Hindawi Publishing Corporation
Advances in Multimedia
Volume 2010, Article ID 658980, 28 pages
doi:10.1155/2010/658980

Research Article

Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception

D. Strohmeier,[1] S. Jumisko-Pyykkö,[2] and K. Kunze[1]

[1] *Institute for Media Technology, Ilmenau University of Technology, 98693 Ilmenau, Germany*
[2] *Unit of Human-Centered Technology, Tampere University of Technology, 33720 Tampere, Finland*

Correspondence should be addressed to D. Strohmeier, dominik.strohmeier@tu-ilmenau.de

Received 29 March 2010; Accepted 7 July 2010

Academic Editor: George Ghinea

To quantify the excellence of multimedia quality, subjective evaluation experiments are conducted. In these experiments, the tradition of quantitative assessment is the most dominating, but it disregards the understanding of participants' interpretations, descriptions, and the evaluation criteria of quality. The goal of this paper is to present a new multimedia quality evaluation method called Open Profiling of Quality (OPQ) as a tool for building a deeper understanding on subjective quality. OPQ is a mixed method combining a conventional quantitative psychoperceptual evaluation and qualitative descriptive quality evaluation based on the individual's own vocabulary. OPQ is targeted for naïve participants applicable to experiments with heterogeneous and multimodal stimulus material. The paper presents the theoretical basis of the development of OPQ and overviews the methods for audiovisual quality research. We present three extensive quality evaluation studies where OPQ has been used with 120 participants. Finally, we conclude further recommendations of use of the method in quality evaluation research.

1. Introduction

To become successful, new multimedia systems and services need to meet the user's requirements, offer pleasurable experiences, and provide higher quality than the existing systems. At the same time, audiovisual systems are becoming more and more complex as technological progress provides new possibilities of presenting content. For example, audiovisual 3D on portable devices requires a high level of optimization of technical resources to handle huge amounts of data, with possible limitations due to transmission channel and device constraints. This can result in perceivable heterogeneous impairments in the value chain from content production to display techniques and influence the user's perception of quality. To assess the experienced quality of these novel systems and services, subjective audiovisual quality evaluation experiments are conducted.

Subjective (~perceptual, affective, experienced, sensorial) quality evaluation is based on human judgments of various aspects of experienced material based on perceptual processes [1–3]. These quality perceptions contain both a low-level sensorial and high-level cognitive processing, including knowledge, emotions, attitudes, and expectations [4–7]. Since 1970s, recommendations for video quality evaluations have offered a good basis for assessing one dimension of quality—its hedonistic excellence [8]. Recently, a broader view to quality has been taken by covering other aspects of active perception in the evaluations including in knowledge, different levels of human information processing, or even contextual behavior [9–15]. Although these evaluations have made a significant contribution for understanding quality, they are still limited to the investigation of quantitative quality preferences. Subjective impressions, interpretations, and experiences as factors to explain and understand the results (constructed in the evaluations of different system factors) beyond the excellence are rarely considered [16, 17]. This can be partly because of a lack of reliable explorative instruments for tackling the descriptive characteristics of quality or even more ambitiously for relating quality preferences and descriptions. The few previous attempts have been suggested to multimedia quality society, while they have constraints in terms of accuracy, complexity, required type of assessors, unimodal evaluations, or emphasis only on qualitative methods [16–21].

The goal of this paper is to present a mixed method called Open Profiling of Quality (OPQ) to understand the multimodal quality of experience. Mixed methods combine both quantitative excellence evaluation and qualitative descriptive research into one single study to compensate for the weaknesses of one method, expand understanding of phenomena, and provide complementary viewpoints [22–24]. For the method development we conduct a literature review in mixed methods, quality evaluation research in multimodal quality, and other related fields (food science, consumer research). The proposed method combines a conventional quantitative psychoperceptual evaluation and qualitative descriptive quality evaluation using the participant's self-defined vocabulary. It is applicable to naïve participants and heterogeneous multimodal stimuli material. We present three multimedia quality studies, where OPQ has been used to make further recommendations of use of the method in quality evaluation research. The method presented helps practitioners to conduct mixed method research in the field of audiovisual quality.

