Report on research methodologies for the experiments
Satu Jumisko-Pyykkö  ■  Dominik Strohmeier
Report on research methodologies for the experiments

Satu Jumisko-Pyykkö, Dominik Strohmeier

Abstract: The report presents a framework of a user-centered quality evaluation method. Current recommended research methods based on a psychoperceptual approach do not take into quality evaluation these factors which are related to real usage in contexts of the system under test. Our method of which a first framework is presented here takes into account the user requirements to check optimization of the system. We first present a literature review of existing research methods and first user-centered approaches in audiovisual quality research and discuss the methods. Then, we present the holistic framework of our method. We will discuss how the main factors of user experience – user, system and context – can be evaluated and integrated in the framework. The framework is the basis for the upcoming experiments within the project and will be adapted then. Finally, we present personas and use scenarios for mobile 3D television and video to also define users and contexts within the user-centered research of MOBILE3DTV continuing the user-centered approach which we have taken since the beginning of the project.

Keywords: Research methodology, user-centered quality evaluation, user experience
Executive Summary

Subjective quality evaluation is a widely used method in the optimization of system performance as a part of end-products. Conventionally used psychoperceptual evaluation methods do not relate the evaluation to factors of the potential use of the system or service. It is often desirable to know whether the quality of the components of system meets the user requirements and that the performance of system is not over or under optimized.

- The major aim of this report is to present a framework for holistic user-centered (audio) visual quality evaluation method.

- Minor aims of this report are to describe the overall plan for the content of the experiments and continue the work of user-center design for mobile 3DTV by describing the relevant use cases.

The first section reviews existing psychoperceptual research methods for quality evaluation. The section reviews existing guidelines for quality assessments and compares them. Then, the different application fields of the different standards are reviewed within 3DTV research. The section is concluded with a discussion of the methods following publication reviews. After the preceding review of studies, this part focuses on a more general discussion about applicability of the methods in 3DTV research.

The section 3 focuses on a review of existing user-centered approaches. The review is mainly based on the user-centered quality evaluation in mobile TV and video to initiate the work in mobile 3DTV and video. User-centered approaches are presented are quality of perception, acceptance threshold and presence as a methods in the behavioural level of evaluation. We also review methods to describe ergonomic aspects of the system and finally methods of descriptive quality. The section reviews both methods as well as studies in the context of mobile TV and 3DTV where the methods were used successfully.

Section 4 includes the development of user-centered quality evaluation method for mobile 3DTV which has two parts. Firstly, we present seven principles for starting points for the development of method. 1) Quality perception is an active process combining different levels of human information processing. 2) Component user experience examines quality of critical system component by reflecting the factors of whole user experience. 3) Quality evaluation experiments are part of human-centered design process. 4) Quality evaluation for system optimization in product development has special characteristics. 5) Positive quality – good enough for use – is in our focus. 6) Low produced quality, noticeable impairments and heterogeneous stimuli requires guaranteeing the level of minimum acceptable quality. 7) Overall quality evaluation approach is suitable for user-oriented quality evaluation.

Secondly, we describe the overview to the holistic framework of user-centered quality evaluation method. The main principles of the method, following the factors of user experience, are 1) user, 2) system/service (including system and content), 3) context, 4) task (including measures of quality preferences in relation to action, impressions of quality and ergonomic approach to quality). For each of these factors our method describes aim and role in the evaluation procedure, how they should be taken into account and why, and finally the difference to the previous work. The holistic framework will be adapted in the upcoming experiments in the project.

We express an overall plan for series of the subjective quality evaluation experiments in section 5. The plan aims at describing 1) the content of each of the experiments and schedule for conducting the experiments, 2) that critical parts of the system will be covered in the experiments, and 3) psychoperceptual and user-centered quality evaluation studies. The detailed plans for each of the
experiments including the final variable selection and method will be documented at the time of planning a particular experiment.

The last section of this report presents use scenarios for mobile 3D television and video. After the development of the new methodology and the description of planned experiments, these use scenarios set up a framework of users and contexts for mobile 3DTV research. This section includes the results of our previous studies about user needs and expectation for mobile 3D television and video and brings these results into a common picture for all researchers within the project, but also to a concrete presentation of future end-users and usage scenarios for public understanding of mobile 3DTV research. The section thereby introduces usage scenarios and personas on a theoretical level and the presents the five different personas and their scenarios that have been development to be included into the project.
**Table of Contents**

Executive Summary .......................................................................................................................... 2

Table of Contents .......................................................................................................................... 4

1. Introduction ............................................................................................................................. 4
   1.1. Definitions of the key concepts of the report ..................................................................... 6
   1.2. Criteria for method development ..................................................................................... 6

2. Related work: Psychoperceptual approach .............................................................................. 9
   2.1. Main approaches and their general description ................................................................. 9
      2.1.1 ITU-Recommendations ............................................................................................... 9
      2.1.2. The EBU method: SAMVIQ ..................................................................................... 11
   2.2. Where have the methods been used? ................................................................................ 12
      2.2.1. Double-stimulus methods ....................................................................................... 12
      2.2.2. Single-stimulus methods ......................................................................................... 12
   2.3. Discussion of the methods ............................................................................................... 13
      2.3.1. Influence of context in double-stimulus methods ...................................................... 13
      2.3.2. Recency effect and duration neglect ........................................................................ 14
      2.3.3. Critique on single-stimulus continuous methods ...................................................... 15
      2.3.4. Brotherton: ACR vs. SAMVIQ .............................................................................. 15

3. Related work: User-centered approaches .............................................................................. 17
   3.1. Standardization activities of ETSI HF STF 354 and ITU-T SG12 ...................................... 17
   3.2. Evaluation methods on behavioral level .......................................................................... 17
      3.2.1. Quality of Perception ............................................................................................... 17
      3.2.2. Acceptance threshold: Uni- and bidimensional measures ........................................ 17
      3.2.3. Measuring the ‘being there’: The concept of Presence ........................................... 18
      3.2.4. Evaluation in the context of use: Quasi-experimental approach ............................ 18
   3.3. Ergonomic approach: Measures of cybersickness as part of quality evaluation ........... 19
      3.3.1. Simulator Sickness .................................................................................................. 20
      3.3.2. Recent studies investigating autostereoscopic displays ........................................... 20
   3.4. Descriptive quality ........................................................................................................... 20
      3.4.1. Qualitative interviews as a part of quality evaluation: Experienced quality factors .... 20
      3.4.2. Sensory Profiling and Mixed Methods approaches .................................................. 21

4. Development of user-centered quality evaluation method .................................................... 25
   4.1. Starting points for developing user-centered quality evaluation method for product development purposes .................................................................................................................. 25
   4.2. Evaluation of experienced (audio)visual quality / User-centered quality evaluation .......... 30

5. Plan for the content of the experiments .................................................................................. 35
5.1. Psychoperceptual experiments

Experiment 1: Impact of different stereo video capture, representation and coding factors on experienced visual quality

Experiment 2: Impact of channel and display factors on experienced visual quality

5.2. User-centered quality evaluation experiments

Experiment 1: Comparison of experienced audiovisual quality of 2D vs. 3D in the controlled and field settings

Experiment 2: Impact of coding and transmission factors on experienced quality of different user groups

Experiment 3: Field testing of the prototype

6. Use Scenarios for mobile 3D television and video

6.1. General considerations

6.2. Personas in MOBILE3DTV

6.2.1. Scenarios

6.2.2. Personas and Scenarios in MOBILE 3DTV

7. References
1. Introduction
This section presents the key concept and criteria for designing user-centered quality evaluation method for MOBILE 3DTV and video evaluation.

1.1. Definitions of the key concepts of the report

- **Psychoperceptual quality evaluation** is a method for examine the relation between physical stimuli and sensorial experience following the methods of experimental research. Origins of these methods roots to classical psychophysiological measures developed in 19th century (Lawless & Heyman 1998) and later these are applied in audio and video quality assessment (e.g. ITU recommendations). Psychoperceptual evaluation of audio, video or audiovisual quality is characterized in the terms of high-level control over the variables and test environment. These include standardized test sequences, procedure and categorization of participants to naive and professional evaluators. The given evaluation tasks judges the overall quality or a certain quality attribute with the full attention in the evaluation task.

- **User-centered quality evaluation** is a collection of factors and methods that relates the evaluation assessment to the potential use of system or service. 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) quality evaluation tasks which are in relation to expected goals of viewing, aims also at understanding the interpretation of quality and includes ergonomic measures. The method is called as user-centered quality evaluation method if any of the four listed factors relating the evaluation to potential use is taken into account in the evaluation research.

- **User centered design (UCD)** is a broad umbrella covering approaches such as traditional human factors and ergonomics, participatory design, human-centered design processes and design for user experience’ (Keinonen 2008). According to Usability Professionals’ Association UPA (2008) defines UCD as ‘an approach to design that grounds the process in information about the people who will use the product. UCD processes focus on users through the planning, design and development of a product.’

- **Multimedia quality** (Jumisko-Pyykkö et al. 2008): Multimedia quality is a combination of produced and perceived quality. Produced quality describes the technical factors of multimedia which can be categorized into three different abstraction levels, called network, media, and content. Perceived quality represents user's or consumer's side of multimedia quality, which is characterized by active perceptual processes, including low-level sensorial and high-level cognitive processes. A typical problem in multimedia quality studies is to optimize quality factors produced under strict technical constraints or resources with as little negative perceptual effects as possible.

- **The experienced audiovisual quality** (Jumisko-Pyykkö et al. 2008) is an integrated set of audio, visual and their interaction specialized audiovisual perceptual aspects which characteristics dependents on presented material, and excellence or distraction of aspects is evaluated according to the goals of usage.
- **Image quality** (Engeldrum X): Measurement of image quality is image evaluation. Objective image evaluation involves physically measuring image quality components with instruments. Subjective image evaluation involves obtaining human judgements of various aspects of image quality.

- **Perceptual measurement** (Bech & Zacharov 2006) is an objective quantification of sensorial strength of individual auditory attributes of perceived stimulus.

- **Affective measurement** (Bech & Zacharov 2006) is an objective quantification of an overall impression of stimuli. This assumes that subject is in some form of integrative set of mind where the influence of the impression for individual attributes, the context, mood, expectations, previous experience, traditions so on are all combined into one single impressions that establishes basis for some form of action of the listener.

- **Subjective evaluation** is based on human judgments of various aspects of experienced material based on perception [adapted from Engeldrum 2000]. Quality can be defined as an integrated set of perceptions of overall excellence of multimedia material [Engeldrum 2000].

- **Subjective quality**: "Any information that originates from users, experts or observers can be considered to be subjective data." [Mullin et al. 2002]. Perceptual evaluation is used as a synonym with subjective evaluation (Zou & Corriveau 1997).

- **Sensory evaluation** is a scientific discipline used to evoke, measure, analyse and interpret reactions to those characteristics of food and materials as they are perceived by senses of light, smell, taste, touch and hearing. (Stone & Sidel 2004)

- **Acceptance testing**: can be combined with other sensory analysis, knowledge of consumer expectations, and product formulation constrains in determining the optimal design for food products (moskowitz, bech et al. 1994) [Lawless & Heymann 1998]

- ** Appropriateness studies**: Appropriateness ratings can be used to assess the effects of context associated with hedonic responses of food (Schultz, 1998) [lawless & heyman p. 467]

- **Evaluation as research**
  - For example, Clarke 2000 defines evaluation research as a form of applied research aiming at producing information about the implementation, operation and ultimate effectiveness of policies and programs designed to bring about change (Clarke 2000 Evaluation research)

  - Evaluation is the general process of judging the worthwhileness of activity regardless of the method employed, and evaluation research is the specific use of the scientific method of the purpose of making an evaluation (Suchman 1967).

  - The functions of evaluations are typically targeted on different program states or functions of evaluations. The program stated can either be the process or its outcome (Chen 1996 in Clarke 2000). Functions of evaluations can be either improvement or assessment (Chen 1996 in Clarke 2000).
1.2. **Criteria for method development**

The principles of good research provides a significant list of details, described in high-level, to be taken into account in the method development. According to Haslam & McGarty (2003) the factors are: 1) Reliability, 2) Validity 3) Cumulativity, 4) Parsimonious, and 5) Public Replication. A study, as a piece of empirical research, is characterized by being both process and a product (Haslam & McGarty 2003). From this broad viewpoint, aim of our research methodology is especially to help for conducting the process of research in a way that the goals of good research are filled. Among the many goals of good research, we present in detail aspects of validity which are very central in the development of research methodology.

**Validity** describes extend to which a given finding shows what it is believed to show. A valid finding is one that has been logically and correctly interpreted (Haslam & McGarty 2003). Validity can be categorised into three different groups – to internal, external, and construct validity.

- **Internal validity** is meant to check whether the independent variables are related to the dependant variables enabling to draw the conclusion of causal impact ($X$). In experimental research, like quality evaluation research, the control is the central factor in internal validity. According to Cambpel and Stanley (1966) good internal validity is the most important class of the validity for the experimental research. If the requirement of the good internal validity is not reached in the experimental research the results are unpredictable. Satisfactory research has both good internal and external validity (Shadish, Cook & Campell, 2002).

