User-centered Quality of Experience of mobile 3DTV: How to evaluate quality in the context of use?

Satu Jumisko-Pyykkö, Timo Utriainen
Tampere University of Technology

ABSTRACT

Subjective quality evaluation experiments are conducted for optimizing critical system components during the process of system development. Conventionally, the experiments take place in the controlled viewing conditions even though the target application is meant to be used in the heterogeneous mobile settings. The goal of the paper is a two-fold. Firstly, we present a hybrid User-Centered Quality of Experience (UC-QoE) evaluation method for measuring quality in the context of use. The method combines quantitative preference ratings, qualitative descriptions of quality and context, characterization of context in the macro and micro levels, and the measures of effort. Secondly, we present results of two experiments using this method in different field settings and compared to the laboratory settings. We conducted the experiments with a relatively low quality range for current and future data rates for mobile (3D) television by varying encoding parameters for simulcast stereo video. The study was conducted on a portable device with parallax barrier display technology. The results show significant differences between the different field conditions and between field and laboratory measures.

Keywords: Quality of experience, mobile 3D television, audiovisual quality, multimedia quality, context, context of use

1. INTRODUCTION

Future multimedia services, such as mobile 3D television, needs to fill the users' requirements in the context of use to become successful. For mobile 3D television, huge amounts of 3D audiovisual data, limited bandwidth, vulnerable transmission channel, and constraints of the receiving devices (e.g. screen size, computational power, battery life-time) set tight technical requirements. The optimization over this chain can cause noticeable impairments influencing on user's experienced quality at the end. In product development, it is desirable to know whether the system or its critical components fulfill the user's requirements in its actual contexts of use in the development phase as early as possible.

Subjective quality evaluation methods can be used to conclude perceptual multimodal acceptability, preferences, and critical quality factors[1],[2],[3]. There are two main approaches to measure the excellence of the stimuli. Psychoperceptual approach, popular among engineering society, follows the quantitative tradition of experimental research (highly valuing control and repeatability) based on methodological recommendations of International Telecommunication Union (ITU)[4],[5]. The second approach, User-centered Quality of Experience (UC-QoE), rooting to the HCI community, attempts to gain a high external validity and realism by relating the quality evaluation to the potential use of system or service[6],[7]. To mention s few aspects of UC-QoE, it stresses the importance of potential users, critical system components, the context of use, and understanding of interpreted quality parallel to the conventional quality evaluation[8]. This paper represents the latter of the approaches and set the special focus on the context of use as a part of evaluation. The context of use can be understood as an entity (CoU) that surrounds the mobile human-computer interaction, contains components of task, physical, social, temporal, and technical and informational contexts, and can be dynamic and heterogeneous[9].

There has been increasing interest to evaluate usability of mobile services and devices in the field during the last five years while the quality evaluations in these circumstances are still rare. The experiments in the contexts of use are carried out to gain ecological validity, validate the results (from controlled circumstances e.g. lab), test the systems which evaluation require natural circumstances (e.g. context-aware systems)[10],[11],[12],[13]. There are two recent quality evaluation studies conducted in the field. Jumisko-Pyykkö & Hannuksela[14] compared controlled and three field settings potential for mobile television (Bus – travel, Cafe – relax, Railway station – wait) by varying residual transmission error rates. Later, Knoche & Sasse[15] compared laboratory evaluations to assessment carried out underground when bitrate and image resolution were varied for mobile TV. The results of both studies showed that lower quality requirements were set in the field on an acceptable quality levels giving the opportunities for efficient optimization of system resources[14],[15].

Though these studies set good starting points for contextual quality evaluation they can be criticized from their methodological weaknesses. Among the most notable, circumstances of the experiments surrounding the causal effects are inadequately reported making hard to understand the possible co-influencing factors and to compare quasi-experimental studies.

The goal of the paper is two-fold. Firstly, we present a hybrid User-Centered Quality of Experience (UC-QoE) research methodological framework for measuring quality in the context of use. The key components of the method are quantitative quality preference ratings, qualitative descriptions of quality and context, analysis of characteristics of context in macro level and situational characteristics of context in a micro level. Secondly, we present results of two experiments using this method in altogether three different field settings and compare the results to those gathered from the laboratory settings.

2. QUALITY EVALUATION IN THE CONTEXT OF USE

Multimedia quality is a combination of produced and perceived quality[16]. Produced quality can be categorized into three abstraction levels: content, media and network, and it describes the factors relating to the content and the system[17]. To be played back on small screens of mobile devices, the content is captured, encoded, and transmitted over the mobile broadcasting channel to be received, decoded and displayed to the end-user. Perceived quality, characterized by active low-level and high-level perceptual processes, represents the end-user's side of multimedia quality[16]. Experienced quality by the end-user can be negatively influenced by impairments in any stage of the broadcasting chain[18]. Additionally, the assessment of usefulness or fitness to purpose of use is an essential part of evaluations of quality, and thus the evaluations are not restricted only to the characteristics of the interpreted stimuli[19].

Subjective quality evaluation methods can be used to evaluate multimodal acceptability, user preferences, and critical quality factors based on human perception[1],[2],[3] and information gained by using them can be used in optimization of the system, such as network, media or content creation parameters. Two main approaches to measure subjective quality can be identified. The psychoperceptual approach follows the quantitative tradition of experimental research and bases itself on the recommendations of the International Telecommunication Union (ITU)[4],[5] valuing control and repeatability over realism. However, their viewpoint to quality is limited and error-centric for system development studies that target for high external validity for actual system usage. The second approach, user-centered quality of experience (UC-QoE) as a quality evaluation method is a collection of factors and independent methods that relate the quality evaluation to the potential use of system or service[6],[7]. It takes into account *1) potential users as quality evaluators, 2) necessary system characteristics including its potential content and critical system components, 3) potential context of use resulting evaluation in the controlled experimental and quasi-experimental settings 4) evaluation tasks in relation to expected goals of viewing*[8]. The method aims also at understanding the interpretation of quality and includes ergonomic measures. If any of the listed factors that relate the evaluation into potential use of the system is taken into account in the evaluation research, the method can be referred to as a user-centered quality evaluation method [6],[7]. In this paper, we focus on quality evaluation in the context of use.

