Quantifying Subjective Quality Evaluations for Mobile Video Watching in a Semi-Living Lab Context

Toon De Pessemier, Katrien De Moor, Wout Joseph, *Member, IEEE*,
Lieven De Marez, and Luc Martens, *Member, IEEE*

***Abstract*—This paper discusses results from an exploratory study in which Quality of Experience aspects related to mobile video watching were investigated in a semi-living lab setting. More specifically, we zoom in on usage patterns in a natural research context and on the subjective evaluation of high and low-resolution movie trailers that are transferred to a mobile device using two transmission protocols for video (i.e., real-time transport protocol and progressive download using HTTP). User feedback was collected by means of short questionnaires on the mobile device, combined with traditional pen and paper diaries. The subjective evaluations regarding the general technical quality, perceived distortion, fluentness of the video, and loading speed are studied and the influence of the transmission protocol and video resolution on these evaluations is analyzed. Multinomial logistic regression results in a model to estimate the subjective evaluations regarding the perceived distortion and loading speed based on objectively-measured parameters of the video session.**

***Index Terms*—mobile video, Living Lab, subjective evaluation, Quality of Experience**

I. Introduction

Over the last decade, Quality of Experience (QoE) has become a very important research topic in our contemporary ICT (Information and Communication Technology) environment. Broadcasters, video providers, and operators can no longer compete solely on the number of channels or the content but increasingly make high definition channels and QoE as a service differentiator [1]. QoE has become a key factor in routing mechanisms and resource management schemes for network operators and IPTV providers [2]. The quest for approaches that enable QoE measurement in the context of ubiquitous, 'always on' multimedia consumption is challenging but crucial. Researchers have already tried to grasp the influence of both static and more dynamic factors upon the quality of people's experiences with ICT products, applications and services for a long time. However, there is no magical formula to solve this complex problem. The increasing collaboration between researchers from different disciplines and epistemological positions is in this respect not only enriching, but also necessary. In this respect, the definition of QoE is a much debated topic in the QoE community: various considerations and contributions have been made to the literature [3-5]. Both from a theoretical and empirical perspective, this concept has been broadened over the last years. In the definition of QoE by the ITU (International Telecommunication Union) [6], it was noted in the margin that 'the overall acceptability of an application or service, as perceived by the end-user' might differ according to the 'expectations' of this end-user, e.g., concerning the application or service in question or concerning its actual use. Moreover, although not further specified in this definition, 'context' might also affect QoE. Previous research and observations however, seem to indicate that the importance of contextual, content-related, and user-related dimensions cannot be overemphasized in this respect and might even be the key to 'ubiquitous QoE'.

This more user-centric perspective fits in the broader theoretical and methodological shift from technology-driven to a more open and user-driven innovation paradigm [7], in which users are increasingly put at the center of the innovation process. As reflected in the importance of the QoE concept, users have become more demanding and expect that products, services and applications address their personal and situational requirements [8], allowing them to have a good and pleasurable (quality of) experience anywhere and at any time. This is especially challenging in the mobile media domain, which is characterized by an exponential growth in the number of mobile devices, services, and applications, by the availability of various new content technologies and access networks, and by the massive adoption of mobile services by users. As mobile applications are used in dynamic and heterogeneous usage contexts, insights in the objective and

Manuscript received xxx, 2011. This work was supported by the IBBT−GRASP project, co-funded by the IBBT (Interdisciplinary institute for BroadBand Technology), a research institute founded by the Flemish Government. W. Joseph is a Post-Doctoral Fellow of the FWO-V (Research Foundation - Flanders).

T. De Pessemier, W. Joseph, and L. Martens are with Ghent University/ IBBT, Dept. of Information Technology, Gaston Crommenlaan 8, B-9050 Ghent, Belgium, Fax: +32 9 33 14 89 9 (e-mail: Toon.DePessemier@intec.ugent.be, Wout.Joseph@intec.ugent.be, Luc.Martens@intec.ugent.be).

K. De Moor and L. De Marez are with Ghent University/ IBBT, Dept. of Communication Sciences, Korte Meer 7-9-11, B-9000 Ghent, Belgium (e-mail: KatrienR.DeMoor@ugent.be, Lieven.DeMarez@ugent.be).

subjective dimensions that may influence users' QoE in these contexts, have become crucial in view of QoE optimization [9].

Over the past years, numerous video quality assessment methods and metrics have been proposed with varying computational complexity and accuracy. Full-reference and reduced-reference media-layer objective video quality assessment methods, whether or not considering natural visual characteristics or perceptual (human visual system) characteristics, are extensively classified, reviewed, and compared [10]. However, these metrics only measure objective parameters, which is insufficient to reliably estimate end-users' subjective overall perception of the quality (i.e. the QoE). Therefore, the most reliable way of assessing and measuring QoE is conducting subjective experiments, in which human observers evaluate a series of video sequences [1].

At the level of subjective measurements, there is a strong tradition in experimental research taking place in controlled laboratory settings. This type of research makes it possible to investigate the relative influence of particular isolated parameters on users' quality perceptions. Yet, especially when the focus is on 'ubiquitous QoE' and its interplay with dynamic contextual and user-related variables, the complementary value of more ecologically valid approaches should be explored.

Although no common or standardized approaches have been developed in this respect, interesting work has already been done in this area, e.g., in the domain of pervasive computing [11], and mobile TV [12]. Various researchers pointed to the relevance of the living labs approach for 'integrating technology components into the complex environment of the wireless world and end-users in their daily life' [13]. In the definition of Følstad [14] living labs are 'environments for innovation and development where users are exposed to new ICT solutions in (semi-)realistic contexts, as part of medium- or long-term studies targeting evaluation of new ICT solutions and discovery of innovation opportunities'. Drawing on the abovementioned open and user-driven innovation rationale, the living lab approach might help to facilitate the continuous and systematic involvement of end-users and to enable researchers to understand the drivers and barriers of QoE in heterogeneous real-life contexts [15]. Moreover, as living labs 'bring the lab to the people' and draw on 'real' experiences from 'real' users, QoE research in such settings will likely yield more accurate results and have a higher ecological validity than research in controlled environments [16]. In this respect, Staelens et al. compared QoE assessment performed in controlled laboratory environments and in the natural setting of people's everyday life context [1]. They discovered significant differences concerning impairment visibility and acceptability. In general, impairments showed to be less visible during real-life QoE assessment. So, conclusions which are obtained using a standardized subjective-quality assessment methodology may not always hold on the case of real-life QoE assessment since user expectations and context influence QoE. In previous research [15], a framework for evaluating QoE in a mobile living lab setting was presented. The exploratory study presented in this paper draws on this framework. The framework monitors context information, subjective user evaluations, and Quality of Service (QoS) aspects in real-life settings.