- **External validity** examines in what extend the research can be generalized into several aspects of research from sample, settings, researcher, and materials to time (Haslam & McGarty 2003.)

- **Construct validity** describes the theoretical accuracy of the research. Researcher needs to be critical to estimate weather the research has arrived at the correct explanation for any cause-and-effect relation that study has demonstrated. Internal validity is the requirements for the construct validity. (Haslam & McGarty 2003.)

In addition to listing the different forms of validity, **threats of validity** are listed in several sources (Campbell and Stanley (1966), Cook and Cambell (1979), Shadish et al 2002):

1. Selection bias: Assignment of initially nonequivalent participants to the groups being compared
2. Selective drop-out: Nonrandom, systematically biased loss of participants in the course of the study
3. History: Potentially important events occurring between early and later measurements in addition to the independent variables being studied
4. Maturation: Naturally occurring changes in the participants as a function of the passage of time during the study
5. Testing: Effects of taking a test upon a performance on a later test
6. Reactivity: Unintended effects of the experimental arrangements upon participants’ responses
7. Instrumentation: Unintended changes in experiments, observers, or measuring instruments in the course of the study
8. Statistical regression: Tendency of initially extreme scores to move toward the group mean upon retesting
9. Low reliability: Errors of measurement in the assessment of independent variables
10. Low statistical power: Low probability of detecting genuine effects because of characteristics of the design and statistical tests
11. Mono-operation bias: Use of a single operationalization of either the independent or dependent variable
12. Mono-method bias: Use of a single experimental method for examining possible relations between the independent and dependent variables

2. Related work: Psychoperceptual approach

2.1. Main approaches and their general description

2.1.1 ITU-Recommendations

Evaluating subjective video quality according to standardized methods has got a long history in quality assessments. Already in 1974, the International Telecommunication Union (ITU) published their recommendation ITU-R BT.500 (2002) "Methodology for the subjective assessment of the quality of television pictures". Until now, this recommendation has been revised several times and ITU-R BT.500-11 (2002) is still common in the field of video quality evaluation. It offers a collection of methods that are applicable for assessment tests. The three most prominent methods of ITU-R BT.500-11 are the double-stimulus impairment scale (DSIS), the double-stimulus continuous quality-scale (DSCQS), and the single stimulus continuous quality evaluation (SSCQE). While DSIS and DSCQS are based on direct comparison of impaired and unimpaired stimuli, SSCQE offers guidelines to assess the perceived quality of a single stimulus without a source reference. With respect to the assessment of stereoscopic image quality, ITU-R BT.1438 (2000) states that the methods of ITU-R BT.500-11 (2002) are also applicable in case of quality evaluation of stereoscopic images or videos. Goal of all these methods and the additional ones that are discussed in detail below is to create a preference order of test sequences to be able to make conclusions about the impact of test parameters on subjective video quality.

In DSIS Variant II which is recommended when “moving sequences are under test” (ITU-R BT.500-11 2002) short clips of the unimpaired and an impaired stimulus with a length of 10 seconds are presented one by one twice. The test participants are asked to rate the impact of impairment on their subjective quality perception on a 5-grade impairment scale reaching from "1" (very annoying) to "5" (imperceptible). The presentation structure is illustrated in Figure 1, the DSIS grading scale can be found in Figure 2.

DSCQS also consists of the same presentation structure like DSIS (cf. Figure 1). Two sequences are presented one after the other with one repetition. But in contrast to DSIS, test participants are asked to rate the quality of each of the two stimuli on a 5-point continuous quality scale reaching from “bad” to “excellent” as can be seen in Figure 2. This method is especially useful, when “it is not possible to provide test stimulus test conditions that exhibit the full range of quality.” (ITU-R BT.500-11 2002) During analysis, the scores on each of the scales are normalized to a range of 0 to 100 and the difference between the scores of reference and test condition is calculated. ITU-R BT.500 (2002) points out that this form of assessment only refers to the difference between reference and test condition and that “it is erroneous to associate the scores with a single quality description term.” (ITU-R BT.500-11 2002)
In contrast to double-stimulus methods like DSIS or DSCQS, ITU-R BT.500 also recommends single-stimulus methods like SSCQE. This method differs in two aspects from the double-stimulus methods mentioned above. Beside the fact that the quality of only one stimulus is rated without reference, SSCQE comprises a continuous quality assessment. While perceived quality is rated retrospectively in DSIS or DSCQS, test participants evaluate the quality of a stimulus with the help of a slider along a similar scale as in DSCQS (cf. Figure 2 5-grade impairment scale of DSIS (left) and 5-point continuous quality scale of DSCQS (right) according to ITU-R BT.500-11 (2002), while they watch the test sequences. This method was developed to provide a possibility which can be used in more realistic home viewing-like conditions. Its goal is to provide possibilities to assess the subjective quality of longer scenes to capture scene-dependent and time-varying effects of impairments.

Besides ITU-R BT.500 for assessing television pictures and the additional standard ITU-R BT.1438 (2000) for stereoscopic cases, ITU-R P.910 (1999) and ITU-R P.911 (1999) offer additional test methods for assessing subjective video and audiovisual quality of multimedia systems. Especially the application of audiovisual quality in ITU-R P.911 (1999) is remarkable as ITU-R BT.500-11 (2002), ITU-R BT.1438 (2000), and ITU-T P.910 (1999) do not include any additional audio in the sequences under test. An important research method, which is included in ITU-T P.910 (1999) and ITU-T P.911 (1999), is Absolute Category Rating (ACR). Like SSCQE, ACR is a single-stimulus method. But while SSCQE captures subjective quality continuously, ACR is used for retrospective quality assessment. In ACR, test sequences are watched one by one without reference and are judged subsequently as depicted in Figure 3. For quality judgment, ITU-R P.910 (1999) and ITU-R P.911 (1999) propose a five-level quality scale similar to DSCQS (cf. Figure 2). However, the recommendations also outline that more detailed scales (9- or 11-level quality scales) can be used if higher discrimination power is needed.

The parameters of all methods recommended by the ITU are concluded in the following table regarding representation order, quality rating, scales, and the duration of sequences under test.
2.1.2. The EBU method: SAMVIQ

The Subjective Assessment Methodology for Video Quality (SAMVIQ) (Blin 2006)(Kozamernik et al. 2005) was developed as a derivate of ITU’s DSCQS method (ITU-R BT.500-11) to offer a test methodology for multimedia. Blin (2006) describes SAMVIQ to be an efficient way in the assessment of a large range of image quality as it provides reliable discrimination at both high and low quality levels. SAMVIQ uses hidden and explicit references in a multi stimulus test environment. In contrast to DSCQS the test participant has got the possibility to access more than two stimuli at the same time. Blin (2006) states that this direct comparison of multiple stimuli makes SAMVIQ to be “able to discriminate low qualities as well as high qualities”.

The test methodology is based on a multi-stimulus approach (Kozamernik et al. 2005). All stimuli are evaluated each after the other on a continuous scale from 0 to 100 with 5 explicit quality levels (Excellent, good, fair, poor, bad). Each stimulus is thereby compared to an explicit reference which determines the best quality that can be achieved in the test. But in contrast to DSCQS not only pairs, but all stimuli are available all time and they can be repeated and reevaluated as often as needed. Additionally to the explicit reference there is a hidden reference which is of the same quality level as the explicit one, but it is not outlined in the test description. So the test participants have to find this hidden one as an explicit part of the test.

With respect to stimuli Blin (2006) and Kozamernik et al. (2005) describe that stimuli of a maximum length of 15 seconds are sufficient to get a stabilized and reliable quality score. The performance of SAMVIQ method was tested in words of reliability and stability (Blin 2006). By comparing the standard errors of a SAMVIQ study to those of a DSCQS based study with the same stimuli under test. Blin (2006) shows that SAMVIQ gives better results with lower standard deviations. The second test was carried out to test the stability of the test results which refers to a comparison of results of a SAMVIQ study that was conducted with two independent groups. The results of both groups show very high correlation.

<table>
<thead>
<tr>
<th>References</th>
<th>DSIS</th>
<th>DSCQS</th>
<th>SSCQE</th>
<th>ACR</th>
<th>SAMVIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double stimulus</td>
<td>Double stimulus</td>
<td>Double stimulus</td>
<td>Single stimulus</td>
<td>Single stimulus</td>
<td>Multi stimulus</td>
</tr>
<tr>
<td>Moment of rating</td>
<td>Retrospective</td>
<td>Retrospective</td>
<td>Continuous</td>
<td>Retrospective</td>
<td>Retrospective – rating can be adapted several times</td>
</tr>
<tr>
<td>Scale</td>
<td>5-grade impairment scale</td>
<td>5-point continuous scale</td>
<td>5-point continuous scale</td>
<td>5-point continuous scale (or higher if required)</td>
<td>5-point continuous scale</td>
</tr>
<tr>
<td>Length of stimuli</td>
<td>10 seconds</td>
<td>10 seconds</td>
<td>Long stimuli (&gt;60 seconds) up to 20 minutes</td>
<td>10 seconds</td>
<td>Maximum 15 seconds</td>
</tr>
</tbody>
</table>

2.2. Where have the methods been used?
The following section will give an overview of cases in which the previously discussed methods of subjective video quality assessment were used in stereoscopic research. We show that both double-stimulus and single-stimulus methods have been used in studies to evaluate subjective stereoscopic video quality. We present an overview regarding also selection of participants, stimuli used and description of test environment.

2.2.1. Double-stimulus methods
Stelmach et al. (2000) studied the perceived quality of mixed-resolution stereo-video sequences and the impact of mixed resolution on perceived quality, depth, and sharpness. The test sequences, namely “Flower Pot” and “Trapeze”, had 720x480 pixels in size at 60Hz interlaced. The test sequences all had a length of 10 seconds and were presented using shutter glasses. The research method used was DSCQS according to ITU recommendations. Two groups of participants took part in the study. One group rated “perceived image quality”; the other group rated “perceived sharpness and perceived depth” (Stelmach et al. 2000). Each group consisted of 21 university students who were all screened for visual acuity, color vision, contrast sensitivity and stereo depth perception.

In another study, Stelmach et al. (2000) examined the impact of asymmetrical quantization and low-pass filtering of left and right view in stereoscopic video sequences. Again, they used DSCQS according to ITU-R BT.500-11 (2002) as research method. Three groups of participants with 21, 26, and 23 members evaluated separately impact of asymmetrical low-pass filtering, quantization and the combination of both, respectively.

Aksay et al. (2005) used DSCQS to evaluate the impact on subjective image quality if color channels of the stereoscopic video sequences are downsampled in order to compress the videos. In two experiments, they downsampled Y, U and V channel of different video sequences (tsu-kuba, sawtooth, venus, acrobats, mountain, teddy bear, touring, purple car, and truck). In the first part, there were 16 participants aged from 20 to 25, in the second experiment, there were 17 participants aged from 19 to 30. Color vision, stereo depth perception and visual acuity of the participants were screened before the experiments. According to Aksay et al., all participants can be regarded as “non-experts in the area of picture quality.” (Aksay et al. 2005)

2.2.2. Single-stimulus methods
Single-stimulus evaluation methods were developed within the European RACE project MOSAIC (Hamberg and de Ridder 1995) to close the gap of retrospective video assessment and variations of perceived quality “as a function of scene content and time” (Aldridge et al. 1998).

Following we present two studies that used SSCQE to evaluate stereoscopic video, both in the context of 3DTV. It is remarkable that the two studies examined completely different attributes of stereoscopic images. Focusing on visual fatigue, Yano et al. (2004) used SSCQE method to study the impact of content factors on visual fatigue in 3D HDTV images. Therefore, test participants watched 3D sequences with a length of 15 minutes and 10 minutes (“Waffen” and “Africa”) in both stereoscopic and non-stereoscopic case. Participants rated the experience of visual comfort on a continuous 5-point scale where 5 corresponded to “good” and 1 to “bad” comfort (Yano et al. 2004). In the analysis, Yano et al. then used negative peaks in the SSCQE graphs to analyze the content at the corresponding moment regarding “only the amount of motion and parallax on the test images.” (Yano et al. 2004)

Interestingly, IJsselsteijn et al. (1998) already introduced an adaptation of the SSCQE method in 1998. They used SSCQE method to evaluate the sense of presence in 3DTV sequences. In this study, 12 participants (all participants had normal or corrected-to-normal vision) watched stereoscopic video sequences using polarized glasses and evaluated three criteria in independent, consecutive repetitions. The evaluation criteria that were assessed by the participants were depth, naturalness of depth and pre-sence. Interestingly, IJsselsteijn et al. did not follow completely the ITU recommendations for SSCQE. Instead of evaluating moment-by-moment quality on a continuous
quality scale. IJsselsteijn et al. asked the participants just to move the slider upwards if the experience of the evaluation criterion increased and to move the slider downwards if the experience of the criterion decreased again. In this case, they use the first derivation of the continuous quality graph to run analysis of the data as one can find the “rate and direction of the subjective judgements of the observers.” (IJsselsteijn et al. 1998) This adaptation will be discussed in the next section.

