perception of video quality and the extent of perceived video quality distortions, we used 11-grade numerical quality scale designed in [10].

To ensure minimum required sample size of 4 responses per one test sequence (according to [10]), 12 DVD copies of each video were made and two questionnaires were inserted into the envelopes (given the possibility that the subjects might watch the video in someone's company). That led to 830 collected questionnaires, but 228 questionnaires were rejected, i.e. the analysis of user QoE was conducted on a sample of 602 test subjects. For each type of video between 6 and 12 questionnaires were collected.

III. RESULTS

A. *Analysis of user QoE*

Average QoE scores and standard deviations (SD) for each of the 72 videos are presented in Tab. 1. It should be noted that the grades on the 11-grade numerical quality scale had the following linguistic meanings: 0-2 *Bad*, 2-4 *Poor*, 4-6 *Fair*, 6-8 *Good*, 8-10 *Excellent*. The boundaries between these five different sets are not firmly determined, because linguistic meanings are given as a help to the test subjects during rating. That feature makes the scale suitable for exploring user opinions which are usually fuzzy in nature.

The results showed that PLR and total duration of inserted clips have limited impact on user QoE in cases when the test sequences contained only one inserted clip (videos number 1 to 18 in Tab. 1). The average QoE scores for these videos varied between [7.46, 8.96], i.e. the quality of 4 videos was evaluated as *Good* while for other 14 videos it was *Excellent*. When the number of inserted clips in the test sequences increases to 4 the QoE score falls under 6 for the first time. However, this only happens when PLR reaches 2% and total duration of all inserted clips equals 28 seconds (video number 36). Similar results are obtained for the videos containing 7 inserted clips. Here the QoE score is for the first time lower than 5, but only for the video with the most quality distortions (video number 54: PLR of 2% and 49 seconds of total clip duration).

The lowest QoE scores are recorded for higher PLR (above 0.5%) and for the videos which contained 10 inserted clips that lasted 4 or 7 seconds each, i.e. 40 or 70 seconds in total (videos number 64 to 66 and 70 to 72). The worst average QoE score (4.16, i.e. *Fair* quality) is recorded for the video with PLR of 2% and 10 inserted clips which last 70 seconds in total.

It is noteworthy to mention that, when the PLR is the same and above 0.5%, lower average QoE scores are recorded for those videos which had more inserted clips, even though the total duration of all inserted clips was the same. For PLR of 1% this can be seen by comparing the average QoE scores for video number 10 and 22, 16 and 40, 34 and 46. For PLR of 1.5% comparison needs to be made between video number 11

and 23, 17 and 41, 35 and 47, while for PLR of 2% video number 12 and 24, 18 and 42, 36 and 48 have to be compared.