In this paper, we explore contextual aspects and subjective quality evaluations related to mobile video watching in a natural environment. Nonetheless, due to the fact that the test users were 1) given an additional device to perform the test, 2) asked to watch a limited and pre-defined content list, and 3) only had one week to finish the test, we label this study as semi-living lab. More specifically, we zoom in on the viewing behavior as observed in a natural research context and on the subjective evaluation of four different video classes, combining low and high-resolution videos with two transmission protocols for video, being Real-time Transport Protocol (RTP) and progressive download using HTTP (HyperText Transfer Protocol) and TCP (Transmission Control Protocol). User feedback was collected by means of short questionnaires on the mobile device, combined with traditional pen and paper diaries. Additionally, this paper proposes an innovative model to predict the subjective quality evaluations based on objectively-measured parameters related to the video session.

The setup of this study is shortly described in the following section. A number of observations and results regarding the user's context and subjective evaluations are shared in Section 3. Section 4 describes how to quantify these subjective evaluations in terms of objectively-measured parameters. Finally, Section 5 is dedicated to our conclusions.

II. User Study

A. Study Setup

The test users were asked to watch 28 pre-defined movie trailers (covering different genres) in their everyday life context (when and where they wanted), but within a time-span of 1 week (weekend included). Table I lists the titles and main genres of the trailers that were used. (This metadata is originating from the Internet Movie Database, IMDb.) All movie trailers were relatively short and had a duration between 2 and 3 minutes. To avoid boredom, the test users had to watch all 28 trailers only once during the experiment. The viewers were able to decide themselves in which order they watched the clips. The list consisted of 7 low-resolution videos using RTP, 7 high-resolution videos using RTP, 7 low-resolution videos using progressive download and 7 high-resolution videos using progressive download. Both RTP and progressive download are often used for the transmission of video content but have different characteristics in terms of possible influence on the user's experience.

TABLE I. TITLES AND MAIN GENRES OF THE MOVIE TRAILERS

Title	Genre (from IMDb)
2012	action, adventure, drama
21	crime, drama
Saw VI	horror, mystery, thriller
The Wrestler	drama, romance, sport
Toy Story 3	animation, adventure, comedy
Twilight 2: New Moon	adventure, drama, fantasy
Valkyrie	drama, history, thriller
Babel	drama
Milk	biography, drama, history
Mr Untouchable	documentary, crime
Prince of Persia: The Sands of Time	action, adventure, fantasy
Rush Hour 3	action, comedy, crime
Self-Medicated	biography, drama
Shrek the Third	animation, adventure, comedy
The Kite Runner	drama, romance
28 weeks later	horror, thriller
Michael Jackson's This Is It	documentary, music
Mr Woodcock	comedy, romance, sport
Quantum of Solace	action, adventure, crime
Sex and the city	comedy, romance
The Dark Knight	action, crime, drama
Then She Found Me	comedy, drama, romance
Alvin and the Chipmunks	animation, comedy, family
Avatar	action, adventure, fantasy
It's complicated	comedy, romance
Ong Bak 2	action
Sherlock Holmes	action, adventure, crime
There will be blood	drama

In the case of streaming media using RTP, video playback does not suffer from interruptions due to rebufferings, but the loss of multiple (consecutive) packets may lead to noticeable distortions for the user. Although intelligent mechanisms in core and distribution networks may prevent congestion and packet loss, video streaming over IP networks is error-prone and subject to a wide range of distortions, artifacts, and degradations during transmission [17].

Progressive download on the other hand is based on a reliable transport layer protocol for host-to-host data transfer (in most cases TCP), which can avoid the loss of packets by means of packet retransmissions. However, this protocol may cause rebuffering interruptions during video playback.

These transmission protocols were combined with two video qualities in order to investigate their impact upon the user's quality evaluation. Table II summarizes the technical parameters of the two quality version of the mobile videos. All videos were coded with an average bitrate and resolution as specified in the table. The video list in the user interface was randomly mixed and the users were not informed about the different qualities and transmission protocols.

TABLE II. TECHNICAL PARAMETERS OF THE MOBILE VIDEO

Low Resolution Video			
Audio		**Video**	
Codec	AAC LC	Codec	H.264/AVC
Bit rate	32 Kbit/s	Bit rate	128 Kbit/s
Channels	2	Resolution	142*80
Sampling frequency	44100 Hz	Frame rate	24 fps
High Resolution Video			
Audio		**Video**	
Codec	AAC LC	Codec	H.264/AVC
Bit rate	128 Kbit/s	Bit rate	384 Kbit/s
Channels	2	Resolution	512*288
Sampling frequency	44100 Hz	Frame rate	24 fps

Every test user was handed over a Google Nexus One mobile phone, running on Android 2.1 as operating system, to watch the videos. In order to gather immediate and explicit user feedback after each watched video, six short questions concerning the content, general technical quality, fluentness of the video, loading speed, eventual distortions, and the user's physical context had to be answered on the device. After the video playback, these questions pop-up on the screen and users have to answer them before the next video can be played. The first four questions are evaluated on a 5-level subjective quality evaluation scale (Absolute Category Rating 5-point scale (ACR)) ranging from 5 (excellent) to 1 (bad). The choice of the rating scale might be seen as an important element in the subjective testing methodology. Nevertheless, a direct comparison between four different rating scales based on experimental data showed no overall statistical differences between the different scales [18]. For the evaluation of the perceived distortion, a 5-point scale was used ranging from 5 (not perceptible) to 1 (perceptible and very annoying). Both the numbers and corresponding labels were shown to the test users. Four options were selectable for the question regarding the physical context of the user: "on the move", "at home", "at work", or "somewhere else". In the case of selecting "somewhere else", the user could specify his or her location.

Additionally, a traditional paper diary was completed by the test users immediately after playback: for every watched video, a diary sheet containing additional (open and closed) questions was filled in. The goal of this paper diary was to give users the opportunity to provide more detailed and qualitative feedback regarding the video session and their experience through some open questions. Since inputting text on mobile phones is difficult and tedious, mobile phones are not the optimal tool to gather detailed feedback. Therefore, we opted for an alternative feedback tool: a small paper diary that can also be used in case of technical problems with the device such as an application crash or a dead battery. Concerning the appreciation of the content, users were firstly asked to indicate whether or not they would want to watch the entire movie (6-point scale) and whether they had already seen it before.