Although Seutiëns et al. don’t refer to ITU recommendations in (Seuntiëns et al. 2007), the method of their study follows the recommendations for Absolute Category Rating. They used this single-stimulus, retrospective method to “investigate the concepts naturalness, viewing experience, and presence in relation to image quality, depth, and Ambilight.” (Seuntiëns et al. 2007) All 20 participants were screened for color and stereo vision. The participants had to rate two different HD video sequences (“Balloon” and “Adventure” (Seuntiëns et al. 2007)) with regard to six attributes, namely perceived image quality, perceived depth, perceived Ambilight, naturalness, viewing experience, and presence on a 5-point numerical scale.

The following table finally presents a summarizing overview of the studies presented above.

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>Scale</th>
<th>Applied for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stelmach et al.</td>
<td>DSCQS (ITU-R BT.500-11)</td>
<td></td>
<td>Impact of mixed resolution on perceived quality, depth, and sharpness</td>
</tr>
<tr>
<td>(Stelmach et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stelmach et al.</td>
<td>DSCQS (ITU-R BT.500-11)</td>
<td></td>
<td>Impact of asymmetrical quantization and low-pass filtering</td>
</tr>
<tr>
<td>(Stelmach et al.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aksay et al.</td>
<td>DSCQS (ITU-R BT.500-11)</td>
<td></td>
<td>Impact of downsampled color channels</td>
</tr>
<tr>
<td>(Aksay et al.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yano et al.</td>
<td>SSCQE (ITU-R BT.500-11)</td>
<td></td>
<td>Study on content features influencing visual fatigue</td>
</tr>
<tr>
<td>(Yano et al. 2004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IJsselsteijn et</td>
<td>Adaptation of SSCQE</td>
<td></td>
<td>Sense of presence</td>
</tr>
<tr>
<td>al. (IJsselsteijn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>et al. 1998)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seutiëns et al.</td>
<td>ACR, although to</td>
<td>5-point numerical scale (ITU-R</td>
<td>Naturalness, presence, and viewing experience in relation to image quality,</td>
</tr>
<tr>
<td>(Seuntiëns et al.</td>
<td>referred</td>
<td>BT.500-11)</td>
<td>depth, and Ambilight</td>
</tr>
<tr>
<td>2007)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Overview of studies into subjective stereoscopic image quality evaluation and the method applied

### 2.3. Discussion of the methods

#### 2.3.1. Influence of context in double-stimulus methods

Corriveau et al. (1999) studied the influence of context on different methods and the relating different scales. Context thereby means the dependency of ratings from the presentation order of the test sequences. Impairment of the preceding test sequence could affect the rating of the actual test sequence. Usually, researchers try to eliminate this effect by randomizing different presentation orders. Corriveau et al. (1999) investigated this effect in the double-stimulus assessment methods.
DSIS Variant II, DSCQS and a “Comparison method” where the two stimuli were compared on a scale from “A is much better” to “A=B” and “B is much better” (Corriveau et al. 1999). In their study, Corriveau et al. (1999) showed that all three methods under comparison track the perceived subjective quality in a similar manner. However, they found that DSCQS only shows slight contextual dependencies. In contrast, contextual effects are clearly present for the comparison method and highly for DSIS Variant II. They hence suggest that DSCQS should be the preferred method “to minimize contextual effect for subjective picture quality assessment” [7].

2.3.2. Recency effect and duration neglect

The recency effect in subjective quality assessment was first described by Aldridge et al. (Aldridge et al. 1995-2) in 1995. In their study, Aldridge et al. studied the applicability of 30 seconds long test sequences in subjective television picture quality assessment. Using DSCQS method participants had to evaluate the quality of two 30 seconds test sequences, whereby both sequences had included a highly impaired short clip (C), one sequence at the end (ABC), the other sequence at the beginning (CDE). Previous evaluation of each sequence separately had shown that participants strongly perceive decreased quality for clip C. However, the analysis of the results of DSCQS study showed that there were significant differences in the ratings for C, CDE, and ABC. Aldridge et al. formulate that “when an impaired section of video was followed by a section of relatively high-quality video, subjects tended to ‘forgive’ the bas section by a kind of averaging of quality over the period of the sequence. However, where good-quality video precedes poor-quality video, subjects appear to rate the sequence as a whole on the basis of the poor-quality end section alone, apparently with little regard to the earlier material.” (Aldridge et al. 1995-2) Aldridge et al. refer to the length of human working memory as a possible explanation for this recency effect. The recency effect in subjective quality assessment was corroborated by Hands and Avons (2001). According to Hands and Avons, the recency effect can be attributed to Hogarth and Einhorn’s belief-adjustment model (Hogarth and Einhorn 1989). Quality estimation is an adaptive process where a current level of perceived quality is revised in the moment that this quality level changes. Quality estimation is hence a “step-by-step strategy” integrating different levels of quality into one (current) quality perception. (Hands and Avons 2001)

Beside the recency effect, there is another interesting effect that can be found in retrospective quality assessment methods. After continuous assessment of subjective video quality was introduced by Hamberg and de Ridder (1995), Hands and Avons (2001)1 compared the results of SSCQE to results of the same test sequences obtained in a DSCQS study. Their interest was how retrospective ratings of the double-stimulus method correspond to the moment-by-moment ratings of the continuous single-stimulus method. In one part of their study, Aldridge et al. used a test sequence of 24 minutes coded with MPEG2 at a constant bit rate of 2.5 Mbit/s. Subjects rated the quality of two different test sequences using SSCQE method according to ITU-R BT.500 (ITU-R BT.500-11). Additionally, they were asked to rate the quality of the whole clip retrospectively. By comparing the correlation between the means of SSCQE and the retrospective ratings, they showed that these results do not correlate in both test sequence cases.

In a second study, they further investigated this finding. Hands and Avons (2001) used 30 seconds long test sequences in which they included 5 seconds or 10 seconds of highly impaired video. The goal was to investigate the influence of impairment depth and duration on subjective quality ratings. All three test sequences (unimpaired, 5 seconds impairment, 10 seconds impairment) were rated by test participants with SSCQE and DSCQS method. Analysis of the results showed that participants neglected the different durations of impairment in the retrospective ratings, although SSCQE graphs clearly revealed that participants realized the different impairment durations. and could even estimate the impairment length in seconds when they were asked after viewing a test sequence. This effect is called duration neglect. Furthermore, Hands and Avons showed that test participants cared only for the

---

1 The results are also presented by Aldridge et al. (1995)
depth of impairment, not its duration, to determine the subjective quality in the DSCQS method. Regression analysis of possible influencing factors showed that the “peak impairment” (Hands and Avons 2001) is the best feature to regress retrospective ratings. Aldridge et al. (1998) finally conclude in their presentation of the same study that retrospective overall quality results might be "disproportionately influenced by the magnitude of negative peaks of distortion, independently of their duration." This "negative peak effect" (Aldridge et al. 1998) is a possible explanation for bad correlation of continuous (moment-by-moment quality) and retrospective results (overall quality) of the same test sequences.

2.3.3. Critique on single-stimulus continuous methods

To counter criticism that continuous assessment of long scenes might be affected by fatigue of the participants or that participants' ratings might drift over time, Aldridge et al. (1998) conducted two experiments. The first one targeted the problem of arising fatigue during continuous long-sequence quality assessment. Two groups of ten participants each had to watch either a 24 minute long video sequence or just the last 10 minutes of it using SSCQE method. At the end of each session, participants were asked to give an additional retrospective quality rating. In both sequences, a 1 minute scene of large impairment was included at the same position after 19 minutes (or 5 minutes, respectively) from the start. The results of this experiment show that the continuous ratings for the impaired sequence were comparable for both groups and the long-sequence group even seemed to be "more sensitive to quality variations." An analysis of the retrospective ratings didn't show any significant differences for both cases, too. Aldridge et al. (1998) conclude "that fatigue has little or no effect on SSCQE ratings for a test session of around 25 minutes. Indeed, the use of long test sessions appears to improve rather than to impair subjects' ability [... to] identify coding errors." (Aldridge et al. 1998)

The second experiment (Aldridge et al. 1998) targeted the effect that ratings might drift over time in continuous assessment of long sequences. A 24 minute long test sequence was therefore evaluated by 24 participants in case a) as a whole and in case b) in short sequences of 3 minutes whereby the participants were asked to initiate the slider to the middle position before a new scene started. Analyzing the ratings of the long sequence and the rating of the corresponding short sequences, statistical tests showed that the ratings of case a) lie within the 95% confidence interval of case b) “for over 70% of the time, suggesting that they are generally comparable.” The experiment shows that drift is not a major problem in continuous quality assessment.

However, there is one main criticism concerning single-stimulus continuous assessment methods. Studies (Hands and Avons 2001) have shown that continuous assessment according to ITU recommendations (ITU-R BT.500-11) can be a too demanding task for the test participants and hence can influence the ratings negatively. This was shown by Hands and Avons (2001) when examining recency effect and duration neglect in subjective video quality assessment tests. They asked participants to track subjective quality continuously and finally give a rating of the overall quality retrospectively. Afterwards, Hands and Avons compared these retrospective ratings to ratings obtained without preceding SSCQE and found out that recency effect were removed for the continuous + retrospective case. They then conclude that continuous assessment influences the quality integration strategy in a way that the “demands of continuous assessment may block the step-by-step strategy and thereby eliminate the recency effect.” (Hands and Avons 2001)

2.3.4. Brotherton: ACR vs. SAMVIQ

Looking for the most efficient method to assess multimedia quality subjectively the Video Quality Experts Group (VQEG) compared single-stimulus ACR and multi-stimulus SAMVIQ (Brotherton et al. 2006). Two laboratories thereby ran both ACR and SAMVIQ based assessment tests of the same stimuli. All apparatus and laboratory characteristics were comparable. 8 seconds long video only test sequences of different content (music, sports, documentary, animation, movie trailer, talking head and advertisement) were selected “to span a wide range of spatial (SI) and temporal (TI) complexity.” (Brotherton et al. 2006) The two laboratories used different user groups for their test. One group
consisted of 15 laboratory staff members, the 25 test participants in the latter lab were recruited from public space.

According to Brotherton et al. (2006) the preferable method is the ITU recommended ACR (ITU-T P.910) for the following reasons:

- ACR shows a very good inter-laboratory reliability as results comparison of the two studies produces a correlation index of 0.98. Brotherton et al. outline that inter-laboratory comparability of the results is a very important finding for the purposes of VQEG because of multiple studies that are run in different laboratories.

- The previous results also show that ACR delivers comparable results for different user groups. This finding is a very important feature for multi-laboratory studies.

- Brotherton et al. (2006) found out that a comparison of SAMVIQ and ACR shows a correlation of the results of about 0.94. However they prefer ACR as it allows testing more test sequences due to the fact that in SAMVIQ test participants can spend a lot of time to switch between the stimuli and adapt their ratings. Brotherton et al. suggest that “ACR tests could present for assessment [at least] twice as many test sequences as the SAMVIQ tests.” (Brotherton et al. 2006)

- SAMVIQ is seen as to artificial compared to ACR. The fact that SAMVIQ allows the participants to replay test sequences and to adapt the ratings as often as they want is criticized by Brotherton et al. (2006). This “review capability” leads to a more artificial test method compared to ACR as “in real viewing situations observers do not normally review content.” (Brotherton et al. 2006)

There is only one case where SAMVIQ produced slightly more critical ratings. The ratings of a stimulus with clearly reduced temporal resolution let Brotherton et al. (Brotherton et al. 2006) suggest that the ability in SAMVIQ to review the stimuli and to verify the errors several times among the stimuli is one minor advantage of SAMVIQ. However, this still means increased artificiality as mentioned above.
3. Related work: User-centered approaches

3.1. Standardization activities of ETSI HF STF 354 and ITU-T SG12

The assessment of the Quality of Experience (QoE) has become more important in audiovisual quality assessment and MOBILE3DTV has started to evaluate user experience for mobile 3D television and video. Within standardization bodies the topic of Quality of Experience has been included into standardization activities. The goal of ETSI HF STF 354 (Specialist Task Force 354 within the Human Factors Group of ETSI) is to provide requirement guidelines for real-time multimedia services aiming at providing a good QoE. Goal is to provide objective and subjective measures of user experience for given communication situations, service prescriptions and levels of QoS (2006). The guidelines as the expression of measure are available in ETSI EG 202 534 (2007) and ETSI TR 102 535 (2007).

Within the International Telecommunication Union ITU standardization activities towards QoE are targeted in ITU-T SG12, the lead study group on quality of service and quality of experience within the ITU. Goal of ITU-T SG12 is to be able to a) measure quality parameters in next generation networks and b) measure their impact on QoE. The challenge of providing good QoE for new multimedia systems is study item of question Q13/12 “QoE, QoS and performance requirements and assessment methods for multimedia including IPTV”3. Main study items are the relation of end-user requirements to system parameters and to identify simple and efficient analysis techniques to measure and monitor QoE. All results will be included into the development of new recommendations.