TABLE I. CHARACTERISTICS OF THE TEST SEQUENCES AND USER RATINGS

Video no.	PLR [%]	Number of inserted video clips	Video clip duration [s]	Total duration of all inserted clips [s]	Average QoE score / SD
1	0.05	1	1	1	8.41 / 1.03
2	0.1	1	1	1	8.08 / 0.69
3	0.5	1	1	1	8.10 / 1.16
4	1	1	1	1	8.81 / 0.69
5	1.5	1	1	1	8.96 / 0.89
6	2	1	1	1	8.40 / 1.43
7	0.05	1	4	4	8.42 / 0.63
8	0.1	1	4	4	7.89 / 0.35
9	0.5	1	4	4	8.10 / 1.40
10	1	1	4	4	8.75 / 0.55
11	1.5	1	4	4	8.26 / 0.75
12	2	1	4	4	8.17 / 1.87
13	0.05	1	7	7	8.32 / 1.28
14	0.1	1	7	7	7.90 / 1.14
15	0.5	1	7	7	7.46 / 1.44
16	1	1	7	7	8.63 / 0.95
17	1.5	1	7	7	7.85 / 0.81
18	2	1	7	7	8.19 / 0.46
19	0.05	4	1	4	8.41 / 0.97
20	0.1	4	1	4	7.84 / 0.66
21	0.5	4	1	4	8.59 / 0.91
22	1	4	1	4	8.23 / 1.10
23	1.5	4	1	4	7.66 / 1.03
24	2	4	1	4	7.67 / 1.03
25	0.05	4	4	16	8.20 / 1.02
26	0.1	4	4	16	7.73 / 0.63
27	0.5	4	4	16	6.83 / 1.17
28	1	4	4	16	7.56 / 0.57
29	1.5	4	4	16	7.06 / 0.97
30	2	4	4	16	6.70 / 0.70
31	0.05	4	7	28	7.74 / 0.61
32	0.1	4	7	28	7.60 / 0.53
33	0.5	4	7	28	7.60 / 1.06
34	1	4	7	28	6.54 / 0.92
35	1.5	4	7	28	6.29 / 1.41
36	2	4	7	28	5.88 / 2.32
37	0.05	7	1	7	8.07 / 0.78
38	0.1	7	1	7	7.89 / 0.94
39	0.5	7	1	7	7.03 / 1.29
40	1	7	1	7	7.80 / 1.12
41	1.5	7	1	7	6.84 / 1.14
42	2	7	1	7	6.02 / 2.27
43	0.05	7	4	28	8.20 / 1.25
44	0.1	7	4	28	7.13 / 1.12
45	0.5	7	4	28	6.84 / 1.61
46	1	7	4	28	6.33 / 1.71
47	1.5	7	4	28	5.88 / 0.88
48	2	7	4	28	5.63 / 1.00
49	0.05	7	7	49	7.63 / 1.46
50	0.1	7	7	49	6.75 / 1.97
51	0.5	7	7	49	6.74 / 1.45
52	1	7	7	49	6.23 / 1.30
53	1.5	7	7	49	5.04 / 0.46
54	2	7	7	49	4.84 / 1.94
55	0.05	10	1	10	7.89 / 0.67
56	0.1	10	1	10	7.87 / 1.14
57	0.5	10	1	10	7.67 / 1.42
58	1	10	1	10	6.89 / 1.17
59	1.5	10	1	10	6.01 / 1.07
60	2	10	1	10	6.03 / 1.49
61	0.05	10	4	40	7.87 / 0.88
62	0.1	10	4	40	7.18 / 1.19
63	0.5	10	4	40	6.62 / 1.78
64	1	10	4	40	6.41 / 0.96
65	1.5	10	4	40	5.17 / 1.10
66	2	10	4	40	5.35 / 1.82
67	0.05	10	7	70	7.59 / 0.71
68	0.1	10	7	70	7.10 / 1.07
69	0.5	10	7	70	5.93 / 1.10
70	1	10	7	70	4.68 / 1.25
71	1.5	10	7	70	4.57 / 1.41
72	2	10	7	70	4.16 / 0.96

The clips which were inserted into the original video signal contained various video artifacts such as jerkiness, frame freeze, blurring, blocking, error blocks, object persistence, edge busyness and mosquito noise. The clips which were degraded with higher PLR (e.g. 1.5 and 2%) contained more delivery artifacts which were easier to notice. This clearly led to lower QoE scores of such test sequences, but it is evident that the other two objective parameters (the number and total duration of inserted clips) also have an impact on user QoE.