Secondly, they were asked to rate their general experience on a 6-point scale ranging from very positive (6) to very negative (1) and to mention aspects that on the one hand influenced their experience (in a positive way as well as in a negative way) and on the other hand, that might help to enhance/improve the experience. In contrast to the questions on the device, a 6-point rating scale was used to evaluate the user's general experience and the desirability of the video content as was done by Kortum and Sullivan [19]. They investigated the effect of content desirability on subjective video quality ratings. By adopting their 6-point rating scale, correlations between the desirability of the movie content and subjective ratings of the video quality can be compared in future research.

The third question of the diary asked the users whether other people were around the user during watching (in a radius of approximately 5 meter) and whether or not the presence of others was perceived as disturbing. Finally through the fourth question, users had to indicate whether the overall technical quality of the video during the watching experience was a) acceptable in every context, b) acceptable but only in the context in which the user watched it or c) not acceptable. Although each user watched each movie trailer in only one context, this question provides insights into the users' experiences and behavior regarding video watching in different contexts.

As already briefly mentioned above, the research design draws on two complementary voting interfaces because of the specific nature of the data that we wanted to collect. The 'on the device' voting interface is very suitable for collecting an immediate, in situ evaluation, as close to the experience as possible. As the short questionnaire on the device was part of the viewing protocol, we are sure that the test subjects rated the videos immediately after viewing. As a result, we were able to limit possible biases on the rating procedure due to memory errors or due to the time elapsed between the watching and the evaluation.

At the same time, we deliberately aimed to limit the number of questions on the device as much as possible in order not to disrupt the user's natural flow when using the smartphone. However, we also wanted to collect additional (contextual) information, for which the diary method is more suitable.

TABLE III. OVERVIEW OF THE QUESTIONS ON THE DEVICE AND IN THE PAPER DIARY

	Digital questions on the device	Possible answers
1	How do you evaluate the content of this movie trailer?	5-point rating scale: 1 = bad; 5 = Excellent
2	How do you evaluate the technical quality of this movie trailer in general?	5-point rating scale: 1 = bad; 5 = Excellent
3	Did you perceived distortion in the video during playback?	5-point rating scale: 5 (not perceptible); 1 (perceptible and very annoying).
4	How do you evaluate the fluentness of the video playback?	5-point rating scale: 1 = bad; 5 = Excellent
5	How do you evaluate the loading speed of the video?	5-point rating scale: 1 = bad; 5 = Excellent
6A	Select your current location. I am …	4 options: on the move, at home, at work, or somewhere else.
6B	(if somewhere else)Where exactly are you?	Open question
	Paper diary questions	Possible answers
1A	Please indicate whether or not you agree with the statement: "I would like to completely view this movie"?	6-point rating scale: 1 = completely disagree; 6 = completely agree
1B	If you have already seen this movie before, please indicate by coloring the button.	Yes or No
2A	Please indicate on the scale below how you have experienced this viewing session. My overall experience was ...	6-point rating scale: 1 = very negative; 6 = very positive
2B	Which aspects did you experience as positive during the viewing session?	Open question
2C	Which aspects did you experience as negative during the viewing session?	Open question
2D	Which aspects could enhance or improve your experience?	Open question
3A	Were other people in a radius of approximately 5 meter around during the viewing session? If so, how many?	No or Yes + number
3B	(If other people were in the immediate surroundings) Did you experience their presence as disturbing?	No or Yes because … (Open question)
4	Please indicate what is most applicable: the technical quality of the video was…	3 options: a) acceptable in every context, b) acceptable but only in the context in which I watched it or c) not acceptable

Before the actual test started for every user, instruction meetings were organized in groups of five users. After some general information on how to switch on/off, use, charge the device etc., it was explained how to access the test application and how to select and watch the videos. Next, it was also shown how to navigate from one question to the next and fill in the questionnaire using the touch screen. At the end of the briefing session, every test user was given a device, a diary, and an instruction leaflet with practical information, screenshots, and relevant instructions related to the grading scales and univocal interpretation of the questions. In total, the data gathering phase took just over three months since the five available devices rotated among the test users.

During the video watching, relevant objective video and network parameters were logged: video quality (resolution and bitrate), transmission protocol (RTP or progressive download), packet-loss rate for the audio and video track, the mean and maximum jitter (i.e., the variability over time of the packet latency across the network) for audio and video, network type (e.g., UMTS, HSDPA, GPRS), number of handovers (i.e., all kinds of radio cell reselections), and inter-system handovers (i.e., different data connection-type cell reselections e.g., between UMTS and HSDPA), and RSSI (received signal strength indicator). In addition, a number of objective parameters concerning the video session and watching

behavior were registered: movement of the device (i.e., the GPS signal to track the mobility during the video watching), early interruption of the video (e.g., due to network disconnection), metadata about the video (id, title, length) and the start and end of the session (timestamp).

B. *Sample Description*

Previous research has already indicated that the appreciation of and interest in the offered content possibly has a major impact on users' QoE [19-21]. Moreover, it has been argued that previous experiences and user-related characteristics should also be taken into account. Therefore, a specific group of users was targeted in this experiment. 30 test users were recruited by an experienced panel manager from IBBT-iLab.o (a research division with a strong expertise in living lab research and panel management). The recruited test users were meeting the three main selection criteria: 1) being a smartphone user, 2) having watched mobile video at least once in the preceding month and 3) having indicated to have an interest in the content category used in this study (movies / movie trailers). Since the idea of a living lab implies staying close to the realistic situation, these criteria were laid down in order to reflect the natural viewing conditions and behavior of the users as much as possible. In total, 29 people (24% female and 76% male) between 20 and 61 years old (M = 33.10, S.D = 9.97) participated in the study. One test user, who had agreed to participate, dropped out just before the actual test period. Due to time constraints, this test user was not replaced. Every test user received a gift voucher of 10 Euro.

The data obtained via the user study were assembled and integrated into one data file containing the subjective evaluations collected through the questionnaires on the device, the paper diary entries for every question, and the logged technical data. Sessions in which video watching was not possible due to the lack of a data connection, had to be removed. Moreover, two additional sessions in which video watching was possible were removed (one outlier with an erroneous value, and one sample in which the user's ratings were missing). After excluding these sessions, 753 data samples were obtained, providing the data to analyze the viewing behavior of the user, and to develop a model for the subjective evaluation of video quality in a mobile context.