The context of use is a multidimensional concept and can be understood as an entity (CoU) that surrounds the mobile human-computer interaction[9]. A recent review by Jumisko-Pyykkö & Vainio[9] describes the mobile context of use consisting of components of physical, task, social, temporal, and technical and information context. In addition to these components, the context of use also contains additional properties: level of magnitude (micro, macro), level of dynamism (static, dynamic), pattern (rhythmic, random), and their typical combinations. The physical context contains spatial location, functional place and space, sensed environmental attributes, movements and mobility and surrounding artefacts. The task context contains multitasking, interruptions and task types. The social context consists of other people present, interpersonal actions and culture. The temporal context consists of duration of use, time of day/week/year, what the user did before/during/after use, actions in relation to time and synchronism. For quality evaluation research in the field, understanding of the components and the properties of the context of use is needed to be able to interpret the circumstances of the study, conclude the comparisons of quality requirements between the different contexts, and finally to compare the results of different experiments.

To measure quality in the context of use a shift in paradigm from experimental research towards quasi-experimentation is needed. An experiment is defined as *a study in which an intervention is deliberately introduced to observe its effects*[20]. Its building blocks are a treatment, an outcome measure, units of assignment, and it contains some comparisons which can be inferred attributes to the treatment[21]. The experiments can be categorized into four different

classes from the high level of control to the high level of realism: 1) randomized experiments in the laboratory, 2) analogue randomized experiments or simulations, 3) quasi-experiments, and 4) natural experiments[10],[20]. The first two types of experiments take a place in the controlled conditions while the last two typically contain ecological context including social context. To evaluate quality in the context of use, quasi experiments offer a ground for drawing conclusions over the causal effects in the natural circumstances (which cannot be done for 'after the fact' type of natural experiments)[20]. Quasi-experiments are experiments where *units are not assigned to the conditions randomly*[20] or where *an experimental intervention is carried out even while full control over potential causal events cannot be exerted*[10]. In literature, the terms field experiments, experimentation in the wild, reality testing in non-traditional environments, in-situ experimentation, are presented parallel to quasi-experimentation[20]. These experiments need special care in the instrumentation, in the selection of contexts, and in the consideration of the threats of validity[10],[20], they are also relatively demanding to design and implement[10],[22],[23]. Unlike the traditional randomized laboratory experiments however, they start to reveal aspects and characteristics of the actual use of the system[22].

There are only a few quality evaluation studies conducted in the field. Studies are conducted for mobile television while there are no previous studies of mobile 3D television in the field circumstances. Jumisko-Pyykkö & Hannuksela[14] compared controlled and three field settings potential for mobile television (bus – travel, cafe – relax, railway station – wait) by varying transmission error rates with television contents. Knoche & Sasse[15] compared laboratory evaluations to an assessment carried out in an underground by varying bitrate and image resolution for mobile TV. In both studies, lower requirements were set in the field to acceptable error rates and bitrates, underlining the possibility of further optimization of the system in the actual context of use. Although the studies offer good starting points for evaluating quality in the context they have unaddressed methodological shortcomings. The documentation of the methods in the field is very limited (How the studies were organized? What kind of circumstances surrounded the evaluation?) They for example assumed that description of the physical location is capable of covering the multidimensional aspects of context being in conflict with what is known about the characteristics of mobile contexts of use[9]. Furthermore, with the former study, the participants had context dependant parallel task to make situation more realistic while with the latter there was no parallel task used.

3. EVALUATION IN THE COU – PLANNING, DATA-COLLECTION AND ANALYSIS

An overview to the methodological framework is given in this section. The research phases are divided into three parts: planning, data-collection and analysis. The context of use is described in a) macro (high-level description of context e.g. whole situation in a certain location) and b) micro (situational e.g. second-by-second) levels. The quality evaluation in the context of use requires a hybrid data-collection and analysis procedure. We present the framework and show examples from our case study.

3.1 Planning phase

The planning phase contains two main actions 1) selection of contexts of use based on user requirements and 2) analysis of characteristics of chosen contexts and analysis of threats of validity as a supporting action. The planning phase targets at the analysis of contexts and their characteristics at the macro-level.

1) Selection of the contexts is based on user-requirements. Due to the practical time constrains of experiments only a certain amount of contexts can be chosen. It is important that chosen contexts represent the most common and diverse contexts of use to gather the heterogeneity of these circumstances. The context of use for mobile television is summarized based on previous field studies and user requirements. a) Physical context - The main locations for viewing mobile (3D) television are while commuting (public and private transportation), in the waiting halls, at home, parks and cafe's (during breaks or lunch)[24],[25],[26],[27],[28],[29]. b) Temporally, the macro breaks with a need to fill extra time are needed for viewing mobile television[26],[27],[29],[31]. The typical length of viewing varies from a couple of minutes up to 40 minutes, on average 10-15 minutes at the time[26],[29],[30]. The prime time is in the early morning, during lunch and early in the evening, before dinner time[25]. c) Social context – Primarily mobile (3D) TV is for one person viewing to minimize solitude, avoid social engagement and create private space[28],[29]. Co-viewing happens in some occasions (e.g. for forming a social group for sharing an experience, jokes) but it can also occur passively and involuntary in public transport or in crowded environments[27],[28],[29]. d) Task, technical and informational context – Focused viewing with enough time to concentrate on viewing characterizes the mobile TV viewing task[29]. Users prefer other media (music, radio) in the circumstances where the surroundings requires active maintaining of attention (e.g. way finding) or during the micro breaks[27],[28],[31].

In our case study, we selected three different contexts to the top of conventional laboratory quality evaluation (see Table 2). The selected scenarios were planned for personal viewing during a bus journey, waiting situation in the railway station, and viewing at a home-like situation. All scenarios were planned to have a macro break situation and low level of completing the attention resources in the contexts (e.g. no/low time pressure before leaving the bus).

2) Analysis of characteristics of chosen contexts. As the user requirements give a semantic description of the context of use, more detailed and systematic analysis of contexts is needed to understand the circumstances where the experiment is going to take place. To capture the main characteristics of contexts in a macro level we used CoU-MHCI form (constructed from recently published model based on review of characteristics of mobile contexts of use[9], see Table 2). The form helps 1) to richly identify and report the features of the contexts, 2) think the diversity between chosen contexts, and 3) systematically to think the potential factors influencing on quality requirements.