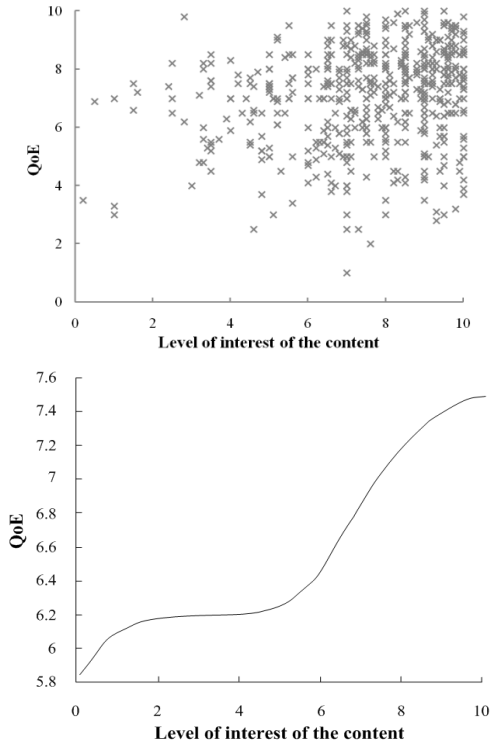

Figure 1. Correlation of the level of interest of the video content and user QoE

Apart from the interrelationship between the three objective parameters and user perception, we also found the non-linear correlation of the level of interest of the video content (documentary film) and user QoE. This is depicted on Fig. 1. The subject's level of interest is also rated on the 11-grade numerical scale (0-2 *Least interesting, boring*, 2-4 *Partially interesting*, 4-6 *Interesting*, 6-8 *Mostly interesting*, 8-10 *Very interesting*). As it can be seen from the figure, the content was interesting to most test subjects. That was desirable result, because in real-life users usually consume the content which interests them. It is also clear that the QoE scores are higher if the content is interesting to the subjects.

B. *User annoyance caused by PLR*

If a packet loss occurs during video streaming session, incoming video signal may contain video artifacts. The advent of video artifacts interferes with the seamless reproduction of the content and increases user dissatisfaction. Fig. 2 depicts the level of user annoyance caused by different PLR. The scores are also given on an 11-grade numerical scale with the following linguistic meanings: 0-2 *Imperceptible*, 2-4 *Perceptible but not annoying*, 4-6 *Slightly annoying*, 6-8 *Annoying* and 8-10 *Very annoying*. It is obvious that higher PLR causes higher annoyance level, although even the highest score (4.38) is well below the worst possible rating (*Annoying* and *Very annoying*).

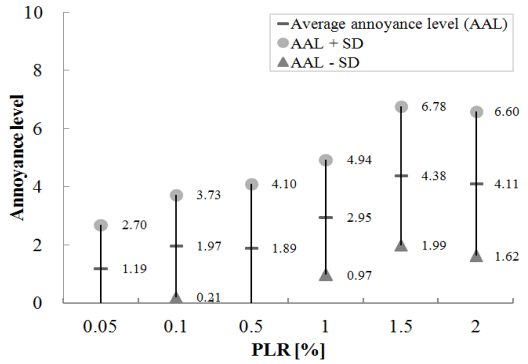

Figure 2. Level of user annoyance caused by different packet loss rate

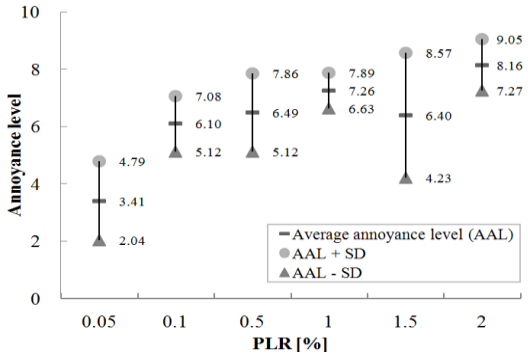

Figure 3. User annoyance caused by different PLR in a video with 10 inserted clips that last 70 s in total

The results showed that lower PLR (e.g. below 1%) is usually not perceived as annoying, while higher PLR is mostly perceived only as slightly annoying. Notwithstanding, note that Fig. 2 depicts the average annoyance scores for all videos with specific PLR, not taking into account different number of inserted clips (i.e. PLR occurrences) nor their total duration. Further analysis showed that the lowest scores, i.e. the highest annoyance level, are recorded for the videos containing 10 inserted clips that last 70 seconds in total (Fig. 3), which was expected given the results presented in Tab. 1. It might be noteworthy to know that the average scores greater or equal to 6 are given to the videos that contained at least 40 seconds of quality distortions.

C. *Visibility of inserted clips*

Test subjects were asked to quantify the instances when they noticed that the video quality was degraded during screening. Fig. 4 brings the results of this analysis. The numbers on the figure are calculated by subtracting the actual number of inserted clips (1, 4, 7 or 10) and noticed number of inserted clips by the subjects. The blue segment of the line indicates how many users failed to notice certain number of inserted clips. Conversely, the red segment indicates the instances when the subjects though that there is more clips than it actually was. For example, 100 test subjects failed to notice one more inserted clip in the video (e.g. instead of 4 clips, they have noticed only 3, i.e. the difference equals +1) and 48 test subjects noticed one additional, nonexistent clip (e.g. instead of 4 clips, they thought that they have seen 5). It can be observed that peeks of the line are all located on the blue segment, indicating that considerable number of test subjects (408) failed to notice some or all PLR occurrences in the video.