III. Viewing Behavior and Subjective Evaluations

In terms of physical context of the test users, we found that most of the videos were watched at home (82.7%) and at work (9.7%). Only 5.2% was watched during travelling. 2.4% was watched somewhere else (including e.g., at the house of a friend or relative, in a café, or in a museum). Although one might expect that more videos would be watched during travelling, this was not the case in this study. In fact only 8 of the 29 users (i.e. 27% of the users) watched videos during travelling. Moreover, previous research on mobile TV points to the same direction: e.g. in [22], the results from a living lab study on mobile TV showed that most viewing occurred at home. In terms of the acceptability of the video quality, no significant differences were found according to the physical context of the users. The reason for this might be that the large majority of the videos (82.7%) were watched at home. The answers on the question regarding the acceptability of the quality were equally distributed. 33% of the videos were evaluated as "acceptable in any context"; 33% was evaluated as "acceptable, but only in the context that I watched it"; and the remaining 34% was evaluated as "not acceptable".

Figure 1 shows the types of data network that were used to transfer the videos to the mobile device according to the physical location of the user. If (inter-system) handovers occurred during the video transmission, the connection type that was responsible for the majority of the video transfer is considered in Figure 1. 7% of the videos is transmitted on a GPRS network (General Packet Radio Service). Only 1% of the videos is using a UMTS (Universal Mobile Telecommunications System) connection. The most used connection type (51% of the videos) is the HSDPA network (High-Speed Downlink Packet Access), followed by the Wi-Fi (Wireless Fidelity) connection (41 %). As shown in Figure 1, the type of data network is closely related to the physical context of the user. E.g., Wi-Fi is almost exclusively used at home.

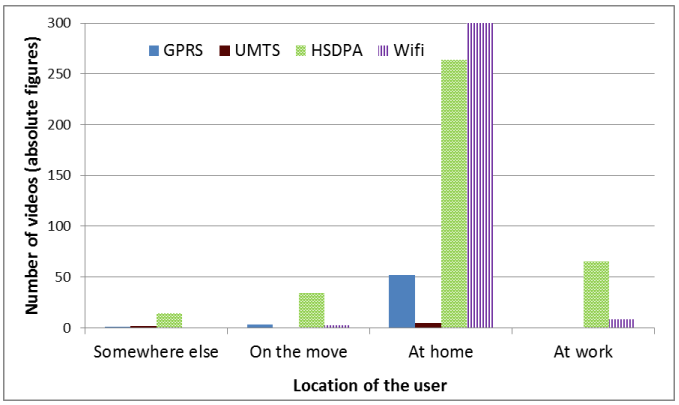

Fig. 1. Type of data network that was used in terms of the location of the user.

Time wise, Figure 2 shows that the evening (from 18.00 till 24.00 o'clock) was the most popular watching time, followed by the afternoon. This is the case both on week days and on weekend days. In absolute numbers, most videos were watched during the week (72.8%), which makes sense since every user had one week to finish the test so only two weekend days, but five weekdays were included in the test period. So users were about equally active during weekend days as during the weekdays.

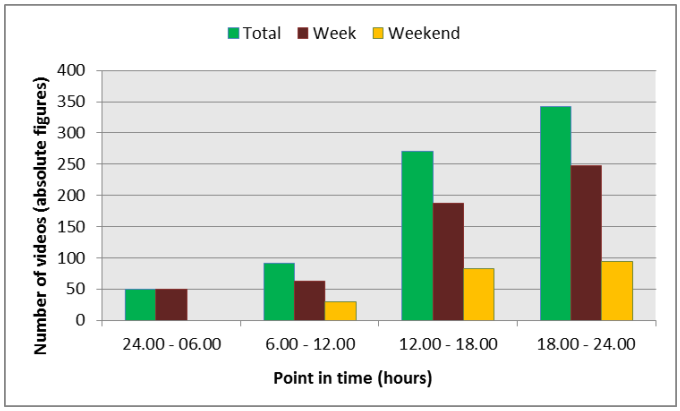

Fig. 2. Viewing behavior in terms of time.

In 61.4% of the cases, no other people were in the immediate surroundings of the user (radius of approximately 5 meter) during the video watching. 22.8% of the videos were watched by the user in presence of one other person. In the majority of the viewing sessions in which other people were in the surroundings of the test user, the presence of these people was not experienced as disturbing (89.8%). In the remaining 10.2%, the talking of the others and noise made by them or coming from other sources (such as the TV) is often mentioned as disturbing factor. However, there is no significant influence of the number of people around (as a variable of the 'social context' of the user) while watching on the overall experience rating.

Figure 3 compares the mean quality ratings for the four technical combinations. Although individual ratings are ranging from very negative to very positive (as illustrated in Table IV and VI for the loading time and distortion), the mean values of the subjective evaluations are all quite positive and range between 2.8 and 4.1.

A one-way Analysis of Variance (ANOVA) relies on the restrictive assumptions of homogeneity of the variances of the distributions and normality of the distributions of the residuals [23]. Also the commonly-used T-test, a statistical hypothesis test which compares the mean values of 2 groups, relies on the assumption that the samples follow a normal distribution [23]. Since the user evaluations are integer values, these assumptions may not apply. Therefore, the four technical combinations were compared using the Wilcoxon rank test as alternative. The Wilcoxon rank test is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. This way, the subjective ratings were compared using the different technical combinations as the grouping variable. Significant differences ($p = .05$) were identified for the rating of the technical quality, distortion, and fluentness.

In the briefing preceding the start of the experiment, the technical parameters that users had to evaluated were explained as follows: "By technical quality, we mean the overall quality of the different technical features that you – as a viewer - can perceive (these include e.g., the sharpness of the image, the synchronization between the sound and image, the fluentness of the video, loading speed, visual artifacts or errors in the video, ...). Other aspects, such as the appreciation of the content of the clip, are not part of this technical quality." A high score corresponds with a positive evaluation of the technical quality; a low score indicates that the user is not at all or not really satisfied with the technical quality. Fluentness was explained to the test subjects as the degree to which the images follow up on each other without delay, interruptions or freezes. Distortion was defined more broadly and different examples of possible distortions were given (e.g., blurriness, blockiness, …). The test subjects were asked whether they experienced a distortion and if so, whether this distortion was annoying or not. The loading speed is evaluating the waiting time between selecting a video and the start of the video playback.