Potential threats to causal interference are listed explicitly in the planning phase to follow the instructions of quasi-experimental research[20]. We analyzed factors of statistical conclusion, internal, construct, and external validity with the aid of[10],[20]. In the experiments in the field settings, the awareness of experimenter expectancies is among the most critical factors. The instructions are given for the participant in a way that he/she leads the situation to make it as a realistic situation as possible and the moderator's task is to follow and shadow the participant. To avoid the mono-operation bias, written instructions for the moderator are needed. The pilot tests play a very central role in the field when checking the relevance of selected situations and testing with portable equipment. Finally, back-up plans are needed as it is very likely that changes occur during the experiment (e.g. as comparable context characteristics as possible to the original plan, extra batteries for camera system etc.).

3.2 Data-collection phase

The data-collection phase contains six parts: 1) Quality evaluation task, 2) Structured observation by moderator, 3) Work-load assessment of quality evaluation in a situation, 4) Interview about quality in contexts of use, 5) Interview about experiences of quality and 6) Situational data-collection - audio-video recording of the experiment. The parts 1-5 describe the macro-level data-collection tools in a context while part 6 targets on micro-level analysis of context.

The data-collection phase can be divided into three parts. The pre-test session contains training for the evaluation task and anchoring including the familiarization of contents and extremes of produced quality. Sensorial acuity can be tested also in this part. Actual test session is similar in each context. Prior to evaluation in the contexts, the moderator gives instructions for the participant. This part can contain a related scenario for the situation to make it more realistic (e.g. travel to the railway station to catch the train) and the participant takes the responsibility of leading the situation on his/her own[22]. The parts 1-4, 6 are carried out in each context while 4-5 are included in the post session.

1) Quality evaluation task: The quality preference ratings are collected retrospectively after each stimulus using a simple evaluation task. In our case study, we collected both overall quality satisfaction ratings on a 11-point unlabelled scale and acceptance of quality for viewing mobile 3D television on a binary scale (yes/no) for ~30 seconds lasting stimuli[5],[16]. For the consumer products, with novel technologies and heterogeneous parameter combinations it is desirable to know weather the presented quality reaches the acceptance threshold[16]. Furthermore, availability of test scenes restricted our content selection, but in the optimal case longer non-repeated stimuli might be preferred to enable a more realistic viewing situation (e.g. [14]). In overall, the challenge in this part is to maximize the viewing and natural behavior and minimize the obstruction of the evaluation task and tools.

2) Structured observation by moderator: During evaluation, the moderator shadows the participant and observes the situation with the aid of a semi-structured observation form based on CoU-MHCI and fills it at the end of context. The goal is to record the characteristics of the situation and make notes about the special attractions during the study. Examples of observation form[32]: For the physical context we collected lighting level, noise level, bus movements, user posture, user's distance to the device measured from the face, device position, movements of the device, and other artifacts. For the social context, other people (few-crowded), people nearby (radius 2m), and moderator position were recorded. For physical and social contexts the 11-point scales with labels in the extremes were used and the dynamisms of the situation were also marked. Social, technical and physical interruptions were collected to capture the task context (none-several; 11-point scale). Temporally, finishing time of task in relation to target and time of day were collected.

3) Work-load assessment of quality evaluation in a situation: After the evaluation task in a context, participant fills a questionnaire about the demands of evaluation in the context using NASA-TLX[33]. It examines the overall workload in the terms of mental demand, physical demand, time pressure, effort, frustration, and performance[33].

4) Interview about quality in contexts of use: Experiences and impressions of quality in the context are shortly interviewed during transition to the next context. These transitions have occurred to be very natural and relaxing situations for short interviews to collect the first impressions of quality in the context.

4-5) Interview about quality and quality in the contexts of use: In the post-experimental session, after contextual evaluations, broader semi-structured interview targeting on experiences about the quality contexts and quality are conducted. Importance of the interview is to build up understanding on the participants own experiences, descriptions of quality in these settings and verify the user requirements. The example main questions: 'What kind of factors you paid attention to while evaluating quality?', 'What kind of factors you paid attention while evaluating quality in these situations?', 'What kind of factors you paid attention in these situations?', 'Would you view mobile 3D television in these types of situations?'

6) Situational data-collection - audio-video recording of experiment: Light weight mobile usability lab containing several mini-video cameras (one for face, one for the user interface and one for the participant's field of view) and audio recording is used to capture the situational data over the whole experiment[34].

3.3 Analysis phase

In this phase, all collected data is first analyzed separately and finally integrated containing the following parts: 1) Characteristics of CoU - To identify the actual characteristics of contexts of use central values of moderator's CoU form are counted and results updated to the planned CoU-MHCI form. Other parts of the analysis targets on focus of the study – on contextual influences on quality. 2) Influence of context of use on quality requirements and workload are analyzed statistically. 3) The analysis of interview data about experience of contexts and situational audio-video recordings is based on data-driven frameworks being applicable to the not-well understood research phenomenon. From latter, it is possible to extract objective data such as attention gaze information[35]. In the next section, we present a selected set of results from our case study and evaluate the appropriateness of the different data-collection techniques.

4. CASE: IMPACT OF EVALUATION CONTEXTS ON EXPERIENCED QUALITY

4.1 Research method

The case study consisted of two parts, referred to as Experiment1=EXP1 and Experiment2=EXP2.

Participants: A total of 60 participants (aged 18-45, 50% male/female) participated to the study. They were divided equally between the two experiments. They were mostly (80%) naïve or untrained participants and a maximum of 20% of the participants were categorized into the group of innovators and early adopters regarding their attitudes towards technology[4],[5],[36].

Test procedure: Pre-test session consisted of sensorial tests (visual (20/40) and stereovision acuity (.6), color vision), demographic data collection and combined anchoring and training where the extremes of quality samples were presented. The whole data-collection task including the quality evaluation task is presented in detail in 3.2. Evaluation task was repeated in different evaluation contexts (EXP1 contained two and EXP2 three contexts, see Table 2).

Parameters and contents: The chosen parameters are listed in (for more details, see[32]). The video bitrates include the entire video stream (two video channels with 3D content, one with 2D) to compare quality of conventional 2D mobile television broadcasts to 3D stereovideo simulcast broadcasts. To provide comparability between 3D and 2D, additional higher video bitrates were selected for 3D to provide at least the comparable bandwidth per video channel.