The paper is organized as follows: in Section 2, we outline the theoretical background of multimedia quality, a state-of-the-art of mixed method research and its use in audio, video, and audiovisual quality evaluation. Section 3 contains the description of the Open Profiling of Quality method. Sections 4–6 present the three studies using the method presented. Finally, discussion and conclusions of the method and its further use are described in Section 7.

2. Related Work

2.1. Understanding Quality Perception. Multimedia quality is characterized by the relationship between produced and perceived quality. Produced quality is determined by the technical factors of multimedia, typically categorized into three different abstraction levels: content, media, and network [25, 26]. The special requirements for produced multimedia quality can result in the juxtaposition of a huge amount of multimedia data and limited bandwidth, a vulnerable transmission channel, and constraints of receiving devices. Perceived (also called experienced, sensorial) quality describes the users' or consumers' view of multimedia quality. It is characterized by active perceptual processes, including both bottom-up and top-down and low-level sensorial and high-level cognitive processing [4, 6]. The relationship between perceived and produced quality for the end-to-end systems is described in terms of Quality of Experience (QoE) as "*The overall acceptability of an application or service, as perceived subjectively by the end-user*" [27]. More broadly, Wu et al. [28] have summarized it "*as a multidimensional construct of user perceptions and behaviors.*"

Quality perception is constructed in an active multilayered process. It contains the extraction of relevant features of the incoming sensorial information in its early stage (e.g., brightness, form, and motion for vision or pitch, loudness, and location for audio) [6]. However, quality perception is not only determined by the stimuli-driven bottom-up processing. The high-level top-down cognitive processing involves individual emotions, knowledge, expectations, and schemas representing reality which can weight or modify the importance of each sensory attribute, enable contextual behavior and active quality interpretation [4–7]. In this stage, stimuli are interpreted according to their personal meaning and relevance to human goal-oriented actions. Related to multimodal perception, one sensory channel can complement and modify the perception derived from another channel [29]. This dependency can be due to clarity of stimuli, as well as content, task, and context [29–32].

Multimedia quality studies aim at optimizing quality factors produced under strict technical constraints or resources with as little negative perceptual effects as possible. Recent multimodal quality evaluation studies have started to underline the characteristics of active and multilayered quality perception. Quality does not only derive from the characteristics of stimuli, but also from usage-, task-, and context-dependent factors [14, 16]. In this paper, we want to continue to work on this track, see human perception as an interesting challenge, and develop explorative tools for understanding underlying attributes of perceived quality.

2.2. Research Methods for Perceived Quality Evaluation

2.2.1. Quantitative Psychoperceptual Evaluation. Psychoperceptual quality evaluation is a method for examining the relation between physical stimuli and sensorial experience following the methods of experimental research. These methods have their origin in classical psychophysics of the 19th century, and they have been later applied in uni- and multimodal quality assessment [2, 3, 8, 33]. In the quality evaluation domain, the applied methods are standardized in technical recommendations by the International Telecommunication Union (ITU) or the European Broadcasting Union (EBU) [8, 33, 34]. The goal of these methods is to analyze quantitatively the excellence of perceived quality of stimuli in a test situation. Psychoperceptual quality evaluation studies are characterized by a high level of control over the variables and test circumstances and can include the use of standardized test sequences, procedures, and the categorization of participants to naïve or professional evaluators to ensure the repeatability of study. As an outcome, subjective quality is expressed as an affective degree-of-liking using mean quality satisfaction or opinion scores (MOS). In psychoperceptual studies, the variety of quality being tested and the research question define the applicable method. Single stimulus methods are useful for evaluations of the large quality range from low to high with detectable differences between stimuli, while pairwise comparisons are powerful when comparing stimuli with small differences [8, 33]. A common single-stimulus method is Absolute Category Rating (ACR). The test method includes a one-by-one presentation of short test sequences at a time. Every test sequences is then rated independently and retrospectively using a 5/9/11-point scale [33]. In multimedia quality assessment, ACR has outperformed other evaluation methods [35, 36].

Table 1: Descriptive perceived quality evaluation.