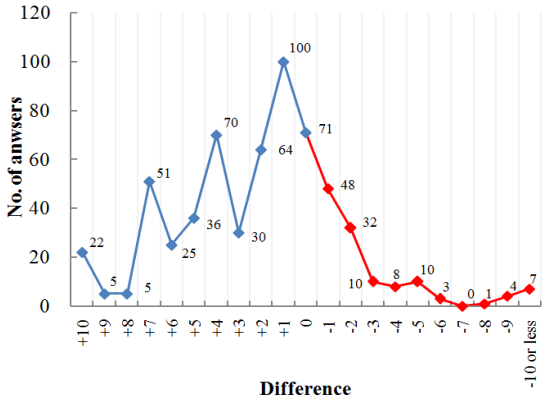

Figure 4. Difference between actual and noticed number of inserted clips

D. *Quantification of total duration of inserted clips*

The subject's ability to accurately quantify the time when the video quality was degraded was also analysed. Results of that analysis are depicted on Fig. 5.

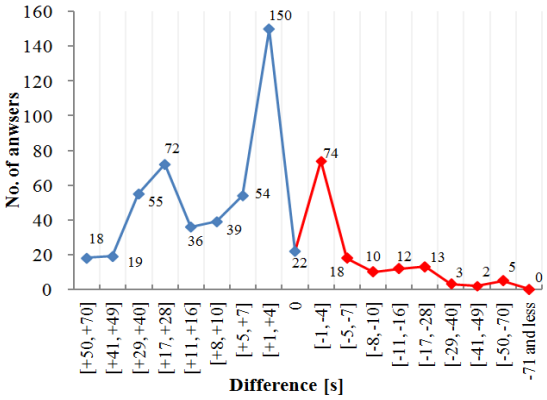

Figure 5. Difference between the actual and quantified duration of inserted clips by the subjects

The numbers on the figure are calculated by subtracting the actual total duration of the inserted clips and the subject's quantification of that duration. The line is also divided on two parts, indicating the instances when the subjects thought that the duration is shorter than it actually was (blue part) and when they thought it is longer (red part). Significant number of test subjects (443 on the blue segment of the line) failed to notice

(or remember) all instances when the quality was degraded, which corresponds with the previous findings (Fig. 4). However, there is one clear peak on the red segment of the line, indicating that 74 test subjects thought that the quality distortion lasted from 1 to 4 seconds longer than it actually was lasting.

IV. DISCUSSION

One hour videos, which were distributed to the subjects, contained one or more segments whose quality was degraded by various PLR. These segments contained various video artifacts (jerkiness, frame freeze, blurring, blocking, error blocks, object persistence, edge busyness and mosquito noise). The subjects expressed higher dissatisfaction for the videos that contained the segments (i.e. inserted clips) which were degraded with higher PLR (1, 1.5 and 2%). These clips had more video artifacts which were easier to notice and that led to lower average QoE scores. However, it is evident that the other two objective parameters (the number and total duration of inserted clips) also have an impact on user QoE. For instance, for the videos with the same PLR, it is shown that lower QoE scores are given if the videos contain more PLR occurrences, even when their total duration is the same.

Under real-life test conditions the majority of test subjects failed to notice significant number PLR occurrences and they were unable to accurately quantify their total duration. These findings confirm that choosing the right test environment is critical for QoE evaluation. The results presented on Fig. 4 clearly illustrate this by showing that 22 test subjects have not noticed all 10 inserted clips of degraded video (difference +10). Additionally, when there is only one inserted clip, the video quality is highly graded, despite of various PLR and duration of the clip (indicated in Tab 1). Owing to these findings it can be argued that in the real-life context occasional drop of network performances will not be adversely perceived by the users of the service.