The used video codec was x264 'Skystrife' b1077 using H.264/AVC baseline profile and the audio codec was Nero AAC 1.3.3.0 with AAC-HEv2 in stereo mode using normalized volume. The clips were muxed into MP4-containers using meGUI 0.3.1.1010 via Avisynth 2.5.7 scripting to avoid compromising quality with intermediate steps. The clips were encoded into a widescreen 16:9 letterbox resolution of 640px x 360px. For 3D, the video channels were squeezed horizontally to half their width and placed side-by-side for encoding. The 2D clips used the left video channel. Audio was set to 75 dBA (+10 dBA for peaks) and presented using in-ear-type headphones. The stimuli used four different potential contents fro mobile 3D television with variable audiovisual characteristics, for details see[32]. The stimuli material was presented centered on the screen covering a 3.3 inch area in diameter and was played using VLC 0.8.6 video player in full-screen mode. The stimuli were presented on a Telson mobile 3D device with a 4.3 inch 800px x 480px touch-enabled transmissive autostereoscopic parallax barrier screen, interleaved at the pixel level. The presentation order of all stimuli was fully randomized.

Table 1. Parameters for the experiments[32].

Parameters	EXP1	EXP2
video bitrate/frame rate/audio bitrate - presentation mode		
160kbps/10fps/48kbps - 3D	V	
160kbps/15fps/48kbps - 2D and 3D	V	
320kbps/10fps/48kbps - 3D	V	
320kbps/15fps/18kbps - 2D and 3D		V
320kbps/15fps/48kbps - 2D and 3D	V	V
768kbps/10fps/48kbps - 3D	V	
768kbps/15fps/18kbps - 3D		V
768kbps/15fps/48kbps - 3D	V	V
1536kbps/24fps/48kbps - 3D	V	

Context of use: Chosen contexts (3) represented the common contexts of use for mobile 3D television. Controlled laboratory evaluation was used in both experiments parallel to the contextual evaluations. See characteristics in Table 2. The order of the evaluation contexts was fully randomized in EXP1 and partly in EXP2 (starting at the station or the lab).

Table 2. Central characteristics of evaluation contexts of use. The presentation follows the model of CoU-MHCI[9], in the table [D] denotes a dynamic property of context.

Components/properties	Lab (EXP1, EXP2)	Home (EXP1)	Bus (EXP2)	Station (EXP2)
Physical context				
Functional place	Laboratory conditions[4],[5]	Simulated home	Local bus	Café, railway station
Sensed attributes	A: quiet	A: quiet	A: noisy [D]	A: noisy [D]
(A=audio, V=visual)	V: calm, indoor	V: calm, indoor	V: noisy, light [D]	V: noisy, light [D]
Movements	M: *	M: none	M: bus [D]	M: none
(M=movement, P=position)	P: straight	P: lean	P: lean [D]	P: lean [D]
Artifacts (other than answer sheet)	*	Accessories (e.g. pillows)	Bag	Bag, refreshments
Temporal context				
Duration	30min	30min, macro	25min, macro	25min, macro
Time of day	Vary	Vary	Vary	Vary
Actions-time	*	Extra time	Extra time, pressure at end	Extra time, pressure at end
Task context				
Multitask 1	Quality evaluation	Quality evaluation	Quality evaluation	Quality evaluation
Multitask 2	*	Relax	Search bus-stop (visual)	Check time for train (audio, visual)
Interruptions	*	None	Physical and social [D]	Physical and social [D]
Task type	*	Entertain	Entertain	Entertain
Social context				
Persons present	Moderator	None	Bystanders, moderator [D]	Bystanders, people near, moderator [D]
Interpersonal actions	*	*	Possible	Possible
Technical and informational context				
Other systems	*	*	Mobile usability lab	Mobile usability lab
Properties				
Level of dynamism	Static	Static	Dynamic (physical, task, social)	Dynamic (physical, task, social)
Other related factors				
Motivations	*	Entertain, pass time, relax	Entertain, pass time	Entertain, pass time
Viewing distance	Fixed (~40cm, 10x picture height[37])	Freedom to adjust	Freedom to adjust	Freedom to adjust
Device volume	Fixed (75dBA[4],[5])	Freedom to adjust	Freedom to adjust	Freedom to adjust

4.2 Characterization of CoU

Results - As a part of the methodological framework, tools based on the CoU-MHCI[9] model provided useful help in the different phases of study. In the planning phase, it acted as a tool for characterizing the selected contexts (Table 2). During the data-collection of EXP2 the characteristics of context were collected using a structured observation. For

EXP1, the form was filled after viewing the recorded videos from the home-like context. To report the main realized features of context the characteristics were updated to the original plan based on the average values of the observed results. In addition, the results of the observation form also gave hints about the aspects to be looked more in detail from situational data-analysis (e.g. user's movements).

Appropriateness of the evaluation technique – The utility of characterization of CoU as a part of evaluation method is to describe the circumstances of the study. These results can be used to explain contextual differences and similarities in different quality requirements and they can help to provide comparability between contextual studies. In the terms of complexity and effort, data is easy to collect and analyze. In future work, some aspects of physical contexts might be collected by using sensors in the data-collection phase.

4.3 Quality preference ratings

Results – The results of quality preference ratings are presented for three comparable parameter cases (shown in Figure 1) between the studied contexts. Only 15 participants were included in the bus context while others have 30 participants.

Acceptance of quality: Quality was experienced being slightly higher in the controlled laboratory (58.9%) analogue home-like context (56.4%; McNemar's test: $p<.01$, EXP1). In the second experiment, the quality was experienced slightly more acceptable in the field conditions compared to the laboratory (Lab: 56.1%, Bus: 57.2%, Station 62.8%; significant difference between station and lab $p<.001$). These results indicate that participants can be more tolerant towards quality in natural field circumstances and confirms the previous results[14].

Overall quality satisfaction: The results show interaction between the studied quality level and context for the comparable parameter cases between the experiments. Over the whole quality range, a good between-test reliability in the laboratory evaluations was shown between the experiments (Mann-Whitney U: $p>.05$) and they were in the same level with the evaluations of home-like context ($p>.05$). Similarly, the evaluations of real-life contexts (bus, station) are equally rated over the quality range (Wilcoxon: $p>.05$). In the low quality range (3D mode), the most critical quality evaluations are given in the artificial laboratory context and home-like context (station vs. home/lab (both studies) $p<.01$; bus vs. home/lab (EXP1) $p<.05$).