The *technical quality* of the combination "high-resolution video – progressive downloading" is perceived as significantly better than that of the other combinations of video resolution and transmission protocol. The technical quality of the high-resolution RTP videos is evaluated as the second best option and is significantly better than the two combinations with low resolution. The technical quality of the low-resolution RTP videos received the lowest evaluation (Mean = 2.72; Standard Deviation = .96).

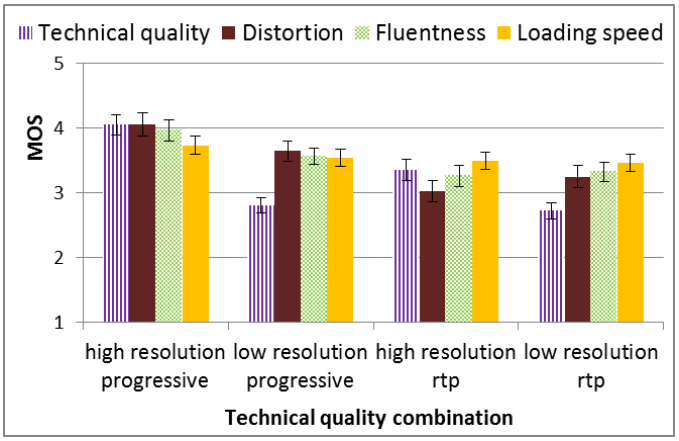

Fig. 3. Average quality ratings according to the 4 technical quality combinations

In terms of the *perceived distortion* (Figure 3), the differences between the high-resolution progressive downloading videos and videos streamed via RTP (both low and high resolution) are significant. High-resolution video sessions using progressive download received the best evaluation regarding the perceived distortion and the difference in MOS with videos streamed via RTP is approximately 1 unit. The difference between the perceived distortion for the low-resolution videos transmitted via progressive downloading and the videos streamed via RTP is also statistically significant (0.62 and 0.39 on the MOS for respectively high and low-resolution RTP videos). This subjectively-observed difference can be explained by the characteristics of the transmission protocol: (multiple) packet

loss may induce audio-visual distortions for video that is streamed using RTP, whereas progressive download based on TCP relies on retransmissions in case of packet loss.

In terms of perceived *fluentness* (Figure 3), the high-resolution progressive downloading videos were perceived as more fluent than the streamed videos. Although the progressive downloading videos may introduce playback interruptions due to rebufferings, many of these video sessions in the experiment suffered only from a small number of short rebufferings which were tolerated by the users. Or in the case of a fast network connection, no rebufferings at all were required. Finally, no significant difference was noticed in terms of perceived *loading speed* for the various combinations of video resolution and transport protocol.

The Wilcoxon rank test, comparing the score for the overall experience which was given in the paper diary as dependent and the resolution / protocol combinations as grouping variables, yields similar results: the high-resolution progressive downloading videos result in a significantly higher QoE ($p = .05$) than the other combinations.. The high-resolution RTP videos provide users the second best QoE and were evaluated significantly better than both low-resolution combinations. Furthermore, the subjective evaluations showed that overall experience of the users was the worst in the case of low-resolution RTP. This negative experience is in accordance with the poor evaluation of the technical parameters of the low-resolution RTP videos.

As the result of a qualitative analysis of the user feedback obtained via the diaries, Figure 4 shows the number of comments in three categories (positive aspects, negative aspects, and things that could be changed to enable a better experience) for the four video combinations.

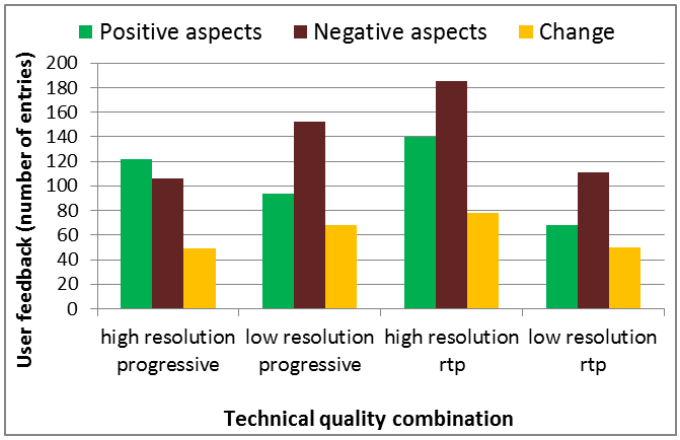

Fig. 4: Overview of the number of qualitative user comments according to the technical condition

Only for the first category of videos in Figure 4 (high resolution – progressive), the number of positive aspects that were mentioned, supersedes the number of negative aspects and proposed changes (122 positive comments, 106 negative comments, and 49 proposed changes). Most negative feedback is given for the high-resolution videos streamed using RTP (185 entries), for which the fluentness and perceived distortion was rated lowest (see Figure 3). The open questions were included in the diary since it is not always clear on which specific aspects user ratings are based. Moreover, the use of numerical expressions of perceived quality is always problematic in a way since these ratings provide little insight in what this really implies from a user point of view. The answers on the open questions contain valuable information on the individual video watching sessions. First of all, they illustrate that the test subjects are precise and detailed and performed the test in a rigorous way, e.g., they make clear distinctions between different technical artifacts in their verbal evaluations. Additionally, the answers revealed that other, non-technical aspects are also considered by test subjects when asked to reflect on positive and negative aspects of the viewing experience. Examples are issues related to the content itself (e.g., good acting, presence of a specific actor, story, emotional impact of the content, associations, …), the sound (e.g., compelling music, aggressive sound, …) , the colors (e.g., too bright or too dark, unnatural, …), etc. Although the technical quality may be negatively perceived, it does not automatically result in a negative viewing experience: the experience can still be rather positive because e.g., the user liked the music, the story, or a specific actor in the trailer. Qualitative user feedback can help to understand how the different combinations were evaluated and why one technical quality condition was preferred over another.

IV. Modeling the Subjective Quality Evaluations

In this section, the subjectively-perceived quality of the video sessions is further investigated in order to model the subjective evaluations based on objectively-measured, technical parameters.

A. *Perceived Loading Speed*

One of the quality aspects that the users could evaluate was the loading speed of the video. Table IV shows the rating options for evaluating the perceived loading speed, the mean of the objectively-measured loading time for each of the rating options, the number of video sessions that received such a rating, and the corresponding marginal percentages of the ratings (i.e. the percentage of the videos which received the specific rating). The loading time is measured as the time period between selecting a video and the moment when the video starts playing.