3.2. Evaluation methods on behavioral level

3.2.1. Quality of Perception

Gulliver & Ghinea (Gulliver and Ghinea 2006) introduce their framework of subjective quality measurements including a two part approach to determine quality. The focus on the technical aspects of the system is described in the Quality of Service. On the other hand, Quality of Perception (QoP) describes the subjective level of quality to measure the "infotainment impact of the presentation." QoP asks about "information transfer' and user 'satisfaction'." The approach allows describing user satisfaction with two different aspects. Quality is determined on a media level independent of the content (i.e. judgment of the users relating to objective QoS), but also on a content level (i.e. the level of users’ enjoyment). Finally, these two levels are expressed by three constructs that can be measured independently contributing the users' judgments: QoP-IA to describe understanding of content, QoP-LoQ to measure subjective level of quality and QoP-LoE to measure the subjective level of enjoyment.

3.2.2. Acceptance threshold: Uni- and bidimensional measures

"In any test procedure it is important to decide acceptability criteria before the assessment is commenced."[ITU-R BT.500-11] Acceptability of sequences under test has become an important factor in subjective quality evaluation. Jumisko-Pyykkö et al. (Jumisko-Pyykkö et al. 2008) present a method called Acceptance Threshold. This method combines continuous acceptance rating and retrospective quality rating. This idea was already introduced by Sasse and Knoche (Sasse and Knoche 2006) who outline that continuous rating along a scale can be too demanding for test participants, but that the binary ratings of acceptable or unacceptable are a good way to control quality acceptance continuously. In Acceptance Threshold, test participants first watch video sequences of 60 seconds and mark parts that they find to be of unacceptable quality by clicking on a MIDI controller. After the

---

test sequence, participants rate their perceived overall quality of the items on an 11-point continuous scale, according to ACR (ITU-R BT.500), and the overall acceptability of the sequence on a binary yes-no scale. In contrast to continuous quality rating, which can influence additional retrospective ratings, Jumisko-Pyykkö et al. (2008) showed that a less-demanding task like their acceptance-marking does not have any impact on retrospective ratings.

Another method to assess the threshold of acceptance is based on Fechner’s Method of Limits (Fechner 1966). McCarthy et al. (2004) used the method of limits to investigate the quality acceptance threshold of small-sized video. In their method they increase and decrease different quality parameters of the test sequences and ask the participants to indicate if they find the current quality acceptable or unacceptable. The goal is to find the critical point “at which quality becomes unacceptable.” (McCarthy et al. 2004) During the test sessions, participants watch 30 seconds long test sequences in which the different quality parameters are varied continuously. Participants just mark if the current quality is acceptable or not and, finally, the ratio between acceptable and unacceptable quality per test sequence is taken as the value of quality.

3.2.3. Measuring the ‘being there’: The concept of Presence

Ijsselsteijn et al. (2000) define the concept of presence as “the sensation of ‘being there’ in a mediated environment”. Presence thereby is seen to be multidimensional and these different determinants of presence need to be identified. This leads to multiple approaches which all are meant to measure presence. These methods can be divided into groups of subjective measurements and objective corroborative measurements. While the objective approaches comprise all different psychophysiological and behavioral measurements, the subjective methods base mainly on presence questionnaires (Hendrix and Barfield 1996, Witmer and Singer 1998), continuous assessments (Ijsselsteijn et al. 1998), or qualitative methods (Freeman and Avons 2000). A good summary of all methods applicable to measure presence is given by van Baren and Ijsselsteijn (2004).

3.2.4. Evaluation in the context of use: Quasi-experimental approach

There is very limited number of studies targeting to conduct evaluation in the potential context of use instead or parallel to the controlled laboratory environment. This approach has become relevant for mobile and ubiquitous technologies. In general, there is juxtaposition between controlled and field measures in the level of control versus realism. The controlled laboratory measures represent an artificial setting, enable accurate control of variables and replicable experiments, but suffer from limited realism and unknown level of realism (Wynekoop & Russo 1997). Quasi-experimentation is a

As an example, McCarthy et al. (2004) varied quantization, frame rate and a combination of both as the three quality parameters.
presented as a new way to conceptualize experimental intervention conducted without full control over potential causal events, control and validity (Shadish et al. 2002, Cooc & Campbell 1979). Oulasvirta (in press) has outlined potentiality of quasi-experimentation to evaluations of mobile applications (e.g. in mobile phone calls and comparisons of mobile maps). There are no publications in user-centered quality evaluation which have applied quasi-experimentation in the current phase.

There are only two studies of quality evaluation of mobile television conducted outside of controlled environment. Jumisko-Pyypkó & Hannuksela (2008) have conducted a comparison between controlled and three field settings for mobile television by varying the contents and residual transmission error rates. The experiment took a place in different physical contexts with related task (Bus – travel to library, Cafe – relax, Railway station – wait for a friend). The data-collection was done using a combination of retrospective bidimensional acceptance threshold method and examination of goals of viewing (entertainment and information assimilation). Their results have shown evaluations were more favorable and less discriminate in the mobile contexts compared to the laboratory.

Similarly, Knoche & Sasse (2008) have conducted a comparison between controlled and field (underground) settings. They varied image resolution with different content as shot types. They applied a acceptance threshold method as such and did not give any additional tasks for the participants to the top of quality evaluation. Their results showed that image size had high value for users in field settings and concluded that current prediction models based on preferred viewing distances for TV and large displays do not predict viewing preferences on mobile devices.

In sum, both of published quality evaluation studies in the field revealed a significant difference between the results controlled and laboratory results. They also revealed that 1) the role of the field evaluations should not be underestimated in the development process and 2) methodological guideline for the contextual experiments are highly needed.

In the field of human computer interaction, some experiments in mobile usage context have been reported. Kaikkonen et al. (2005) compared usability in a laboratory test with usability experienced during a short-term travel task (crossing a street, taking the subway and escalators, as well as finding one's way). The field results remained equivalent compared to the laboratory results, but there were also differences. E.g. long loading times in web browsing were mentioned less in the real usage contexts than in the controlled environment (Kaikkonen 2005).

Oulasvirta et al. (2005) has examined attentional resources in the controlled and field settings. In their work participants conducted a mobile web browsing task in a laboratory and in a way finding task (e.g. in a busy street, travelling in a bus/ metro, chatting while having a coffee, standing in a busy railway station and while waiting for a metro). One of the main finding was that attentional span in the silent lab situation can last up to 16 seconds while in the field it can be only 4 seconds.

There are also several studies on mobile text legibility and entry on move. These studies are mostly conducted in a controlled environment with tasks of walking in a certain speed on a marked walking route, standing or sitting (Mustonen et al. 2004, Mizobuchi et al. 2005, Vadas et al. 2006, Barnard et al. 2007, Brewster 2002). In addition, impact of light representing the factors of physical context has been studied (Barnard et al. 2007).

3.3. **Ergonomic approach: Measures of cybersickness as part of quality evaluation**

Many people have experienced symptoms of motion sickness at some point in their lives. It is mainly evoked by transportation, e.g. ships, airplanes or automobiles, and is a result of a certain excitation of the vestibular system (LaViola 2000). The symptoms experienced by the persons concerned can be grouped into three categories: Disorientation, oculomotor disturbances, and nausea.
3.3.1. Simulator Sickness

In 1957, Havron and Butler described motion-sickness like symptoms with users of military flight simulators. Although the symptoms are very much alike, simulator sickness is not the same as motion sickness, as it can be caused by visual stimuli alone (Kolasinski 1995). The simulator sickness symptoms are usually subjectively measured by the Simulator Sickness Questionnaire (SSQ), which was developed by Kennedy et al. (Kennedy et al. 1993) after a review of the data of 1200 simulator. The SSQ consists of 26 Questions, assessing subjective symptoms of three dimensions of simulator sickness: nausea, oculomotor disturbances and disorientation. Each of the questions consists of a 4-point scale (none, slight, moderate, severe) to assess the severity of each symptom, and a weighted average—the total score (TS)—is used to reflect the general discomfort (Kennedy et al. 1993). The SSQ is filled in by the participants and in most cases only once after the exposition to the sickness evoking situation. A few other studies use it as a relative measurement tool as the SSQ is filled in twice, before and after the experiment (Häkkinen 2002, Strohmeier 2007). Aside from the need to check how the SSQ is applied when comparing different studies it has to be discussed whether the relative score of the SSQ provides more useful results than the absolute. To our knowledge this discussion has not yet taken place openly, but we believe that the relative measurement has its advantages, since it eliminates the participants' immediate pre-test conditions (e.g. agitation).

3.3.2. Recent studies investigating autostereoscopic displays

The three-dimensional presentation of stimuli by an autostereoscopic display is similar to the one of a VE, where usually a Head Mounted Device (HMD) is used. Furthermore, motion sickness-like symptoms are reported by participants using autostereoscopic displays. This could lead to the assumption, that the symptoms experienced with these displays are more or less the same than within a virtual environment. Surprisingly, only few studies could be found in which SSQ was used to assess stereoscopic displays (Strohmeier 2007, Häkkinen et al. 2006).

3.4. Descriptive quality

3.4.1. Qualitative interviews as a part of quality evaluation: Experienced quality factors

As a background, subjective quality is not seen as a determinant of several predefined variables any longer. In contrast to Reiter and Köhler who write that "the term overall quality is a fuzzy one which is often individually interpreted by test subjects and its meaning depends on subjects personal background" (Reiter and Köhler 2005), Jumisko-Pyykkö et al. (2007) state that human behavior in determining subjective quality must be seen as a challenge and, hence, test methods must be open to understand underlying quality attributes. Following, we present a review of several test methods that support this point of view on subjective quality. All these methods aim to identify "experienced quality factors" (Jumisko-Pyykkö et al. 2007) or hidden structures which are used by the participants to form their quality.

This idea to use interviews to examine individual quality factors of the users was used in two studies by Jumisko-Pyykkö et al. (Jumisko-Pyykkö et al. 2007, Jumisko-Pyykkö et al. 2007-2). In both studies a combination of quantitative ratings according to the ACR method (ITU-T P.910) and subsequent interviews with the test participants was used to reveal underlying factors of subjective quality. Jumisko-Pyykkö et al. call these quality evaluation criteria "experienced quality factors" (Jumisko-Pyykkö et al. 2007). While in (Jumisko-Pyykkö et al. 2007) the interview was just conducted retrospectively after the quantitative part, the study presented in (Jumisko-Pyykkö et al. 2007-2) included a pairwise comparison test in which the interviews were used to investigate differences between each pair of stimuli.
3.4.2. Sensory Profiling and Mixed Methods approaches

Interpretation Based Quality

The approach of Interpretation Based Quality (IBQ) was introduced by Nyman et al. (2006) and is fully described by Radun et al. (2008). IBQ was used by Häkkinen et al. (2008) to assess the viewing experience of stereoscopic content.

IBQ follows a two-step approach. It consists of a first part comprising a Classification Task using Free-sorting and a Description Task using Interviews. The second part then is the psychometric assessment where assessors evaluate one perceived attribute of the stimuli.

Free-sorting tasks as an alternative to descriptive analysis with naïve assessors was introduced by Picard et al. (Picard et al. 2003) and Faye et al. (Faye et al. 2004). Especially the study of Faye et al. (Faye et al. 2004) aimed to compare the method of perceptive free-sorting with naïve assessors to experts’ descriptive mappings. In this study, naïve assessors were asked to first group 26 “rectangular translucent plastic pieces” (Faye et al. 2004) according to the individually perceived similarities of the stimuli. Assessors were free to form in at least 2 groups and each group consisted thereby of 2 to 25 stimuli. This ensured that a group structure was build. After sorting the stimuli into groups the assessors described the created groups. In the analysis the co-occurrence matrix representing the similarity of the stimuli in a stimuli × stimuli-matrix and the coded qualitative expressions were analyzed using Multidimensional Scaling. Comparing these results to those obtained in a descriptive mapping with expert listeners Faye et al. show that both results are comparable in terms of describing the same sensations and the related wording of the attributes (Faye et al. 2004). Faye et al. (Faye et al. 2004) discuss that the results of the free-sorting task provide more nuances of one term whereas expert assessments with fixed vocabulary are more precise in the definition of one certain term. According to (Faye et al. 2004) the main advantages of free-sorting tasks are lower costs due to naïve assessors, missing training sessions and a fast assessment of a large number of stimuli compared to descriptive profiling with experts. But although naïve assessors’ attributes and wordings can help to understand better consumers’ perception, the large amount of words produced by these assessors require more complex data analysis and finally data interpretation. Faye et al. underline that detailed
knowledge about the used stimuli and their perceptual differences was helpful in this task of interpreting the results (Faye et al. 2004).