Appropriateness of the evaluation technique – Quality preference ratings are the primary data source in the field settings, but they require other types of data for explaining the results (e.g. CoU analysis). In the terms of complexity and effort, this quality evaluation data is easy to collect and analyze.

4.4 Workload

Results – Overall workload is the average of the six TLX factors (mental demand, physical demand, time pressure, effort, frustration, and performance)[33]. Evaluation tasks in different contexts did not have an influence on overall workload in any of the experiments (Table 3) or between them ($F(4,130)=0,298$, $p=0,879$, ns).

Table 3. Influence of context on overall workload.

CONTEXT	OVERALL WORKLOAD (NASA TLX) - Mean (SD)	EFFECT
Lab (EXP1)	4.03 (3.04)	t=-0.545, df=29, p=0.590
Home (EXP1)	4.12 (3.11)	
Lab (EXP2)	4.46 (2.57)	F(2,27)=1.454, p=0.232
Bus (EXP2)	4.46 (2.29)	
Station (EXP2)	4.75 (2.42)	

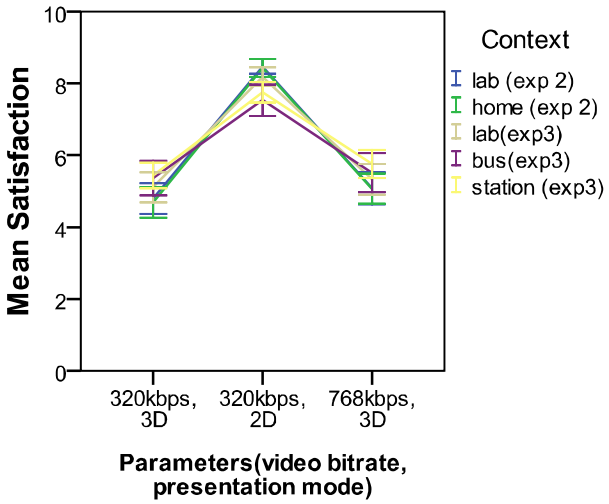

Figure 1. Interaction of parameters and context on mean overall satisfaction. The bars show 95% CI of mean.

Appropriateness of the evaluation technique – Utility of workload data-collection is to support the understanding of tasks that surround the quality evaluation in the field. For mobile (3D) television (used in extra time situations with easy viewing task, light parallel tasks when coping with surrounding and not under extremely harsh physical surrounding) it can be assumed that there should not be differences between the studied contexts. In the terms of complexity and effort, data-collection procedure is easy as is its analysis. We recommend including overall workload analysis as a part of quality evaluation experiments in the field.

4.5 Interview: experiences of quality in context

Results – The analysis is based on the interview of experiences of quality in the contexts from the experiment 2. Analysis of transcribed interviews was conducted following the principles of Grounded Theory framework through systematical steps of open coding, concept development and categorizing[38]. It is well-applicable in research areas with little prior knowledge, such as experienced quality in context, and when aiming at understanding the meaning or nature of a person's experiences[38]. A total of 15 participants were included into this data-analysis and one mention per participant for each subcategory was counted.

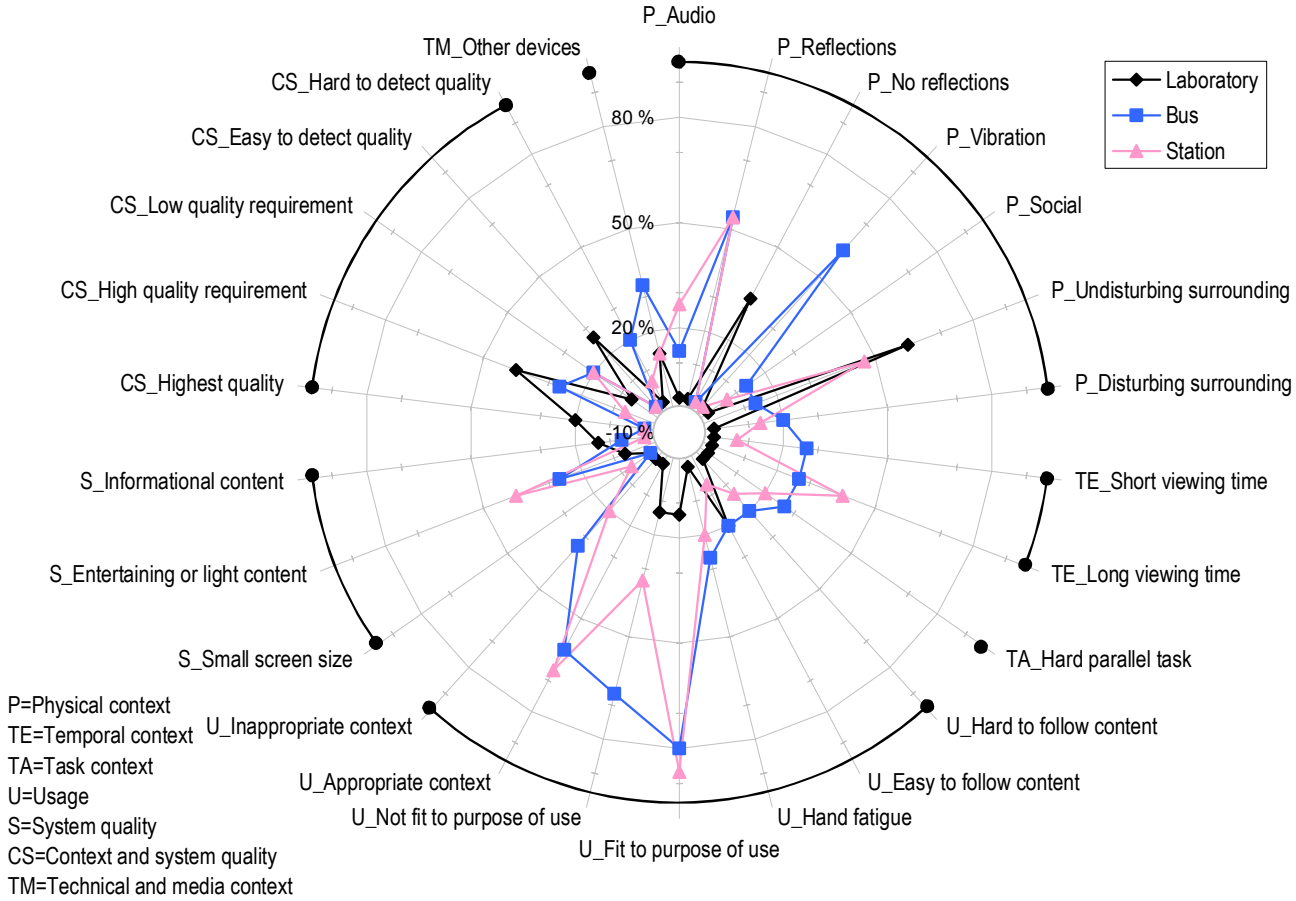

Figure 2. Experienced quality in the context. The scale shows the percentage of all participants mentioning that category.