Methodological approach	Interview-based approach	Consensus vocabulary profiling approach	Individual vocabulary profiling approach
Methods using this approach	Interpretation-Based Quality evaluation [17], Experienced quality factors [16]	Flavor Profile Method [3], Texture Profile Method [3], Quantitative Descriptive Analysis [3], RaPID [18], ADAM [45]	Free-Choice Profiling [47], Flash Profiling [48]
Vocabulary Elicitation	Interview	Group discussions and consensus attribute list	Individual attribute list, supporting task like Repertory Grid Method applicable
(Statistical) Analysis	Coding (e.g., Grounded Theory) and Interpretation	ANOVA, MANOVA, PCA	GPA, PCA
Participants	15 or more naïve test participants	Around 10 highly trained participants	Around 15 naïve test participants
Used in Mixed Method approaches	Yes	No	No
Applied in audiovisual quality research	IBQ [17, 49], Experienced Quality Factors [16]	RaPID [18], ADAM [45]	IVP [19, 20]

Recently, conventional psychoperceptual methods have been extended from hedonistic assessment towards the evaluation of appropriateness to use- and goal-oriented actions (cf. overview [37]). Quality is measured as a multidimensional construct of cognitive information assimilation or satisfaction constructed from enjoyment and subjective, but content-independent objective quality [11, 12, 31, 38]. Furthermore, evaluations of acceptance act as an indicator of service-dependent minimum. The useful quality level has been established parallel to the assessment of quality satisfaction in the laboratory and natural contexts of use [14, 15, 39–41]. However, all quantitative approaches lack the possibility to study the underlying quality rationale of the users' quality perception.

2.2.2. Descriptive Quality Evaluation. Descriptive quality evaluation approaches focus on a qualitative evaluation of perceived quality. The basic idea is that test participants are asked to describe their quality factors or the reasons for a certain quality rating. In more advanced methods, these expressions or quality attributes are used to rate test items in a subsequent task. We identified three main approaches: (1) interview-based approach, (2) consensus vocabulary profiling, and (3) individual vocabulary profiling which differ in terms of vocabulary elicitation methods, methods of analysis, and characteristics of participants (Table 1).

Interview-Based Evaluation. In the existing interview-based methods, naïve participants describe explicitly the characteristics of stimuli, their degradations or personal quality evaluation criteria under free-description or stimuli-assisted description tasks [16, 40–42]. The goal of these interviews is the generation of terms to describe the quality and to check that the test participants perceived and rated the intended quality aspects. Semistructured interviews are commonly used. They are especially applicable to relatively unexplored research topics, constructed from main and supporting questions, and, compared to open interviews, they are less sensitive to interviewer effects [43]. The frameworks of data-driven analysis are applied and the outcome is described in the terms of the most commonly appearing characteristics [16, 17, 21].

Consensus Vocabulary Profiling. The "RaPID perceptual image description method" (RaPID) is based on a descriptive analysis assuming that image quality is the result of a combination of several attributes and that these attributes can be rated by a trained panel of assessors [2, 18, 44]. Its purpose is to develop a consensus vocabulary. Later, trained test participants rate quality, based on the vocabulary. A multistep procedure contains (1) extensive group discussions where panel members first develop a consensus vocabulary of quality attributes for image quality; (2) a refinement discussion where the panel then agrees about the important attributes and the extremes of intensity scale for a specific test according to the test stimuli available; (3) an evaluation task where each test participant applies each attribute for a set of stimuli in a pair comparison of the test stimulus and a fixed reference. RaPID requires extensive and time-consuming panel training, can be sensitive to context effects, and requires an experienced researcher to conduct the experiments [18]. A comparable methodology is used for audio evaluation in the Audio Descriptive Analysis and Mapping (ADAM) technique [45].

Individual Vocabulary Profiling. In contrast to consensus vocabulary profiling, Lorho's Individual Profiling Method (IVP) is a descriptive quality evaluation for naïve participants. His work was the first approach in multimedia quality assessments to use individual vocabulary from test participants to evaluate quality. The procedure contains four steps. (1) Familiarization—participants become familiar

Table 2: Mixed method designs according to Creswell and Plano Clark [50].