IBQ method (Nyman et al. 2006, Radun et al. 2008, Häkkinen et al. 2008, Radun et al. 2006, Radun et al. 2007) adapts the approach of Faye et al. (Faye et al. 2004) and extends the free-sorting and decrption task with a psychoperceptual assessment of the stimuli. Radun et al. (Radun et al. 2008) divide IBQ into two main parts as it is depicted in Figure 6. The first part covers the "classification (free-sorting) and description (interview) tasks" and the second part is a "psychometric evaluation." Originally, IBQ was developed to evaluate subjective quality of still images (Nyman et al. 2006, Radun et al. 2008). As in Faye et al.'s work (Faye et al. 2004), participants are first asked to sort the test images into groups according to the perceived differences in a way that "each group differs from other groups on some attribute." (Radun et al. 2008) After the participants classified the pictures into groups, they describe each of the groups (description task). They are interviewed by the researcher about the criteria which the participants used to classify the groups and which of the groups the participants prefer.

The second part then covers the psychometric evaluation. All stimuli, not the groups, are analyzed on an 11-point scale. The participants have to estimate one attribute in all images and for reference an image of best and worst value of the attribute are presented before.

Figure 6 Structure of the IBQ method (Radun et al. 2008)

Häkkinen et al. (2008) adapted the previously described method in a way that the qualitative data is collected based on a pairwise comparison. Häkkinen et al. presented the stereoscopic and the non-stereoscopic version of several video sequences. After each pair presentation, the participants were interviewed which version they prefer and why. In Häkkinen et al. (2008), the quantitative part of the study is still missing so that analysis is only based on the qualitative attributes following grounded theory. The results outline interesting content factors which can contribute to a good perceptual quality of stereoscopic content. The results are included in the literature review about user needs and expectations for stereoscopic video in (Strohmeier et al. 2008)

**Free Choice Profiling**

In 1984, Williams and Langron (Williams and Langron 1984) presented a new approach in sensory profiling that allowed assessors to use their own vocabulary to describe the characteristics of products or attributes with explaining it. The advantage is a massive reduction of time compared to other methods like descriptive analysis as the task of vocabulary development and group agreement is missing. Williams and Langron outline that their approach allows the assessors to be individuals having "their own differing sensitivities and idiosyncrasies" (Williams and Langron 1984) to describe products.
Free Choice Profiling has been used in different areas of sensory evaluations (Williams and Langron 1984, Jack and Piggott 1992, Wemelsfelder et al. 2001) since its first presentation in 1984. However the application field has changed among the studies, the test methodology has remained the same as it was presented in the work of Williams and Langron (Williams and Langron 1984). Assessors first develop their own non-hedonic attributes to describe sensory characteristics of the samples under test. The whole task of attribute generation is done with the help of at least a subgroup of the test samples to represent sensory differences and the range in differences. To support this attribute finding task researchers can present general terms describing the product. In a second stage, current studies have introduced a training session which, after vocabulary generation, allows the assessors to adapt their attributes, to remove or to add terms (Wemelsfelder et al. 2001). Each attribute is then attached to a scale which is mostly unlabeled (Williams and Langron 1984, Wemelsfelder et al. 2001).

The second part of the procedure then is related to the sensory evaluation itself. Participants evaluate the sensory characteristics of the test items (samples) according to their own vocabulary by marking the relating sensations of each attribute on the scales. This data is then analyzed using Generalized Procrustes Analysis (GPA).

As participants are free to use their own vocabulary and definitions are not given in the vocabulary generation task, general statistical methods of data analysis are not applicable in this case. Gower (1975) introduced a method called Generalized Procrustes Analysis which allows to first scale FCP data according to an inter-assessor consensus which then can be analyzed by Principal Component Analysis. As GPA has got high calculation costs due to the complex scaling algorithms Kunert and Quannari (1999) introduced an alternative approach which is applicable for FCP data, too. As a result of both methods of analysis researchers obtain an Item Factor Map depicting the sensory space that is spanned among the items under test and additionally PCA like correlation plots of all assessors' attributes. Both of these results can be used for interpretation of the sensory space to explain product characteristics.

In audiovisual research the FCP approach was adopted by Strohmeier (Strohmeier 2007). In his study on different perception of audiovisual autostereoscopic and non-stereoscopic stimuli, Strohmeier used FCP to collect sensory data about the perception of audiovisual stimuli. Additionally Strohmeier (Strohmeier 2007) collected quantitative preference data of the same stimuli that were combined with the sensory characteristics following the theory of Mixed Methods Research (Creswell and Plano Clark 2007).

Strohmeier's approach (Strohmeier 2007) consists of a three step evaluation of the test sequences as can be seen in Figure 7. In a first step, the overall quality of the test sequences is assessed using quantitative methods according to ITU recommendations. Strohmeier (Strohmeier 2007) used ACR (ITU-T P.910) so that quantitative data collection and FCP data collection both are done using retrospective, single-stimulus methods. The second step comprises the introduction to FCP and the development of the participants' vocabulary. To familiarize participants with the method and to motivate them to use their own vocabulary, Strohmeier uses training with cookies to explain the method. Participants need to learn that quality attributes are everything that they perceive and that they need to name the attributes preferably with an adjective. After the training, the individual attributes are developed. Therefore, a subgroup of test sequences is presented one by one and test participants write down their attributes.

Before the third step is run, the attributes are transformed into a questionnaire. In contrast to IBQ (Radun et al. 2008) where qualitative interview data is analyzed according to grounded theory, the qualitative data of the FCP method is transformed and analyzed as quantitative data (Wemelsfelder et al. 2001). Therefore each attribute in the questionnaire is tagged with a 10cm long line and the ends of the lines are marked with 'min' and 'max' for 'minimum perception of the attribute' and 'maximum perception of the attribute', respectively.
Then, in the third step, the test sequences are assessed by the participants using their individual questionnaire including all the attributes that were developed in the second part. Therefore, the test sequences are again watched one by one and participants tick the line of each attribute according to their perception of the attribute on the corresponding place between min and max.

![Three step evaluation of FCP according to Strohmeier (2007)](image)

The combined data of quantitative and qualitative assessment of the same test sequences is finally analyzed with the help of Generalized Procrustes Analysis (Gower 1975) and External Preference Mapping (Mattila 2001). This allows mapping quantitative ratings and experienced quality factors in one common room for better interpretation.
4. Development of user-centered quality evaluation method

4.1. Starting points for developing user-centered quality evaluation method for product development purposes

This section presents the starting points for developing a new evaluation method for quality evaluation of mobile 3DTV. The aim is to broaden reader understands about the underlying themes and the related work done in past.

Following principles has been taken as a starting point for development of our user-centered quality evaluation method (each of them is presented in detail):

1. Quality perception is an active process combining different levels of human information processing.
2. Component user experience examines quality of critical system component by reflecting the factors of whole user experience.
3. Quality evaluation experiments are part of human-centered design process.
4. Quality evaluation for system optimization in product development has special characteristics.
5. Positive quality – good enough for use – is in our focus.
6. Low produced quality, noticeable impairments and heterogeneous stimuli give requires guaranteeing the level of minimum acceptable quality.
7. Overall quality evaluation approach is suitable for user-oriented quality evaluation.

1. Quality perception is an active process combining different levels of human information processing.

Active perception integrates always low-level sensorial processing and high-level cognitive processing including personal knowledge, attitudes, expectations and emotions (Dedtweiler-Bedell et al. 2006, Fiske & Taylor 1991, Keltner et al. 1993, Neisser 1976, Oatley & Jenkins 2003, Silva et al.2006). Each sensory modality has its special characteristics that depend on the physical dimensions of stimuli. The purpose of early sensory processing is to extract relevant visual features from the incoming sensory information. In vision, sensory processing uses brightness, form, color, stereoscopic and motion information in creating the early perceptual experience whereas pitch, loudness, timbre and location are the attributes of auditory processing (Ts’o, & Roe 1995, Grill-Spector & Malach 2004, Evans 1992). However, the final quality judgment is always a combination of low-level sensorial and high-level cognitive processing. In cognitive processing stimuli are interpreted and their personal meaning and relevance to intentions and goals are determined. For example individual emotions, knowledge, expectations and schemas representing reality affect the weight that each sensory attribute is given and these factors enable human contextual behavior and active quality interpretation (Jumisko-Pyykkö et al. 2007, Jumisko-Pyykkö 2008, Nyman et al. 2006, Radun et al. 2006).

Multimodal perception has its special characteristics. The unified multimodal experience of audiovisual material is created when the information from audio and visual channels are combined. Different modalities can complement and modify the perceptual experience created by other perceptual channels and therefore multimodal experience is more than the simple sum of two different perceptual channels (Hands 2004). The McGurk effect is a classical example of audiovisual integration in speech perception in which the mismatched visual and acoustical materials are integrated into unified experience which differs from both presented material (McGurk & MacDonald 1976). The detailed integration process of audiovisual material itself is complex and still relatively
unknown. Traditionally audiovisual perception is examined as isolated processes of different channels (also studied in separate research societies) and multimodality is a combinatorial mechanism on the top. However, the multimodality has been also approached from the behavioral and early combination point of view (Coen 2001). Even though the processing is not known in depth, synthesis of auditory and visual material is requirement for unified perception. The proper temporal synchronization between the sources is needed in synthesis. Inadequate synchronization reduces the clarity of message and distracts the viewer from intended content (Reeves & Nass 1996).

2. Component user experience examines quality of critical system component by reflecting the factors of whole user experience.

For understanding experienced quality of critical system component, we present the basic principles of holistic user-experience (UX) at first. User experience research examines ‘the big picture’ or wholeness of quality. Compared to cognitive and task-oriented concept of usability (ISO 13407 1999), user experience aims at shifting the focus to the direction of hedonistic and emotional aspects and from negative to positive experiences. Among the many relevant definitions Hassenzahl and Tractinsky (2006) define it as:

_UX is about technology that fulfils more than just instrumental needs in a way that acknowledges its use as a subjective, situated, complex and dynamic encounter. UX is a consequence of a user's internal state --, characteristics of designed system -- and the context -- within the interaction occurs._

This definition not only underlines similar characteristics as human perceptual processes, but also summarizes the building blocks of user experience: user, system and context. Later, Roto (2006) has followed the work of Hassenzahl and Tractinsky (2006) and presented an extensive model for UX of mobile browsing. The UX model presents UX factors and systematic definitions for each of them offering a vocabulary for the field. We continue applying this categorization of factors in our work, assuming that mobile browsing has the same characteristics as mobile 3D television at an abstract level (Jumisko-Pyykkö et al. 2008). The main factors of UX are user, system and context:

User is defined as a person controlling or manipulating the system and she/he can be described as having the characteristics of needs, motivations, experiences, expectations, mental state and resources (Roto 2006).

System is defined as the system required for the product under examination to work or to be useful (Roto 2006). From the user’s viewpoint the mobile system can contain a device, browser or player, connection and site or content (adapted from (Roto 2006)). We use the term content to refer any type of moving image or video. The concept of service including, e.g. commercial and a service model, is often overlapping with the term system from the user’s point of view. In this paper, we treat service features as part of the definition of system (e.g. Koivumäki et al. 2008, Mallat et al. 2008).

Context represents the circumstances under which the activity (mobile browsing) takes place (Roto 2006). Physical, temporal, social, task factors and changes within these factors are the essential components of the mobile context (Roto 2006, Belk 1975, Hiltunen et al. 2002, Tamminen et al. 2003, see Jumisko-Pyykkö et al. 2008 for more detailed descriptions).
Critical component user experience

Component user experience approach examines the quality of the certain system component taking into account the factors of holistic user experience (Figure X). Critical system component refers to the part of the whole system which performance can impact negatively or prohibit the utility of the whole system from the user's point of view. The original idea of experienced quality of critical system components has been recently introduced (Jumisko-Pyykö et al. 2008). It aims at ensuring that experienced qualities of components, developed in isolation from the end-product, are not barriers for adoption of end-products and therefore their acceptability should be studied in their optimization process (Jumisko-Pyykö et al. 2008).

There are also factors necessary to take into account from holistic UX. We use recent work of mobile TV and video quality evaluation studies as examples. 1) User factors: selection of potential viewers and impact of their background factors, like interests in the content and technology attitude has been as a part of quality evaluation studies (Jumisko-Pyykö & Häkkinen 2008), 2) System – content: Potential television content has used as a stimuli material (e.g. Jumisko-Pyykö & Häkkinen 2005, Knoche 2005), 3) Context: the first studies to introduce contextual quality evaluation procedure parallel to controlled laboratory measures has been conducted (Jumisko-Pyykkö & Hannuksela 2008, Knoche & Sasse 2008), 4) User-system interaction: There has been an attempt to understand the interpretation of constructed quality (Jumisko-Pyykkö et al. 2007), 5) User-system interaction: Acceptance of quality has been examined to figure about weather the presented quality is acceptable for viewing mobile television (Jumisko-Pyykkö et al. 2008).