Although the subjective evaluations show some inconsistencies, the results indicate that the loading time of the majority of the video sessions (62.4%) is evaluated as “good” or even “excellent”. Conversely, for a considerable part of the video sessions (15.4%), the subjectively-perceived loading time is “poor” or “bad”.

TABLE IV. SUBJECTIVE EVALUATIONS AND MEAN OBJECTIVE MEASUREMENT OF THE LOADING TIME

Rating of the Loading Speed	Mean loading time (s)	Number of sessions	Marginal Percentage
5 = Excellent	2.9	125	16.6%
4 = Good	3.5	344	45.8%
3 = Moderate	5.7	167	22.2%
2 = Poor	18.7	56	7.5%
1 = Bad	29.3	59	7.9%
Total	7.1	751	100%

Therefore, the influence of the objectively-measured loading time on the subjective evaluation of the perceived loading speed is investigated. Besides the loading time, the duration of the video might also influence the subjective evaluation of the loading speed. But since all videos of the experiment had approximately the same duration, this parameter is not included in the analysis.

An important aspect during the selection of the most appropriate statistical technique is the type of data that has to be analyzed. Although the answers on the multiple choice questions consist of a verbal description and a corresponding number, these ratings have to be considered as ordinal numbers. This means that it is possible to rank the values, but the real distance between categories is unknown. E.g., the difference between "excellent" and "good" is not treated the same as the difference between "good" and "moderate".

Given the ordinal nature of the subjective ratings, traditional statistical techniques, such as linear regression and Pearson correlation, are not suitable for investigating the effect of objective parameters on the rating behavior of the users. One candidate technique to analyze the subjective ratings is ordinal logistic regression. Ordinal logistic regression is an extension of a binary logistic regression model (which is a model used for prediction of the probability of occurrence of an event by fitting data to a logistic function) [23]. Ordinal regression modifies the binary logistic regression model to incorporate the ordinal nature of a dependent variable by defining the probabilities differently. Instead of considering the probability of an individual event, this technique considers the probability of that event and all events that are ordered before it [24].

However, one of the assumptions underlying ordinal logistic regression is that the relationship between each pair of outcome groups is the same. In other words, ordinal logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. This is called the proportional odds assumption or the parallel regression assumption [24]. However, this test of parallel lines showed that this assumption was not valid for the obtained subjective evaluations. Therefore, different models have to be defined to describe the relationship between each pair of possible ratings by multinomial logistic regression. *Multinomial logistic regression* is also a generalization of binary logistic regression and allows more than two discrete outcomes. This regression model is used to predict the probabilities of different possible outcomes of a dependent variable (in our case the subjective rating), given a set of independent variables which may be real-valued, binary-valued, categorical-valued, etc. (in our case the objective parameters) [25].

Multinomial logistic regression compares the probability of a specific event against the probability of a reference event. For this analysis, the subjective evaluation of the loading speed was selected as dependent, the objectively-measured loading time is an independent (covariate), and the reference event was the evaluation of the loading speed as "moderate". So for each rating, the regression model provides a function for the ratio of the probability of obtaining that specific rating, e.g., P(excellent) and the probability of obtaining the reference rating P(moderate), in terms of the objectively-measured loading time, i.e., LT. Table V lists the results of this multinomial logistic regression analysis: The probability ratios as exponential functions in terms of the objectively-measured loading time (in seconds). The likelihood ratio chi-square of 164.7 with a p-value < 0.0001 and 4 degrees of freedom tells us that our model as a whole fits significantly better than a model without the loading time as predictor. (The chi-square statistic is the difference in 2-log-likelihoods between the final model and a reduced model.)

TABLE V. THE RESULTS OF THE MULTINOMIAL LOGISTIC REGRESSION ANALYSIS WITH THE SUBJECTIVE EVALUATION OF THE LOADING SPEED AS DEPENDENT AND THE OBJECTIVELY-MEASURED LOADING TIME AS A COVARIATE (LT= LOADING TIME).

Probability Ratio	Estimated Function
P(excellent) /P(moderate)	Exp(0.261-0.143*LT)
P(good) /P(moderate)	Exp(1.075-0.081*LT)
P(moderate) /P(moderate)	1
P(poor) /P(moderate)	Exp(-1.652+0.060*LT)
P(bad) /P(moderate)	Exp(-1.800+0.068*LT)

Figure 5 visualizes these probability ratios for an objectively-measured loading time between 0 and 40 seconds. The graph shows that for short loading times (less than 10 seconds), a high probability exists that users will evaluate the loading speed as "good" or "excellent". Given the high marginal percentage of video sessions evaluated as "good" (45.8% in Table IV), the probability of obtaining "good" as subjective evaluation is higher than the probability of obtaining "excellent". If the measured loading time is more than 13 seconds, users are more willing to evaluate the loading speed as "moderate" than to rate it as "good". For short loading times, users are not inclined to give low evaluations like "bad" or "poor". However after a loading time of approximately 27 seconds, ratings with the label "bad" or "poor" are more likely than the reference rating, i.e. "moderate". And for instance after 40 seconds of waiting time,

it is 2.5 times more likely that users perceive the loading speed as "bad" than that users perceive it as "moderate" (Figure 5).

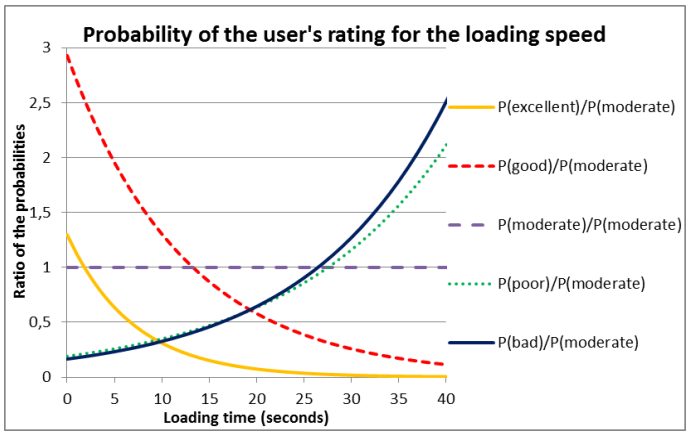

Fig. 5. The probability ratios of the possible ratings for the perceived loading speed.