Mixed method design	Design pattern	Purpose
Triangulation design	Independent collection of QUAN and QUAL data. Interpretation based on both data sets.	Comparison of QUAN and QUAL results for a broad interpretation of the results
Embedded design	One data set is used in a supplemental role in studies primarily based on the other data set.	Additional qualitative expressions about quantitative results (e.g., supporting decisions about further studies or tasks)
Explanatory design	Two-step design. First collection of QUAN, then QUAL.	QUAL data may be needed to explain unexpected results or to detect errors in the QUAN research design.
Exploratory design	Two-step design. First collection of QUAL, then QUAN.	QUAL data may be needed to explain unexpected results or to detect errors in the QUAN research design.

with describing the attributes of stimuli, and they develop their individual vocabulary in two consecutive tasks. (2) An attribute list is generated in a triad stimulus comparison using an elicitation method called Repertory Grid Technique. (3) The developed attributes are used to generate scales for the evaluation. Each scale consists of an attribute and its minimal and maximal quantity. (4) Test participants train and evaluate quality according to the attributes developed. The data is analyzed through hierarchical clustering to identify underlying groups among all attributes and Generalized Procrustes Analysis [46] to develop perceptual spaces of quality. Compared to the other descriptive methods, the four-step procedure for individual vocabulary training can be time consuming. However, analysis of IVP is relatively easy and the location of the researcher's interpretive process is at the very end compared to interview-based methods. Although the paper shows that there are various methods to study perceived multimedia quality quantitatively and qualitatively, the methods do not combine both approaches (in Table 1). We see the challenge of modern evaluation methods also in the combination of both data sets.

2.3. Mixed Method Research

2.3.1. The Theory of Mixed Method Research. Fundamentally, mixed method research has its roots in pragmatic philosophy, represents the third wave of research methods, and is suitable for applied research, such as quality evaluation [22]. It is defined as the class of research in which the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts, or language into a single study [22]. The major characteristics of traditional quantitative (QUAN) research are a focus on deduction, confirmation, theory/hypothesis testing, explanation, prediction, standardized data collection, and statistical analysis [22]. The major characteristics of traditional qualitative (QUAL) research are induction, discovery, exploration, theory/hypothesis generation, the researcher as the primary "instrument" of data collection, and qualitative analysis [22]. To combine these two research traditions, mixed methods are used to provide complementary viewpoints, to provide a complete picture of phenomena, to expand understanding to phenomena, and to compensate for the weaknesses of one method [23]. The core of mixed method theory is the combination of quantitative and qualitative methods into one final result. There are four main design patterns to fuse these methods with slight differences in the emphasis of dominating method, their interdependency, and the purpose (Table 2) [50].

Triangulation is the most common and important mixed method design (Table 2) [50]. In triangulation, data collection and analysis are carried out independently for QUAN and QUAL methods with no preference, and the final inference aims at creating a broad picture of the phenomenon [50]. Three possible outcomes can be expected in these studies: (1) the convergence of results where both results lead to the same conclusions, (2) the complement of results where the different results highlight different aspects of the same phenomenon, or (3) the results can be divergent or contradictory [24]. The ideas of triangulation and other mixed method designs (Table 2) have been used in quality evaluation research although researchers have not explicitly expressed the relationship to this methodological approach, for example, [16, 17, 42]. Following, we present a review of the existing methods using mixed quantitative and qualitative evaluation of quality.

2.3.2. Mixed Methods in Audio, Visual, and Audiovisual Quality Evaluation. In multimedia quality evaluation methods, triangulation is the applied mixed method design. Jumisko-Pyykkö et al. [16] have introduced an approach of combined quantitative psychoperceptual evaluation and posttask interviews to explore experienced quality factors for audiovisual quality with naïve test participants. Psychoperceptual evaluation follows the ITU recommendations to collect overall quality [8, 33]. The experienced quality factors were collected using a Semistructured interview with a free-description task to describe the quality evaluation criteria used during the quantitative evaluation. Data-driven analysis, following the framework of Grounded Theory, was used in the interview analysis [51]. The results have underlined that experienced quality is constructed from the impressions of (1) low-level features of stimuli (e.g., audio, video, audiovisual impairments), (2) high-level factors (e.g., relationship of quality to use, content), and (3) the most varied variable representing the peaks or extremes of quality

[16, 40, 42]. Finally, the interpretation of both quantitative and qualitative data was firstly carried out independently, and the interpretations were integrated to support each other's conclusions. This method may suffer from inaccuracy as the descriptions are related to a set of stimuli instead of single stimulus. However, the descriptive task is fast to conduct and can be easily adapted to the quality evaluations in challenging circumstances (e.g., field) [40].