3. Quality evaluation experiments are part of human-centered design process.

Broadly speaking, ‘user-centered design is a design process that views knowledge about users and their involvement in the design process as a central concern’ (Preece 1994). It is characterized by active user involvement to understand user’s requirements, iterative design and evaluation and moreover, a multidisciplinary approach (ISO 13407 1999). An overall idea of cyclic human-centered design process is described in Figure 2. The gained benefits of user-centered design are listed in the terms of better end-user satisfaction and lower costs of system development (Maguire & Bevan 2002).
MOBILE3DTV project is characterized as a system/product development project and therefore high-level emphasis is given for the user requirement elicitation at the beginning (mobimedia). To become successful, the system needs to satisfy its end-users. The results of our past study form initial requirements for mobile 3DTV and video (Jumisko-Pyykö et al. 2008). User requirements include any externally visible function, non-functional property or constraint that is required in order to satisfy user needs (Kujala 2002). From the viewpoint of UX this approach is limited, and therefore at the beginning of the product development, we have examined requirements more broadly and not only as goal or task related issues. Our requirement elicitation was based on three user studies, surveys, focus groups and a probe study, to form an initial user’s idea of mobile 3D television and video. The results are expressed as user, system and service, and contextual requirements.

The main requirements are summarised (Jumisko-Pyykkö et al. 2008):

**User - Mobile 3D needs to provide the following:**
- Fulfill entertainment and information needs. Users also want to relax, to spend time, and to learn by using mobile 3D services.
- Increased realism and naturalness and evoke an emotional relation and a greater feeling than existing systems of being inside.

**System & Service - Mobile 3D system needs to offer the following:**
- TV content (e.g. news, series, sport, documentaries) as well as other video contents (e.g. games, tailored 3D content, interactive guidance, navigation, product presentation).
- Both on-demand and push services, and both pay-per-view and monthly payment options.
- A device with a display size of 3 inches or larger and probably 4-5 inches
- Both mono- (audio or visual only) and multimodal (audiovisual) presentation modes and an easy shift between multimodal and visual 2D-3D presentation modes.
- Interactive possibilities including saving, receiving, sending, and recording.

**Context - Mobile 3D viewing has the following attributes:**
- Takes place in public and private locations and in outdoor and indoor environments, potentially on public transport, or in parks, cars, cafes, waiting rooms or at home.
- Is primarily for private and focused viewing, but there is also a need for shared viewing.
- It is well-suited to waiting situations, during transport trips in coffee or lunch breaks, and for short time viewing from a couple of minutes to 15 minutes or half an hour.
When improving and evaluating these critical non-functional system components we apply gathered user requirements to simulate the characteristics of end-product. The requirements offer guidance concerning users, content and context selection for these quality evaluation experiments. Our aim is also to use early stage prototyping parallel to the critical component improvement and evaluation to iteratively validate understanding of user requirements in greater depth (ISO 13407 1999, Maguire & Bevan 2002, Gotchev et al. 2008). User requirements in the earliest development phase typically contains inaccuracies, e.g. due to the incomplete data and the fact that imaging something that does not exist yet is a difficult task (Maguire & Bevan 2002).


Stages for (subjective) quality evaluation and related factors according to product readiness have been described by Jumisko-Pyykkö & Väänänen-Vainio-Mattila (2005). Quality evaluation stages can be divided into psychophysical assessment, quality optimization assessments and usability testing (FIGURE X, Jumisko-Pyykkö & Väänänen-Vainio-Mattila 2005, Reiter & Köhler 2005). In the product development, the emphasis is given for two latest. The usability tests require the highest degree of product readiness to be extensive. Some of the typical features can already be applied earlier stages referring to the optimization studies. In usability testing stage, potential users, contents and usage contexts (especially in the field tests) can be known in the applicable degree as well as the measurement tasks. Quality optimization assessment is also a part of typical testing in product development. However, from the viewpoint of final product its product readiness is not as high as in the usability tests. To reach the acceptable quality level as early as possible in product development, the same factors as in usability testing should be in focus in quality optimization stage.

5. Quality evaluation focuses on low produced qualities with possible noticeable impairments in heterogeneous stimuli, but aims at guaranteeing the acceptance.

We use a term low quality to refer to presentation with perceived noticeable impairments, and to make a distinction to perceptually impairment-free high qualities (e.g. top-end multichannel audio or...
This distinction is made because the evaluation methods of low and high qualities can be different requiring for example different type of evaluation tasks, experimental procedure or evaluators (naive vs. professional) (Jumisko-Pyykkö et al. 2008, Bech & Zacharov, ITU-R BT.500-11 2002, ITU-T P.911 1998)

The noticeable impairments, or artifacts, can be resulted from single or combined factors throughout the whole value chain (from content production and packaging, its delivery and transmission and reception including device and its display due to the required high-level of optimization of the system (Gotchev et al. 2008). This can result very heterogeneous stimuli material for the experiments. For low quality and for these circumstances, the measurements of acceptance threshold indicating minimum useful quality is needed (Figure X, Jumisko-Pyykkö et al. 2008). The goal is to ensure that the low produced quality is set in a way that constitutes no obstacle to the wide audience acceptance of a product or service (Jumisko-Pyykkö et al. 2008).

![Figure 5 Levels of produced and perceived quality. In low produced qualities, perceptually minimum accepted quality level is a threshold for the useful quality. (Jumisko-Pyykkö et al. 2008)](image)

6. Overall quality evaluation approach is suitable for user-oriented quality evaluation.

In overall or global quality evaluations, participants assess stimuli as whole. This approach is suitable for naive participants and under the usage of heterogeneous stimuli with different type of impairments and multimodal presentation. This approach also assumes that human information processing is integrative in nature including low-level sensorial and high-level cognitive processing (Bech & Zacharov 2006). In contrast to overall evaluations, perceptual evaluations of certain attribute (e.g. jerkiness) can be conducted with trained assessors.

4.2. Evaluation of experienced (audio)visual quality / User-centered quality evaluation

The goal of this section is to present an overview to user-oriented quality evaluation method. In the current phase, it is not meant to be a detailed methodological guideline giving step-by-step instructions for conducting the experiments. It is rather a holistic framework which can be applied and adapted into the different user-centered quality evaluation studies. The detailed methodological guidelines will be provided parallel with the reports of upcoming quality evaluation studies.

The main principles of the method are categorized following the factors of user experience: User – System/Service – Context – Task [Figure]. Each of these factors presents its aim and role in the
evaluation procedure, how they should be taken into account and why and the difference to the previous work.

![Diagram](image.png)

Figure 8 The principles of evaluation of experienced quality / user-centered quality evaluation

1. User - User/Participant selection
   - Sample selection: Aim at selecting the sample which is representative to the potential users of the system. When designing the new system or service sampling is done based on the user group definitions of user requirements. Potentiality can also contain many aspects, like relation to the content, service, or technology (Jumisko-Pyykkö & Häkkinen 2008).

   - External validity: Sample selection is related to the external validity of the research: Does the results generalize beyond the sample tested to some broader population of interests in MOBILE 3DTV and video?

   - Threats of external validity: As a difference to psychophysical approach, it is worth of noticing that the conventional categorization of participants to naive or professional evaluations in quality evaluation research can threat the external validity as well as assuming that the students are representative to all user groups (e.g. Jumisko-Pyykkö & Häkkinen 2008).

   - Demographic and psychographic data-collection: Collect broadly demographical and psychographic factors about users in the quality evaluation experiments for understanding quality and re-defining the user requirements. The collected data may contain: ADD. Test the participants’ sensorial sensitivity of the modalities under the investigation (e.g. tests: Ishihara test for color-vision, Randot stereovision test, visual and auditory acuity)

   - Reporting sample selection: Reporting needs to describe sampling in a level that the study can be replicated. The description needs to contain basic demographic variables (age, gender, education, age, gender, technology attitude (DIS scale), current habits related to content, service, and technology).

2. System/Service – Selection of produced quality factors (independent variables)

2.1 Content
Content selection: Aim at selecting the test contents which are representative to the potential contents of the system and representative to measure the phenomena under investigation. In the level of genre, the user requirements describe the most potential contents. Audiovisual characteristics need to also replicate the characteristics of desired genre and also be representative to measure technical parameters. Additionally, the length of the content should aim at representing the length of the potential viewing of one meaningful episode.

External validity: Content selection is related to the external validity of the research: Does the results generalize beyond the content tested to whole mobile 3DTV television or video content of interests?

Threats of external validity: As a difference to conventional psychoperceptual methods, test materials contain a thread for external validity in two ways for user-oriented quality evaluation. Firstly, for example, standardized clips in video quality (VOEG 2000), they represent the range of motion, content and shooting distance, but they are not representative for the mobile TV quality research due to missing audio, non-covering different contents and limited camera shots within one test clip (Knoche & Sasse 2005). Secondly, the short test contents (10s), are hardly representative to any meaningful segment of television content and are not suitable for measuring all video quality related phenomena (e.g. transmission parameters).

Reporting content selection: Reporting needs to describe the content in a level that the study can be replicated. The description should contain: the chosen genre, describe the story of the chosen content containing meaningful segment of content, its audiovisual characteristics (e.g. motion, details, speech/ music, text, cuts) and length.

2.2 System parameters

System parameter selection: Aim is to select system parameters or their combinations framing the meaningful and representative unit of the whole system. Meaningful unit for viewing may contain two modalities and combinations from whole value chain (production and packaging, its delivery and transmission and reception including device and its display).

External validity: Selection of system parameters is related to the external validity of the research: Does the results generalize beyond the parameters tested to whole mobile 3DTV system?

Threats of external validity: Threats of external validity is to choose parameters form limited part of the system, while their impact form the viewpoint of whole system can be very small. For example, drawing conclusions for the whole experienced quality just by examining coding would be potential thread. Similarly, concluding results for multimodal service just focusing in a conventional way on one modality at the time describes the potential thread and can lead to inefficient optimization of resources of the whole system (Winkler 200X, Knoche 2005, Jumisko-Pyykkö 2005).

Reporting parameter selection: Reporting needs to describe parameters in a level that the study can be replicated. The description should contain definitions of parameters and the used values.

3. Context – Contextual evaluation

Contextual evaluation: Aim of the contextual quality evaluation is to assure that the produced quality meets the requirements in the actual context of use. It can also aim, as in MOBILE 3DTV case, to confirm the results of quality evaluations from highly controlled environment. This
approach is important as the contexts of mobile devices are heterogeneous and dynamic in comparison to static settings of television located at home.

- Context selection: Aim at selecting the contexts which are representative to the potential usage contexts of the system. User requirements define the most potential contexts. The context selection needs to replicate characteristics of physical, temporal, social, task and changes of context during the usage according to the requirements (Roto, 2006, Belk 1975, Hiltunen et al. 2002, Tamminen et al. 2003)

- External validity: Context, i.e. selection of setting is also related to the external validity of the research: Does the results generalize beyond the setting used in the research to real-life settings of interest?

- Threats of external validity: Conventionally quality evaluation experiments have taken a place in the highly controlled and sensorial optimal laboratory settings. It has been shown that the drawn requirements from these 'perfect' settings differ significantly being more demanding compared to the requirements from the noisy mobile settings (Jumisko-Pyykkö & Hannuksela 2008). The field experiments have been used to prove the results of controlled environment.

- Contextual quality evaluation (i.e. field experiments) is not equal to 'lab' experiment as such to the field settings. It requires shift from experimental research to quasi-experimental research method, understanding and reporting of related threats of causal interference (Shadish, Cook & Campell, 2002).

- Reporting context selection: Reporting needs to describe context of research in a level that the study can be replicated. The description should contain definitions and descriptions of physical, temporal, task and changes of context during the use.

4. Task – Selection of evaluation task

4.1 Preferences of quality: Acceptance, satisfaction (dependant variables)

- Evaluation in relation to action: Aim is to define the evaluation task in relation to the actual viewing task with minimal distraction on it. We recommend using retrospective ratings of based on acceptance threshold for relating the quality evaluation to the actual application and ACR methods (Boherton et al. 2006, Jumisko-Pyykkö et al 2008.). The involvement in the viewed content is the part of the evaluation and evaluator's task is not especially concentrate on errors. If the interaction with application is the part of the user action, then it should be included in the evaluation procedure (Reiter & Jumisko-Pyykkö 2007). E.g. in navigation the main task is way finding, so the way finding should be included in the evaluation.

- Reporting the data-collection of evaluation task: Evaluation task and procedure needs to be reported including the presented question for the participants.

4.2 Descriptive interpretations and impressions of quality

- Understanding of quality: Aim at measuring experienced quality qualitatively is to understand the evaluators' interpretations and impressions of quality. When studying 1) novel heterogeneous stimuli without knowing its perceptual effects in detail and 2) using overall evaluation approach with naive participants, it is important to understand what kind of aspects of stimuli has been paid attention. We recommend to use two either lightweight version of 1) retrospective interview of quality evaluation criteria (Jumisko-Pyykkö et al. 2007). This approach can result for example the description of the most distractive factors, but also unexpected ideas for the further development and confirm the reliability of the study. Or, 2) Free-choice profiling or other mixed methods to understand the constructs of perceived
quality more in detail and relate these constructs to quality preferences (Radun et al. 2007, Strohmeier 2007).

- Reporting the data-collection: The task and procedure for gathering interpretation of quality needs to report including the main questions presented during the research.

4.3 Ergonomic approach to support quality evaluations

- Make sure low negative ergonomic impact: We recommend using ergonomic measured to check the reliability of settings of experiments. For task reliability, especially in quasi-experimental settings, NASA Task Load Index (Hart & Staveland 1988) can be used and Simulator Sickness Questionnaire (SSQ) (Kennedy et al. 1993) for checking low level of symptoms of cyber sickness in the experiments.