As emphasized in [11] one of the factors influencing the subjects reasoning is the human short term memory. After watching one hour video, certain number of test subjects simply forgot the quality distortions which they might have noticed during screening. However, it is also clear that the videos with the most degradation (PLR above 0.5%, 7 or 10 inserted clips and more than 40 seconds of quality distortions) tend to be perceived as videos with *Fair quality*, while the extent of their distortions is *Slightly annoying* or *Annoying*.

The research revealed that the video content, which was used in the analysis, was interesting to most test subjects. This corresponds with the real-life conditions because in real-life users do not watch the videos which are not interesting to them. Furthermore, the non-linear correlation of the level of interest of the documentary film and user QoE disclosed that the quality is evaluated higher when the content is interesting for the subjects. It can be stated that if the content awakes higher interests of the users of the service, they become more forgiving to the perceived quality distortions.

V. CONCLUSION

In this research the subjective evaluation of video quality was conducted in real-life environment using the one hour video. Test subjects were uninformed about the purpose of the test before and during screening. Hence, instead of focusing on video quality, they were concentrated on video content.

This paper confirmed that test environment and longer test sequences can have significant influence on perceived service quality. The results indicate that users usually did not maintain negative attitude about quality distortions even when the PLR reached relatively high values for this type of service. This implies that certain level of flexibility exists when trying to match particular Quality of Service (QoS) demands of different services in IP networks.

For our future research it was important to understand the relationship between the chosen objective parameters. Specifically, it is shown that the impact of PLR on user perception depends on the number of instances when the PLR occurs and total duration of those instances. This knowledge will be used to develop a model for objective evaluation of QoE of video streaming service.

REFERENCES

[1] Y. Bai, Y. Chu, and M. R. Ito, "Dynamic end-to-end QoS support for video over the Internet," International Journal of Electronics and Communications (AEÜ), vol. 65, no. 5, pp. 385-391, 2011

[2] M. Matulin, and Š. Mrvelj, "State-of-the-practice in evaluation of Quality of Experience in real-life environments," Promet–Traffic&Transportation, vol. 25, no. 3, pp. 255-263, 2013

[3] P. Reichl, P. Fröhlich, L. Baillie, R. Schatz, and A. Dantcheva, "The LiLiPUT prototype: a wearable lab environment for user tests of mobile telecommunication applications," In Proc. of the Human Factors in Computing Systems, New York, USA, pp. 1833-1838, 2007

[4] S. Jumisko-Pyykkö, and M. Hannuksela, "Does context matter in quality evaluation of mobile television?" In Proc. of the 10th Conference on Human-Computer Interaction with Mobile Devices and Services, Amsterdam, Netherland, pp. 63-72, 2008

[5] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. De Meester, "Assessing the perceptual influence of H.264/SVC signal-to-noise ratio and temporal scalability on full length movies," In Proc. of the 1st International Workshop on Quality of Multimedia Experience, San Diego, USA, pp. 29-34, 2009

[6] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. De Meester, "Assessing Quality of Experience of IPTV and Video on Demand services in real-life environments," IEEE Transactions on Broadcasting, vol. 56, no. 4, pp. 458-466, 2010

[7] P. Frohlich, S. Egger, R. Schatz, M. Muhlegger, K. Masuch, and B. Gardlo, "QoE in 10 seconds: are short video clip lengths sufficient for Quality of Experience assessment?" In Proc. of the 4th International Workshop on Quality of Multimedia Experience, Yarra Valley, Australia, pp. 242-247, 2012

[8] D. Hands, and S. Avons, "Recency and duration neglect in subjective assessment of television picture quality," Applied Cognitive Psychology, vol. 15, no. 6, pp. 639-657, 2001

[9] P. Datta, L. Izdebski, N. Kumar, and K. Suh, "It came to me in a stream The upward arc of online video, driven by consumers," Cisco, 2012

[10] ITU-T: "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union (Rec. ITU-T P.910), 2008

[11] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, "Quality of Experience of VoIP service: A survey of assessment approaches and open issues," IEEE Communications Surveys & Tutorials, vol. 14, no. 2, pp. 491-513, 2012