Experienced quality in the context was described with the aid of seven main categories, called as: 1) physical context, 2) temporal context, 3) task context, 4) usage, 5) system quality, 6) context and system quality, and 7) technical and media context (Figure 2). Physical context consists of audio factors (noise, announcements), light reflections (from outside, from lighting) or lack thereof on the device screen, vibrations (trembling and movements of the bus), social (presence of other people), and disturbances (other than audio) or lack thereof from the surroundings. Temporal context consists of short and long viewing times (under or over 20 minutes in length). Task context consists of mentions of the hard parallel task making evaluating quality harder. Usage contains easy or hard concentration on viewing the content, hand fatigue from needing to hold the device, fittingness to the purpose of use, and appropriateness to the purpose of use (does the

participant encounter similar situations often). System quality contains mentions of too small screen size of the device and mentions of entertaining or light, and informational content. Context and system quality contains descriptions of quality being experienced as highest in a certain context, low or high demands for system quality due to the surrounding context and easiness or hardness of detecting quality differences. Finally, technical and media context describes the presence of other devices or services for accessing similar content.

The laboratory context was described as being a calm environment for viewing content, free from light reflections on the screen, but suffered from artificiality. It offered good circumstances for viewing, allowing easy following of content and detection of quality differences. Even though quality was described to be the highest there, the artificial settings caused the participants to feel more demanding towards quality.

The field conditions (bus and station context) were described with similar descriptions. Both received high numbers of mentions of distracting factors: reflections from light sources on the device screen, movements and presence of other people, and surrounding noise. In the bus context, the movements and vibrations of the bus provided extra distractions. In both contexts, the extra task (keeping an eye on the timetable or the correct bus stop) made the viewing task harder than in the laboratory. Both contexts were mentioned to be appropriate for viewing entertaining or light content, mostly suitable as viewing situations and mostly relevant to participants' lives.

Appropriateness of the evaluation technique – There are five main functions of interviews of quality in the context of use: It describes 1) subjectively important characteristics of the context of use, 2) quality requirements on a general level, 3) critical system factors which can influence on the usability of the system, 4) confirm and facilitate the creation of new design ideas, 5) underline needed improvements for further experiments. From the viewpoint of complexity and effort, the interview is easy to collect, but it is laborious to analyze. Based on good utility-cost ratio, we recommend including interviews as a necessary part of the quality evaluation experiments for understanding the quality characteristics[32] and quality in the context of use.

4.6 Situational data-characteristics

Results - The goal of the situational video analysis was to explore momentary changes of user's gaze, movements and actions and environment during the experiment. We focused on gaze and user's movements in the analysis. The gaze behavior reveals information about human attention as selective allocation processing resources (overview[35]). In the use of mobile applications in urban areas, attention is actively shared between device and environment[35]. We also explored user's movements needed to improve the viewing conditions.

We applied the following coding scheme in the video analysis done in an accuracy of one second: 1. General information: time (timestamp, accuracy of 1 sec), context (lab, bus, station), video (number of clip, or not playing). 2. Coded target of gaze: video, answer sheet, on timetable or watch, surroundings, other people, somewhere else (e.g. eyes closed). 3. User movements and actions: moving head and/or device, other movement, user-produced sound (talk, other), adjustments of volume on the device. 4. Environment: changes in physical environment (audio, lightning level, people near by (radius 2 m)), haptic environment (vibrations, acceleration/breaking of bus), bus movement (moving, not moving (at the bus stop, traffic lights)). Data-coding scheme was developed in an iterative process by two researchers prior to the start of the actual coding and a combination of tools (Microsoft Office Excel 2003, InqScribe 2.0.5) was used in the coding. The analysis included 12 participants with complete recordings in all three contexts

Analysis of gaze: Video viewing in the field circumstances shows active sharing of attention between the surroundings and viewing task, similarly to other mobile HCI tasks ([39], not walking[35]). Higher number of gaze shifts were done during the video playback in the field compared to the laboratory settings (Table 4: $F_R = 9.5$, $df = 2$, $p<.01$, lab vs. others $p<.05$; bus vs. lab $p>.05$). Similarly, the length of continuous gaze span (including the answer time) was shorter in the field ($F_R = 9.7$, $df = 2$, $p<.01$; Lab vs. others $p<.05$, Bus vs. station $p>.05$). In the field conditions (bus and station contexts), glances were usually either at the answering sheet or at the surroundings, while in the laboratory, the glances were mainly targeted on the answering sheet. In the field, participants used their time to watch the surroundings, navigate and orientate (bus) and keeping track of the time for catching the train (station) while there were no other stimuli or parallel tasks to capture users' attention in the laboratory. Surrounding actions such as a change in the movement of the bus (e.g. acceleration, slowing down, intensive trembling or a sharp turn), passing people in a close distance to user, particularly sharp and loud sounds in both contexts were connected to the gaze shifts.

Adjustments of viewing conditions: To improve the viewing conditions during the video playback, higher number of movements appeared in the field conditions (Table 4: $F_R = 10.7$, $df = 2$, $p<.01$). The result appeared for both 2D and 3D

visual presentation modes with equal bitrates of 320kbps (Figure 3: 2D - F_R =13.32, df=2, p<.01; 3D - F_R=8.35, df=2, p<.05). Within context comparisons also show a tendency where higher number of adjustments for improving the viewing conditions are needed for 3D presentation mode (Lab: Z=-1.96, p=.05; Bus: Z=-1.84, p=.07; Station: Z=-3.06, p<.01). In the field, participants held the device in their hands and were under variable movements, vibrations and lightning conditions which may result to more changes to maintain the optimal viewing position. In the laboratory, the device was placed in a stand. For further design, it is important that users can reach the sweet spot for 3D easily in the field after the multiple gaze shifts and natural movements.