B. *Perceived Distortions*

In contrast to progressive download, which relies on packet retransmissions in case of packet loss, video streaming via RTP might suffer from audio-visual distortions if packets are lost during transmission. Therefore, the influence of packet loss on the subjectively-perceived distortion during mobile video watching was investigated for the video sessions which are streamed via RTP. Table VI shows the rating options for evaluating the perceived distorting during video watching, the mean of the objectively-measured packet-loss rate for each of the rating options, the number of video sessions that receive such a rating, and the corresponding marginal percentages of the ratings. This analysis was based on the data samples obtained for the mobile video sessions using RTP (high- and low-resolution videos).

Table VI shows that sessions which receive a positive evaluation regarding the perceived distortion ("not noticeable" or "noticeable, not annoying") are characterized by a low packet-loss rate (mean values of 0.8% and 0.4%). In contrast, low ratings for the perceived distortion ("noticeable, annoying" or "noticeable, very annoying") are typically due to high packet-loss rates (mean values of respectively 18.9% and 32.5%). Therefore, the influence of this packet-loss rate on the subjectively-perceived distortion during mobile video watching is further investigated.

TABLE VI. SUBJECTIVE EVALUATIONS OF THE DISTORTION AND MEAN OBJECTIVE MEASUREMENT OF THE PACKET-LOSS RATE

Rating of the Distortion	Mean packet loss rate (%)	Number of sessions	Marginal Percentage
5 = Not noticeable	0.8	88	23.7%
4 = Noticeable, not annoying	0.4	68	18.3%
3 = Noticeable, slightly annoying	3.1	78	21.0%
2 = Noticeable, annoying	18.9	67	18.0%
1 = Noticeable, very annoying	32.5	71	19.1%
Total	10.5	372	100%

For the same reason as in the analysis of the loading speed, a multinomial logistic regression analysis was performed to estimate the probability of obtaining a specific rating as a function of the packet-loss rate. For this analysis, the subjective evaluation of the perceived distortion was selected as dependent, the objectively-measured packet-loss rate is an independent (covariate), and the reference event was the evaluation of the distortion as "noticeable, slightly annoying". For each rating option, Table VII lists the ratio of the probability of obtaining that specific rating, e.g., P(not noticeable), and the probability of obtaining the reference rating, P(noticeable, slightly annoying), in terms of the objectively-measured packet-loss rate, i.e. PL. The likelihood ratio chi-square of 149.3 with a p-value < 0.0001 and 4 degrees of freedom tells us that our model as a whole fits significantly better than a model without the packet-loss rate as predictor.

Figure 6 visualizes the probability ratios of Table VII for a packet-loss rate ranging from 0% to 40% (using a logarithmic scale). Video sessions with a limited packet-loss rate have a higher probability to obtain a positive rating regarding the perceived distortion ("not noticeable" or "noticeable, not annoying") than to receive the reference rating (i.e. "noticeable, slightly annoying"). In contrast, if more than 2.6 % of the packets are lost during transmission, the probability that users are slightly annoyed by distortions is higher than the probability that users do not notice these distortions (full decreasing line versus dashed horizontal line in Figure 6). If the packet-loss rate during video watching is higher than 30%, the probability of receiving a positive evaluation from the user is very small (less than 5% of the probability of receiving the reference rating).

TABLE VII. THE RESULTS OF THE MULTINOMIAL LOGISTIC REGRESSION ANALYSIS WITH THE SUBJECTIVE EVALUATION OF THE DISTORTION AS DEPENDENT AND THE OBJECTIVELY-MEASURED PACKET-LOSS RATE AS A COVARIATE.

Probability Ratio	Estimated Function
P(not noticeable) /P(noticeable, slightly annoying)	Exp(0.302-0.115*PL)
P(noticeable, not annoying) /P(noticeable, slightly annoying)	Exp(0.147-0.287*PL)
P(noticeable, slightly annoying) /P(noticeable, slightly annoying)	1
P(noticeable, annoying) /P(noticeable, slightly annoying)	Exp(-0.609+0.058*PL)
P(noticeable, very annoying) /P(noticeable, slightly annoying)	Exp(-0.903+0.072*PL)

Negative evaluations of the perceived distortion are less likely than the reference rating for low values of the packet-loss rate. However, the rating options "noticeable, annoying" and "noticeable, very annoying" are more likely than the reference option "noticeable, slightly annoying" as soon as the packet-loss rate is higher than respectively 10.5% and 12.5%.

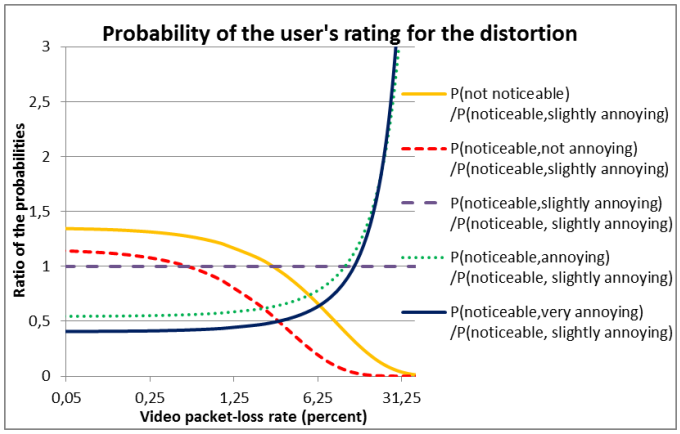

Fig. 6. The probability ratios of the possible ratings for the perceived distortion.

V. Conclusion

In this exploratory study, we investigated Quality of Experience (QoE) related to mobile video watching in a semi-living lab environment. 28 video trailers were watched by the test users in random combinations of two video resolutions (high and low) and two data transfer protocols for video (Real-time Transport Protocol and progressive download using TCP/HTTP). The participants were able to watch the videos when they wanted, where they wanted and user evaluations were gathered by means of questionnaires on the device, complemented with traditional pen and paper diaries. The results illustrate that most videos were watched at home and in the afternoon and evening. In most cases, no other people were around during the watching session. The presence of other people did not have a significant influence on the overall experience rating and was in 90% of the cases, not perceived as a disturbing factor.

We compared the subjective quality ratings for the four technical quality combinations. Both the qualitative and quantitative feedback showed that the high-resolution progressively downloaded videos yield a significantly better experience than the streamed videos in terms of perceived technical quality, distortion, fluentness, and overall experience. The technical quality of the low-resolution videos using RTP was evaluated as the worst. Analysis of the qualitative user feedback could help to understand which aspects influenced the overall QoE in a positive and negative way in the four technical quality combinations.