- Reporting the data-collection: The tools and procedure for gathering collecting ergonomic impact (task, cyber sickness) needs to reported.
5. Plan for the content of the experiments

An overall plan for the subjective quality evaluation experiments is described in this section. This plan 1) aims at describing the content of each of the experiments and schedule for conducting the experiments in high-level, 2) describing that critical parts of the system will be covered in the quality evaluation experiments, and 3) describing psychoperceptual and user-centered quality evaluation studies. The detailed plans for each of the experiment including the final variable selection and method will be documented at the time of planning a particular experiment.

5.1. Psychoperceptual experiments

Experiment 1: Impact of different stereo video capture, representation and coding factors on experienced visual quality

Research question:
How the different stereo video capture, representation and coding factors impact on visually experienced quality?

Independent variables:
Formats: simulcast, intr-view, video-depth, mrsc;
Bitrates: 4-5
Framerates: 3
Artifact related variables (e.g. blurring): 3
Contents: 5 (~10s)

Number of participants:
80

Schedule:
0 1.11.2008-1.11.2009

Reporting:
November 2009

Unit in responsibility:
TUI

Other involved units:
TUT, HHI

Experiment 2: Impact of channel and display factors on experienced visual quality

Research question:
How the transmission channel and display factors impact on experienced visual quality?

Independent variables:
Error rates (e.g. MFER%): 3-4
Error concealments: 2-3
Error resilience methods: 2-3
Code rate: 2-3
Display parameters: 4-5
Temporal mismatch between views: 3-5
Contents: 5 (~60s)

Number of participants:
80

Schedule:
1.6.2009-1.3.2010
5.2. **User-centered quality evaluation experiments**

**Experiment 1:** Comparison of experienced audiovisual quality of 2D vs. 3D in the controlled and field settings

**Research questions:**
How the experienced qualities differ between audiovisual 2D and 3D presentations on small portable devices? How these experienced qualities differ between controlled and field settings?

**Plan:**
Appropriate selection of audiovisual coding parameters (bitrate, framrate, sampling rate, picture size) will be chosen to the experiments. The chosen parameters will follow parameters of psychoperceptual experiment 1. The contents will be customized from the contents available.

**Independent variables:**
- Video presentation: 2 (2D, 3D)
- Audio presentation: 2 (mono, stereo)
- Video bitrate: 3
- Audio bitrate: 3
- Video framerate: 2
- Audio sampling frequency: 2
- Contexts: 3 (Lab, 2-3 contexts of use)
- Contents: 5 (>10s)

**Number of participants:**
60

**Schedule:**
0 1.11.2008-1.6.2009

**Reporting:**
June 2009

**Unit in responsibility:**
TUT

**Other involved units:**
TUI

**Experiment 2:** Impact of coding and transmission factors on experienced quality of different user groups

**Research questions:**
How experienced audiovisual quality is impacted by coding and transmission factors and how experiences differ between appropriate user groups?

**Plan:**
Appropriate selection of audiovisual coding and transmission parameters (codec, bitrate, error rate) will be chosen to the experiments. The contents will be customized form the contents available. The chosen parameters will follow the central parameters of psychoperceptual experiments 1-2. Depending on the outcome of User-centered quality evaluation experiment 1, the comparison between laboratory and field experiment can be considered.
Independent variables:
Video bitrates: 3
Audio bitrates: 3
Error rates: 3
Code rate: 2-3
Contents: 5 (min. 60 s)
Number of participants: 60
Schedule: 01.06.2009-01.03.2010
Reporting: May 2010
Unit in responsibility: TUI
Other involved units: TUT, METU, HHI

Experiment 3: Field testing of the prototype

Research questions:
How the existing prototype and presented contents are experienced in the field settings and how acceptable they are?

Plan:
Goal of this study is to make sure that the system developed reaches the predefined acceptance thresholds (e.g. by 80% of users) in longer term usage. At the time of conducting the field experiment, the available contents and user-interfaces can critically impact on the final plans.

Number of participants: 30
Schedule: 01.05.2010-01.11.2010
Reporting: November 2010
Unit in responsibility: TUI
Other involved units: TUT, METU, HHI
6. Use Scenarios for mobile 3D television and video

6.1. General considerations

To make user requirements and user needs understandable for a development team, it is useful to derive user profiles and usage scenarios for a certain degree of visualization. The goal of these user profiles and usage scenarios is to model specific archetypes and exemplary usage stories out of general user requirements data (Kuniavsky 2003). One method to present your user profiles are personas.

Personas were created as an interaction design technique. Cooper (Cooper 1999) introduced personas as key part of his "Goal-Directed Design" method in 1999. According to Cooper (Cooper 1999) a persona is a models, not real people. The success of your product development is to "design for specific types of individuals with specific needs." (Cooper et al. 2007) A persona so samples a user archetype of your system or service. Personas are both an individual and a sample of one specific user class. The persona thereby is created within your user profiles and so the development process relies on user requirements that were collected at the beginning of the user-centered development process (Strohmeier et al. 2008). First being rough sketches of how users will look like, personas are more and more described in detail to support other quantitative and qualitative methods.

Cooper (Cooper et al. 2007) describes the strength of personas as a design tool in terms of:

- "Determining" the functionality of your product to direct the design effort.
- "Communicating" in a common language with all parties of the development process.
- "Building consensus and commitment" by using a common language that can help to create common understanding of design questions. Personas as models of real users are more explicit than data or flowcharts.
- "Measuring" the effectiveness of your design. Relating design progress and new features to personas allows discussing design problems on the whiteboard. This makes a solution process faster and cheaper as it would be if you'd need to ask real people.
- "Contributing" to all kind of product-related efforts. Personas can also be used to advertise your product or at least to deliver information to marketing or strategic planning as personas give a detailed impression about the final end-user.

Pruitt and Grudin (Pruitt and Grudin 2003) share these strengths of Persona design. They outline comparable advantages as follows:

- If personas are used as a complementary task among other methods during the user requirement elicitation process personas can be a useful tool to raise the effectiveness of those methods. They make the view on the users more explicit by adding focus on your future target groups.
- Personas can communicate information of different research partners in one project. Personas are understandable for designer, engineers and managers. Design questions can be better discussed by referring to the personas. ("Would our personas use this feature?")
- Personas are explicit. They are not data or abstract ideas.
- As personas are explicit, they can be used to transport results within the project and also to the public audience.

6.2. Personas in MOBILE3DTV

Within the project MOBILE3DTV personas are a very interesting design concept to make our user requirements more explicit. Mobile 3D television and video systems and services do not exist yet. However, first studies on user requirements have shown that future users have very clear ideas of what mobile 3DTV will be like, how it will be used, and where they will use it. The results have given
very complex knowledge of usage contexts, preferred contents, and system and service requirements of the users. For us, personas are a suitable possibility to communicate user requirements within the development process, but also to present user archetypes for the public to answer the question: "For whom do you develop mobile 3DTV and how will I use it?"

6.2.1. Scenarios

As described above Personas are individuals but sample a specific user class at the same time. As individuals every persona has got own motivation, expectations, and usage goals of the product under development. These characteristics are composed of real data measured among different user classes. This is why Cooper (Cooper et al. 2007) calls personas also composite user archetypes. But how can we now communicate these different motivation and expectations of each single individual to communicate design requirements?

"Scenarios are stories that describe how a person behaves or thinks about a task or a situation." (Kuniavsky 2003). "Persona-based scenarios" (Cooper et al. 2007) bring personas to life. They show personas interacting with your product during the personas' daily life. In the early development stage of a product development, scenarios are a suitable way to present future users' usage motivations and expectations about a new product. Cooper calls these early stage scenarios context scenarios that "describe the broad context in which usage patterns are exhibited and include environmental and organizational [...] considerations." (Cooper et al. 2007).

Goal of the context scenarios is to visualize the earlier elicited user requirements. So scenarios present the critical tasks in product development focusing always on the usage goals that will be achieved at the end of the whole development process. They are not communicating a status quo. This fact also relates to the need of more than just one scenario. Context scenarios are as multiple as will be the usage contexts of the end product. And as do personas the scenarios will grow in parallel to the product development adapting to new requirements, but also including constraints or exceptions. (Cooper et al. 2007)

6.2.2. Personas and Scenarios in MOBILE 3DTV

We already said in the previous section that persona development and later the formulation of context scenarios requires real data to relate personas and scenarios to user requirements. In MOBILE3DTV these user requirements were collected in a methodological triangulation of online survey, focus groups, and probe study resulting in a broad view on user requirements. These requirements can be seen as first results of user requirements in the context of mobile 3D television and video and were used in a first step to derive design guidelines for user experience. (Strohmeier et al. 2008, Jumisko-Pyykkö et al. 2008).

In a next step we now derive personas and context scenarios from these user requirements. The main focus is thereby on presenting the different usage motivations relating also to multiple usage contexts and contents of mobile 3D tv and video. Following we present five personas as a first description of the future system and services of MOBILE3DTV.

39
## Sanna Virtanen

### Person description
- **Age:** 28
- **Status:** long-distance relationship since 4 month with Rául Gomez, a working colleague from Madrid
- **Location:** Tampere, Finland; Owns a 2-room apartment
- **Profession:** Managing Consultant at WeCan OY – a medium sized software company – in her 2nd year. Mari is responsible for project acquisition and project coordination.
- **Income:** 50'000 €/year

### Psychographic data
- **Social network:** Some few old friends, spread all over the world, otherwise loose social contacts
- **Personality:** very self-confident, influencer, active,
- **Beliefs:** Projects her own demands onto others; Only the best is good enough
- **Attitude:** Behave according to your status
- **Expect the best from yourself—and others—and products
- **Lifestyle motivations:** goal-oriented; early adopter
- **Hobbies/Interests:** Pilates, Shopping, Design, Gossip

### Usage motivation for mobile 3DTV
- **Sanna is a highflyer in her job. She loves the job, although it is very exhausting and she’s always busy at work. But in the same manner in which she loves her job, she lives her life after work. She is a lifestyle woman, always up-to-date and trendy. She bought her mobile 3D phone 5 months ago. Her last phone contract was ending and as she had the possibility to choose a new phone, she decided for mobile 3D. The advertisement and the latest reports in the lifestyle magazines convinced her that this phone fulfills all her demands on functionality and quality.**

### Scenario
At the moment, Sanna is just about to leave from work. It’s 5:30pm and her Pilates workout will start at 8pm. It’s a nice possibility to meet some friends at the near coffee shop where Sanna often takes a cup of coffee before returning home. Before leaving from work, while synchronizing her calendar with the online backup she checks the available downloads of the latest “Sex in the city” 3D episode which she missed yesterday. Meeting friends also means the need to be informed about the latest gossip, of course. After she started the download, she leaves to the coffee.

“I enjoy sitting here and watch the latest lifestyle magazines or all these cute shows on TV like this ‘Finland’s next supermodel’. 3D makes everything so realistic and I enjoy diving into the content so that I can forget the day for a short time.” Having a cup of Iced Moccha she chills on the sofa and watches the series. As the clips ends and the girls are about to arrive in 10 minutes, she checks the best-of-clips of the brand new talent show online as the café offers free high-speed WiFi access.

Sanna’s day ends after her Pilates class at 10pm. It was nice to meet the girls again, to chat, to exchange news and of course to check the latest trends of cloths and gimmicks online. It’s fun to compare products in 3D and as her boyfriend got another mobile 3D phone, it is even more fun to compare two products parallel. But now, it’s time to call Rául. Secretly, the 3D videophony was the final reason to choose the mobile 3D phone. “You know, it is just this higher intimacy that I experience..."
when Rául appears in front of me three-dimensionally. For a short time I even forget the annoying distance between us”, Sanna says with a smile on her face.

**Mikko Laaksonen**

|Person description| Age: 24  
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Status:</td>
<td>Single</td>
<td>Location:</td>
<td>Lives in Lappeenranta, Finland in a student accommodation in the suburb of the city</td>
<td></td>
</tr>
<tr>
<td>Profession:</td>
<td>Master Student at University of Applied Sciences</td>
<td>Income:</td>
<td>10,000 EUR/year</td>
<td></td>
</tr>
</tbody>
</table>

**Psychographic data**

|Social network: | Few selected friends; some intimate chat friends | Personality: | Introverted; good teacher; keen to debate; knowledge oriented |  
|Beliefs: | Everything can be made better; Only believe what you’ve seen by yourself | Attitude: | Be curious towards innovation; push the progress; share knowledge with others |  

|Lifestyle motivations: | High interest in technological progress | Hobbies/Interests: | Scientific literature, computer games, 3D cinema, modern entertainment media |  

**Usage motivation for mobile 3DTV**

Mikko has followed the development of mobile 3DTV for a long time. Already when he read about it on the news ticker for the first time 3 years ago, he was interested in the project and the development progress. Technical progress is Mikko’s favorite interest and the possibility to have a three-dimensional screen on his mobile device has been his dream since he was in a 3D cinema for the first time.