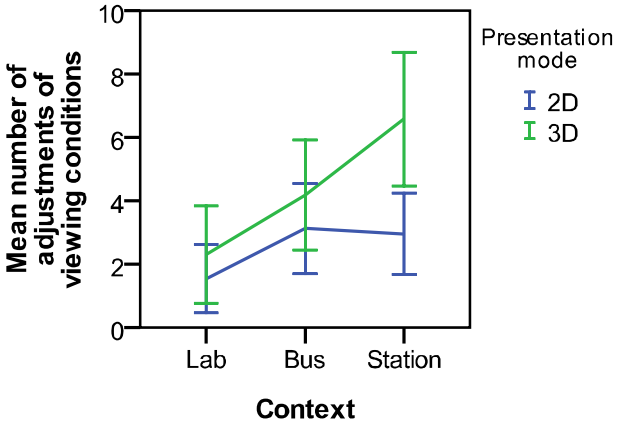

Figure 3. Mean number of adjustments of viewing conditions in different contexts and presentation modes.

Table 4. Situational analysis of gaze and movements.

Analyzed unit	**LAB** Mean(SD)	**BUS** Mean(SD)	**STATION** Mean(SD)
Gaze shifts during a video clip (n of times)	3.3 (0.9)	4.4 (1.3)	4.3 (1.4)
Continuous gaze span on screen (in seconds)	11.3 (3.1)	9.0 (2.9)	9.2 (3.1)
Adjustments of viewing conditions per clip (n of times)	2.6 (2.7)	4.2 (2.6)	5.5 (2.5)

Appropriateness of the evaluation technique – The utility of situational data is to provide deeper understanding on actual events in the field. The data can result in fundamental understanding on contextual behavior and bring out ideas for improvements of the system in a usability level and in the data-collection procedure. In the terms of complexity and effort, the data-collection procedure is relatively easy with modern light-weight camera systems. The required effort in the analysis depends on its depth. The analysis in the level of gaze patterns is laborious, but might not be appropriate for all field experiments. The automatic tools for gaze detection would decrease the effort in the analysis. From the viewpoint of gaze analysis, in further work, it would be desirable to know the natural patterns for gaze behavior with minimal obstruction of the evaluation task (e.g. non-repeated long stimuli) and explore the applicability of simulations of these patterns in the laboratory or analog circumstances. For further field experiments, we recommend the use of situational data collection while the depth of analysis can be decided based on the purpose of study.

5. DISCUSSION AND CONCLUSIONS

This paper presented a hybrid User-Centered Quality of Experience (UC-QoE) research methodological framework for evaluation of quality in the context of use. The novelty of the method was to underline the characterization of context of use as an important part of quality evaluation in the field circumstances. In this paper, we also described experiments where the method was used for evaluation of quality in three contexts of use parallel to the controlled laboratory assessment of mobile 3D television. Based on our analysis of appropriateness of different evaluation techniques as a part of the method, we recommend using a hybrid methodological procedure in the quality evaluations in the field including quantitative quality preference ratings, qualitative descriptions of quality and context, analysis of characteristics of context in a macro level and situational characteristics of context in a micro level. As the method described was especially targeted on the contextual quality evaluation for the mobile 3D television, applicability of the framework to the other application fields (quality in other types of applications, usability and user experience evaluation) needs to be addressed in future work.

ACKNOWLEDGEMENTS

This work is supported by the European Commission within the ICT program of FP7 under Grant 216503 with acronym MOBILE3DTV (http://www.mobile3dtv.eu/). Satu Jumisko-Pyykkö's work is funded by the User-Centered Information Technology (UCIT) graduate school. The authors wish to thank Cinovent, Red Star Studio, Stereoscape, and Centre of Computer Graphics and Visualization from the University of West Bohemia for providing stereoscopic content.

REFERENCES

[1] Winkler, S. and Faller, C., "Audiovisual quality evaluation of low-bitrate video," Proc. SPIE/IS&T Human Vision and Electronic Imaging, vol. 5666, San Jose, United States of America, 139-148 (2005).
[2] Knoche, H., de Meer, H. and Kirsh, D., "Extremely economical: how key frames affect consonant perception under different audio-visual skews," Proc. 16th World Congress on Ergonomics IEA2006, 6 pages, 10-14 July, Maastricht, The Netherlands (2006).
[3] Meilgaard, M. C., Civille, G. V. and Carr, B. T., [Sensory evaluation techniques], CRC Press, New York, NY, USA, (1999).
[4] ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunications Union – Radiocommunication sector, (2002).
[5] ITU-T P.911, "Recommendation P.911 Subjective audiovisual quality assessment methods for multimedia application," International Telecommunication Union – Telecommunication sector, (1998).
[6] Jumisko-Pyykkö, S. and Strohmeier, D., "Report on research methodologies for the experiments," Technical report, November 2008, (2008), http://sp.cs.tut.fi/mobile3dtv/results/tech/D4.2_Mobile3dtv_v2.0.pdf
[7] Gotchev, A., Jumisko-Pyykkö, S., Boev, A. and Strohmeier, D., "Mobile 3DTV System: Quality and User Perspective," Proc. 4th Mobimedia, Oulu, Finland, 7.-9.7.2008, 5 pages, (2008).
[8] Gotchev, A. , Smolic, A., Jumisko-Pyykkö, S., Strohmeier, D., Akar, G. B., Merkle, P. and Daskalov, N., "Mobile 3D television: Development of core technological elements and user-centered evaluation methods toward an optimized system". Proc. IST/SPIE Conference on Electronic Imaging, Volume 7256, 72560J (2009), 3D Video Delivery for Mobile Devices (2009), doi:10.1117/12.816728.
[9] Jumisko-Pyykkö, S. and Vainio, T., "Framing the context of use for mobile HCI," International Journal of Mobile-Human -Computer-Interaction (IJMHCI), in press.
[10] Oulasvirta, A., [Field experiments in HCI: Promises and challenges]. In P. Saariluoma, H. Isomaki (eds.), Future Interaction Design II, Springer, (2009).
[11] Abowd, G. and Mynatt, E., "Charting past, present and future research in ubiquitous computing," ACM Transactions on Computer-Human Interaction, Vol. 7(1), 29-58, (2000).
[12] Schilit, B. N. and Theimer, M. M., "Disseminating active map information to mobile hosts," IEEE Network, Vol. 8(5), 22-32, (1994).
[13] Kjeldskov, J. and Stage, J., "New techniques for usability evaluation of mobile systems," International Journal of Human-Computer Studies, Volume 60, Issues 5-6, HCI Issues in Mobile Computing, May 2004, 599-620, (2004), doi:10.1016/j.ijhcs.2003.11.001.
[14] Jumisko-Pyykkö, S. and Hannuksela, M. M., "Does context matter in quality evaluation of mobile television?" Proc. MobileHCI '08, ACM, New York, NY, 63-72, (2008), doi:10.1145/1409240.1409248.
[15] Knoche, H. and Sasse, M. A., "The big picture on small screens delivering acceptable video quality in mobile TV," ACM Trans. Multimedia Comput. Commun. Appl. (TOMCCAP) 5, 3 (Aug. 2009), 1-27 (2009), doi:10.1145/1556134.1556137.
[16] Jumisko-Pyykkö, S., Malamal Vadakital, V. K. and Hannuksela, M. M., "Acceptance threshold: bidimensional research method for user-oriented quality evaluation studies," International Journal of Digital Multimedia Broadcasting, Volume 2008 (2008), Article ID 712380, 20 pages, (2008), doi:10.1155/2008/712380.
[17] Wikstrand, G., "Improving user comprehension and entertainment in wireless streaming media, introducing cognitive quality of service," Department of Computer Science, Umeå University, Umeå, Sweden (2003).
[18] Boev, A., Hollosi, D. and Gotchev, A., "D5.1 Classification of stereoscopic artefacts," Technical report, July 2008, (2008), http://sp.cs.tut.fi/mobile3dtv/results/tech/D5.1_Mobile3DTV_v1.0.pdf