The influence of the objectively-measured loading time on the subjective evaluations of the loading speed was evaluated via a multinomial logistic regression analysis. The resulting model showed that if the loading time increases from 10 to 30 seconds, the subjective evaluations of the loading speed gradually evolve from mainly positive to mainly negative.

For video sessions using RTP, we investigated the subjectively-perceived distortion during mobile video watching as a function of the video packet-loss rate. The probability of receiving a positive rating is rapidly decreasing if packet-loss occurs during video watching and video sessions with a packet-loss rate of more than 10% are in general evaluated as “annoying” or even “very annoying”.

The presented study can be seen as an example of QoE research in a real-life, semi-living lab setting. Given the increased emphasis on contextual variables and subjective, user-related characteristics of QoE, new context-aware tools and measurement approaches should be explored to take these dimensions into account. Whereas research in controlled settings is very valuable to assess the influence of particular, isolated parameters, research in more natural and ecologically valid settings might help to better understand the interplay between different parameters and their relative influence on the overall QoE.

References

[1] N. Staelens, S. Moens, et al. "Assessing Quality of Experience of IPTV and video on demand services in real-life environments," *IEEE Trans. Broadcasting*, vol.56, no.4, pp.458-466, Dec. 2010.

[2] S. Balasubramaniam, J. Mineraud, et al. "An evaluation of parameterized gradient based routing with QoE monitoring for multiple IPTV providers," *IEEE Trans. Broadcasting*, vol.57, no.2, pp.183-194, June 2011.

[3] P. Reichl, “From charging for quality-of-service to charging for quality-of-experience”, *Annals of. Telecommunications*, vol. 65, no. 3-4, pp. 189-199, 2010.

[4] S. Möller, K.-P. Engelbrecht, C., Kühnel, I. Wechsung, B.Weiss, “A taxonomy of quality of service and quality of experience of multimodal human-machine interaction”, in *First Int. Workshop on Quality of Multimedia Experience (QoMEX’09)*, July 29-31, San Diego, CA, July 2009.

[5] D. Geerts, K. De Moor, et al. “Linking an integrated framework with appropriate methods for measuring QoE”, In *Proceedings of the Second International Workshop on Quality of Multimedia Experience*, pp. 158-163, June 2010.

[6] ITU-T. “Definition of quality of experience (QoE)”, International Telecommunication Union, Liaison Statement, Ref.: TD 109rev2 (PLEN/12), Jan. 2007.

[7] E. Haddon, E. Mante, et al., *Everyday innovators: researching the role of users in shaping ICT's*, Dordrecht: Springer, 2005.

[8] T. De Pessemier, K. De Moor, et al. “Investigating the influence of QoS on personal evaluation behaviour in a mobile context,” Multimedia Tools and Applications, Springer Netherlands, 2011.

[9] K. De Moor and L. De Marez, "The Challenge of user- and QoE-centric research and product development in today's ICT-environment," *Innovating for and by users*, vol. 1, no. 3, 2008.

[10] S. Chikkerur, V. Sundaram, M. Reisslein, L.J. Karam, "Objective video quality assessment methods: a classification, review, and performance comparison," *IEEE Trans. Broadcasting*, vol.57, no.2, pp.165-182, June 2011.

[11] L. Li-yuan, Z., Wen-an, and S. Jun-de, “The research of quality of experience evaluation method in pervasive computing environment”, In *Proceedings of the 1st International Symposium on Pervasive Computing and Applications*, pp. 178-182, Aug. 2006.

[12] S. Jumisko-Pyykkö, and M.M. Hannuksela, “Does context matter in quality evaluation of mobile television?”, In *Proceedings of 10th International Conference on Human Computer Interaction with Mobile Devices and Services*, pp. 63-72, 2008.

[13] M. Ponce de Leon, M. Eriksson, S. Balasubramaniam, and W. Donelly, “Creating a distributed mobile networking testbed environment - through the living labs approach” In *Proceedings of the 2nd International IEEE/Create-Net Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities*, pp. 135-139, July 2006.

[14] A. Følstad, “Living Labs for innovation and development of information and communication technology: a literature review”, *The Electronic Journal for Virtual Organizations and Networks Special Issue on Living Labs*, vol. 10, pp. 99-131, Aug. 2008.

[15] K. De Moor, I. Ketyko, et al. “Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting,” *Mobile Networks and Applications*, vol. 15, no. 3, pp. 378-391, 2010.
[16] J. Schumacher, V-P. Niitamo (eds.), “European living labs – a new approach for human centric regional innovation”, Wissenschaftlicher Verlag Berlin, Berlin, 2008.
[17] J. Asghar, F. Le Faucheur, and I. Hood, "Preserving video quality in IPTV networks," *IEEE Trans. Broadcasting*, vol.55, no.2, pp.386-395, June 2009.
[18] Quan Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcasting*, vol.57, no.1, pp.1-14, March 2011.
[19] P. Kortum and M. Sullivan, "The effect of content desirability on subjective video quality ratings", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 52, no. 1, pp. 105-123, Feb. 2010.
[20] P. Kortum and M. Sullivan, “Content is king: the effect of content on the perception of video quality,” in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, 2004, vol. 48, pp. 1910–1914.
[21] S. Jumisko, V.P. Ilvonen, and K. V¨a¨an¨anen-Vainio- Mattila, “Effect of tv content in subjective assessment of video quality on mobile devices,” in Proceedings of SPIE, 2005, vol. 5684, pp. 243–254.
[22] D. Schuurman, T. Evens, and L. De Marez, “A living lab research approach for mobile TV”. In *Proceedings of the seventh European conference on European interactive television conference (EuroITV '09).* ACM, New York, NY, USA, pp. 189-196, 2009.
[23] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied linear statistical models*, McGraw-Hill/Irwin, 5th edition, 2005.
[24] “SPSS Data Analysis Examples Ordinal Logistic Regression” UCLA: Academic Technology Services, Statistical Consulting Group. from http://www.ats.ucla.edu/stat/spss/dae/ologit.htm (accessed November 23, 2011).
[25] T. F. Liao, “Interpreting Probability Models, Logit, Probit, and Other Generalized Linear Models”, *Quantitative Applications in the Social Sciences*, vol. 7, 1994.