**Scenario**

This morning, Mikko goes to university by bus. He is still tired as he slept just few hours last night. Mikko is currently working on his student project and he had the idea to visualize their results on his 3D screen for better visualization. Although it cost time and sleep, he is now very proud of the results and his abilities to work with the mobile 3D presentation. Mikko quickly checks the presentation once more before he plugs in his headphones to watch the last episode of CSF, the crime serie that was broadcasted for the first time on ‘3D Crime’ last week. “3D is a good possibility to relax. Mostly in the morning when I am still tired I am quite bored of all the people around me. In these moments I just switch on my mobile 3D device and watch some movie clips, trailers or crime series – just to dive into the content.”

Mikko first doubted that 3D would work on a mobile device. He already knew 3D from the cinemas. To buy a mobile 3D device was just the logical result of his interest in the research. Arriving at school, Mikko visits his student work supervisor to present his work. The presentation goes well and Mikko honestly admits that he loves his device more and more.

“It’s not only the 3D experience, it is a question of good content, good services and the fascinating experience”, says Mikko as he arrives at home back from work. His device automatically connects to Mikko’s WiFi and Mikko gets all new emails pushed onto his mail application. One mail contains a short 3D video clip that a friend who visits the Games Convention in Germany sent to show the new
big autostereoscopic displays. Mikko is sure that he will need one of these large screens, too. But he doesn’t want to miss his mobile 3D device.

**Marcel Jung**

| Person description | Age: 17  
| Status: Single  
| Location: Lives in a suburb of Nuremberg (pop. approx 500’000), Germany, still with his parents  
| Profession: 1st year cook apprentice at “Koch und Bring”, a mid-class catering service  
| Income: 550 EUR/month  

| Psychographic data | Social network: Got new friends (circle of friends) through his apprenticeship; sparse relation to his family  
| Personality: healthy self-confidence, influenceable, calm  
| Beliefs: Possession can determine the social role in a group; Advertisements show modern way of living; Gaining independence is an important value from child to adult  
| Attitude: Spend the money you earn; show your financial independency; show your status  
| Ideology: work to get money to buy entertainment  
| Lifestyle motivations: Status-oriented and action-oriented; early adopter (in his financial opportunities)  
| Hobbies/Interests: Cars, meeting friends (his clique), motor sports, basketball  

| Usage motivation for mobile 3DTV  

**Scenario**

It’s Wednesday, at 15.30h. Marcel Jung, our 17-year-old apprentice, comes home from work. On the walk from the bus-stop to his family's home he gets a call from a colleague from vocational school. He is invited to drop by at 17.00h at the Jakobsplatz in the city of Nuremberg, the place where he and his friends use to meet. Marcel happily agrees, because he had an argument with his father the day before and does not want to be at home for too long. After he comes home he rewarms some food his mother prepared for him. While he eats, he prepares his mobile3DTV device to impress his friends. As some kind of ritual, he likes to check the 3D video platforms and to have the latest 3D clips on his phone to show them around.

As he leaves the house, he turns on the music player on his device as he needs to hurry to catch the bus. On the bus, he switches the display on to watch the corresponding music videos. But although he really likes the song, the clip bores him. So he decides to rather play some three-dimensional RISTET, a very popular game since mobile 3DTV is available on the market.

Of course, his friends are eager to watch the 3D clips he downloaded as he is the first one in his clique to own such a device. Marcel enjoys the prestige that he gains through his device. Impressed of the latest trailer of the action movie "Wetman: Light Knight", they’re thinking about going to the MAXI cinema after Marcel showed around the trailer. But as it’s the end of month everyone is rather short of money. They therefore decide to postpone it to the next month. So the talk turns to the same topic as always: cars, one of their favourite pastimes. Suddenly, one of Marcel's friends, who is already 18 and
dreams of buying his own car soon, remembers that there is a three-dimensional presentation of the new Terault Lico, the car he favours most, available on the manufacturer’s website. Marcel lends him his device, of course, and his friend watches the presentation to have a close look at a 360°-model of the car. Everyone agrees that the 3D presentation gives much better experience than the photos of the car in the magazines. 

Back home, Marcel manages to avoid his parents and goes to bed early, as his working day will start early tomorrow. While lying in bed, he reviews the car presentation and fascinated wonders, what car he is going to buy when he will be 18.

Hannah Siebert

| Person description | Age: 32  
|-------------------|---------|--------|------------------|-----------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Status:           | Single  
| Location:         | Lives in Berlin in a shared flat in "Prenzlauer Berg"  
|                   | Profession: Graphic designer at "Kunze und Partner", an advertising agency  
| Income:           | 35.000 EUR/year

| Psychographic data | Social network: Lots of Facebook friends, knows a lot of people loosely  
|--------------------| Personality: Athletic, very creative, open-minded, mature personality, goal oriented  
|                    | Beliefs: It’s possible to combine duties and hobbies; one should be living here and now; friends are important—the more, the better  
|                    | Attitude: be willing to try something new; be open to new experiences; leave the city as often as possible  
|                    | Ideology: Spend your money on adventures  
|                    | Lifestyle motivations: Tries to combine job and hobby  
|                    | Hobbies/Interests: Traveling, surfing, dancing

| Usage motivation for mobile 3DTV | Hannah has always been looking for new experiences. She saw mobile 3DTV for the first time some months ago during her last surf vacation to the USA. One of her friends checked the local weather forecast on his mobile 3D phone and Hannah was quite excited about the 3D experience and the reality of the web clips that were presented for all locations. Hannah bought her mobile 3D phone together with the entertainment and travel packages that were offered by her provider. At first, Hannah used the device most of the times to get some time to relax by watching some short animations when she was hanging around with her friends. But Hannah has learned to take advantage of her phone and she even uses it for 3D navigation purposes during her journeys.

Scenario

This Friday afternoon, Hannah is sitting on the sofa at home in her room. Her next holiday was confirmed this morning and Hannah plans her trip. Earlier, she always had plenty of catalogues and brochures around her to make plans. Because of the travel package she has got access to ‘Trip3D’, an online community where other users post their 3D videos of the last holidays and discuss different locations. Hannah loves to just surf through the websites and to get inspired of the videos. "This experience of reality and naturalness of the videos just causes a feeling of already being there while
sitting on the sofa. I just can imagine walking along the beach or taking my surfboard and catch the waves. If I feel very comfortable with a certain location, then I just decide to go there”, says Hannah. To further discuss the surf spots she found in ‘Trip3D’ Hannah starts the 3D online community ‘RealLife’ where she always finds some friends to chat. Luckily, Frank and the others are already online and Hannah proudly reports that she is planning her next trip. Her virtual friends in ‘RealLife’ are very important for Hannah as she knows that they all share her interests and hobbies. "Although they are spread all over the world I feel as to meet them every time when I am online. The three-dimensional world connects us and, you know, it’s just a great experience to see your friends more real that just flat avatars.”

**Stefan Weber**

<table>
<thead>
<tr>
<th>Person description</th>
<th>Age: 36</th>
<th>Status: Married to Nicole Weber, 31 (bookseller), 1 child (Maximilian, 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location:</td>
<td>Lives in Overath near Cologne (pop. approx. 27'000), Germany in a semidetached house</td>
<td></td>
</tr>
<tr>
<td>Profession:</td>
<td>Sales representative at “Hofmann &amp; Söhne”, a manufacturer of ceramic tableware</td>
<td></td>
</tr>
<tr>
<td>Income:</td>
<td>38'000 €/year</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychographic data</th>
<th>Social network: Nice neighborhood with other young families, close relation to other parents, sports club</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality:</td>
<td>Calm, warm and friendly, eloquent</td>
</tr>
<tr>
<td>Beliefs:</td>
<td>Values family; Safety thinking and future planning is very important</td>
</tr>
<tr>
<td>Attitude:</td>
<td>Care for family and child; Travels a lot with family</td>
</tr>
<tr>
<td>Ideology:</td>
<td>Work to get money to buy entertainment</td>
</tr>
<tr>
<td>Lifestyle motivations:</td>
<td>Outdoor activities; late mainstream</td>
</tr>
<tr>
<td>Hobbies/Interests:</td>
<td>Hiking, Culture (mainly Museums), Football, Reading</td>
</tr>
</tbody>
</table>

| Usage motivation for mobile 3DTV | Stefan got his mobile 3D device from his company. The sales department decided to take advantage of three-dimensional product presentations and changed from printed catalogues to virtual 3D presentations. Stefan has been equipped with his mobile 3D video device since 6 months. As he has been very impressed from the 3D presentations and the features that the device offers additionally, he asked the sales department for private usage which was accepted. He introduced the device to his family and Maximilian immediately liked the short videos that his father showed. In contrast, Nicole has been rather skeptical. But the more she has used the different applications and services the more she likes the device. Especially the information guides for cities and locations appeal to her. |

**Scenario**

It’s Saturday and sunny weather is expected for the whole weekend. Stefan and Nicole decide to go to the zoo in Cologne as Maximilian likes the animals. For Stefan, it is important to spend the weekend with his family. Before they leave, Nicole reminds Stefan to check if there is a visitors’ guide available in the 3D tourist application. Zoo Cologne indeed advertises their service ‘3D safari’ – a three-
dimensional guidance through the zoo which allows mobile 3D users to watch short documentary clips about different animals. The service will be included in the entrance fee for the family pass. While driving to the zoo, Maximilian is allowed to watch 'Experiment', a teaching program for children and Maximilian's favorite TV show, "because it is in 3D." Usually, he is not allowed to use Stefan's mobile 3D device, but this program is the only exception. Arriving at the zoo, Stefan buys the tickets and he asks about '3D safari'. He gets an access code for the service so that he can download the application at the info terminal. Well equipped, they enter the zoo and '3D safari' guides them via the device included GPS module. Signs in front of different animal areas show that additional information is available. "It's amazing that Maximilian can watch these small documentaries and inform himself about the animals in the wild life. He has always loved the 3D experience. Look how fascinated he is of this lion video." Stefan wants to keep these nice family moments for his parents. Luckily his mobile 3D device includes a stereo video camera. He has used the video camera of his device for a long time to record short clips of the family which he presents to his parents whenever they meet. For Stefan, the 3D videos are more realistic and he was proud when his mother said that it is almost as being part of the family activities.
7. References


Cooper, Alan. The Inmates Are Running the Asylum. Indianapolis, IN: Sams, 1999.

Corriveau, P.; Gojmerac, C.; Hughes, B.; Stelmach, L.: All subjective scales are not created equal: The effects of context on different scales. Journal of Signal Processing, Volume 77, Number 1, p. 1-9, 1999


IJsselsteijn, Wijnand; Ridder, Huib de; Hamberg, Roelof; Bouwhuist, Don; Freeman, Jonathan : Perceived depth and the feeling of presence in 3DTV. In: Displays, Jg. 18, H. 4, S. 207–214.


Jumisko-Pyykkö, S. "I would like to see the subtitles and the face or at least hear the voice: Effects of Picture ratio and Audio-video Bitrate Ratio on Perception of Quality in Mobile Television.". Multimedia Tools and Applications (Journal), vol 36, No 1-2, 2008. DOI - 10.1007/s11042-006-0080-9


Nyman, Gote; Radun, Jenni; Leisti, Tuomas; Oja, Joni; Ojanen, Harri; Olives, Jean-Luc et al. : What do users really perceive: probing the subjective image quality. In: Proc. SPIE, Jg. 6059, H. 1, S. 605902–605907.


Usability Professionals’ Association UPA. 2008
Wynekoop, J. L., Russo, N., L, Studying system development methodologies: an examination of research
Yano, S., Emoto, M., Mitsuhashi, T., Two factors in visual fatigue caused by stereoscopic HDTV images, Displays, Volume 25, Issue 4, Pages 141-150, November 2004
MOBILE3DTV - Mobile 3DTV Content Delivery Optimization over DVB-H System - is a three-year project which started in January 2008. The project is partly funded by the European Union 7th RTD Framework Programme in the context of the Information & Communication Technology (ICT) Cooperation Theme.

The main objective of MOBILE3DTV is to demonstrate the viability of the new technology of mobile 3DTV. The project develops a technology demonstration system for the creation and coding of 3D video content, its delivery over DVB-H and display on a mobile device, equipped with an auto-stereoscopic display.

The MOBILE3DTV consortium is formed by three universities, a public research institute and two SMEs from Finland, Germany, Turkey, and Bulgaria. Partners span diverse yet complementary expertise in the areas of 3D content creation and coding, error resilient transmission, user studies, visual quality enhancement and project management.

For further information about the project, please visit www.mobile3dtv.eu.

Tuotekehitys Oy Tamlink
Project coordinator
FINLAND

Tampereen Teknillinen Yliopisto
Visual quality enhancement,
Scientific coordinator
FINLAND

Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V
Stereo video content creation and coding
GERMANY

Technische Universität Ilmenau
Design and execution of subjective tests
GERMANY

Middle East Technical University
Error resilient transmission
TURKEY

MM Solutions Ltd.
Design of prototype terminal device
BULGARIA