[19] Jumisko-Pyykkö, S., Häkkinen, J. and Nyman, G., "Experienced Quality Factors - Qualitative Evaluation Approach to Audiovisual Quality," Proc. IST/SPIE conference Electronic Imaging, 65070M (2007), Multimedia on Mobile Devices, (2007).
[20] Shadish, W., Cook, T. and Campbell, D., [Experimental and quasi-experimental designs], Houghton Mifflin, Boston, MA, (2002).
[21] Cook, T. and Campbell, D., [Quasi-experimentation: design & analysis issues for field settings], Houghton Mifflin, New York, (1979).
[22] Jambon, F., "User evaluation of mobile devices: in-situ versus laboratory experiments," International Journal of Mobile Computer-Human Interaction. 1, 2 (2009), 56-71, (2009).
[23] Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T. and Kankainen, A., [Will laboratory test results be valid in mobile contexts?] In: Joanna Lumsden (ed), Handbook of Research on User Interface Design and Evaluation for Mobile Technology, chapter LIII, 897-909, Information Science Reference, (2008).
[24] Mäki, J., "Finnish mobile TV results," Research International Finland, August 2005, (2005).
[25] Oksman, V., Ollikainen, V., Noppari, E., Herrero, C. and Tammela, A., "'Podracing': experimenting with mobile TV content consumption and delivery methods," Multimedia Systems, Vol 14 (2), 105-114, (2008).
[26] Södergård, C. (ed), [Mobile television – technology and user experiences], Report on the Mobile –TV Project. Espoo: VTT Publications 506, (2003).
[27] Cui, Y., Chipchase, J. and Jung, Y., "Personal television: a qualitative study of mobile TV users," Lecture Notes in Computer Science. Vol. 4471, 195-204, (2006).
[28] O'Hara, K., Mitchell, A. S. and Vorbau, A., "Consuming video on mobile devices". Proc. CHI '07, ACM, New York, NY, 857-866, (2007).
[29] Jumisko-Pyykkö, S., Weitzel, M. and Strohmeier, D., "Designing for user experience: what to expect from mobile 3D TV and video?" Proc. 1st international conference on Designing interactive user experiences for TV and video, October 22-24, 2008, Silicon Valley, California, USA, UXTV '08, vol. 291. ACM, New York, NY, 183-19 (2008), doi:10.1145/1453805.1453841.
[30] Carlsson, C. and Walden, P., "Mobile TV - To Live or Die by Content," Proc. 40th HICSS, IEEE Computer Society, Washington, DC, 51, (2007), doi:http://dx.doi.org/10.1109/HICSS.2007.382 .
[31] Oksman, V., Noppari, E., Tammela, A., Mäkinen, M. and Ollikainen, V., "News in mobiles. Comparing text, audio and video," VTT, (2007), http://www.vtt.fi/inf/pdf/tiedotteet/2007/T2375.pdf
[32] Jumisko-Pyykkö, S. and Utriainen, T., "D4.4 v2.0 Results of the user-centred quality evaluation experiments", Technical report, November 2009, (2009).
[33] Hart, S. G. and Staveland, L. E., [Development of NASA-TLX (Task Load Index): results of empirical and theoretical research]. In Hancock, P. A. & Meshkati, N. (eds.), Human Mental Workload, 139-183, North-Holland, Amsterdam, (1988).
[34] Oulasvirta, A. and Nyyssönen, T., "Flexible Hardware Configurations for Studying Mobile Usability," Journal of Usability Studies, Volume 4, Issue 2, Feb 2009, 93-105, (2009).
[35] Oulasvirta, A., Tamminen, S., Roto, V. and Kuorelahti, J., "Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI," Proc. CHI '05, ACM, New York, NY, 919-928, doi:10.1145/1054972.1055101.
[36] Rogers, E. M., [Diffusion of Innovations], 5th edition, Free Press, New York, NY, USA, (2003).
[37] Knoche, H., McCarthy, J. and Sasse, A., "Can Small Be Beautiful? Assessing Image Resolution Requirements for Mobile TV," Proc. of the 13th annual ACM international conference on Multimedia, 829-838, (2005).
[38] Strauss, A. and Corbin, J., [Basics of qualitative research: techniques and procedures for developing grounded theory], 2nd edition, Sage, Thousand Oaks, CA, (1998).
[39] Chen, T., Yesilada, Y. and Harper, S., "RIAM D2.6: How do people use their mobile phones while they are walking? A field study of real-world small device usage (EPSRC-EP/E002218/1)," Research Report School of Computer Science, University of Manchester, November 2008, (2008), http://hcw-eprints.cs.man.ac.uk/98/1/RIAM_D2_6_Field_Study.pdf