Triangulation is also applied in the method called Interpretation Based Quality (IBQ) [17, 21], adapted from [52, 53]. IBQ also follows a two-step procedure with naïve participants: (1) a classification task using free-sorting and an interview-based description task and (2) the psychoperceptual evaluation based on one quality attribute. In the perceptive free-sorting task, test participants form groups of similar items and describe the characteristics of each group. The free-sorting task with naïve participants produces comparable results to consensus vocabulary approach with expert participants in terms of describing the same sensations and the related wording of the attributes [52]. However, the costs of free-sorting are lower because of naïve test participants, missing training, and fast assessment of a large test set [52]. Extending the idea of a free-sorting task, IBQ allows combining preference and description data in a mixed analysis to better understand preferences and the underlying quality factors in a level of a single stimulus [17]. However, the analysis of interview-based methods for large data sets is time consuming as it requires a multistep procedure and interrater reliability estimations. In contrast to the original definition of the method [17, 21], the term IBQ has been inconsistently used later to refer to monomethodological designs and variable procedures of descriptive tasks [49, 54]. In this paper, we refer to IBQ as it was originally presented as a research method.

Summarizing the review of related work, quality evaluation research has slowly started to extend its approach from quantitative excellence evaluation towards descriptive and mixed methods to create a broader understanding of quality. There are two main approaches in the descriptive quality research: interview and vocabulary-based approaches, both applicable to naïve participants. However, the most accurate versions of these methods have been only applied for the assessment of unimodal quality. The goal of this paper is to develop a new quality evaluation method, which uses the mixed method approach to create a deeper understanding of multimodal quality and which is applicable to naïve participants.

3. Open Profiling of Quality

Open Profiling of Quality (OPQ) is a mixed method that combines the evaluation of quality preferences and the elicitation of idiosyncratic experienced quality factors. It therefore uses quantitative psychoperceptual evaluation and, subsequently, an adaption of Free-Choice Profiling. OPQ is "open" in terms of being "free from limitations, boundaries, or restrictions" [55] and "accessible to new ideas" [56] to understand the participants' construct of overall quality without restricting or constraining their descriptions. The term "profile" refers "to represent the outline (of something)" [56], targeting some kind of identity, characteristics, descriptions, and structure for the phenomenon under study. Finally, our method aims at capturing the dualistic nature of excellence and characteristics of quality according to its two central meanings as "the degree of excellence of something [56]" and as "a distinctive attribute or characteristic possessed by—something [56]". The specific goals of an OPQ study are

(1) to define the excellence of overall quality for different stimuli using quantitative psychoperceptual evaluation methods;

(2) to understand the characteristics of quality perception by collecting individual quality attributes using qualitative sensory profiling methods;

(3) to combine quantitative excellence and qualitative sensory profiling data to construct a link between preferences and quality attributes;

(4) to provide a test methodology that is applicable to use with naïve test participants.

Following, we will present the OPQ method step-by-step, introduce its theoretical background, and describe the test procedure to conduct an OPQ study.

3.1. General Considerations. Open Profiling of Quality as a research method consists of three subsequent parts (see Figure 1): (1) psychoperceptual evaluation, (2) sensory profiling, and (3) external preference mapping. The studies with the first two methods are independently conducted and their data can be combined in the last method.

3.1.1. Test Participants. OPQ is designed to be applicable for naïve test participants with predefined sensorial acuity criterion. Naïve is defined as not meeting any particular selection criterion for assessment tests, neither has experience in the research domain nor in the evaluation task [1, 16, 57]. Naive participants are expected to give holistic quality evaluations and produce unbiased results due to lack of knowledge about the test stimuli and their production [47]. In contrast, the expert assessors are trained for accurate, detailed, and domain-specific evaluation tasks (e.g., visual artifacts) [58]. A certain sensorial acuity level is required from participants to make sure that the results are not biased by sensorial inaccuracy (e.g., the sensorial acuity tests used myopia, hyperopia (Snellen index: 20/40), color vision according to Ishihara, hearing threshold with respect to ISO 7029 [59], and, in our cases, 3D vision using Randot Stereo Test ($\leq$60 arcsec)).

More broadly, the sample selection contributes to the external quality of the study and defines how well the results from the sample tested generalize to some broader population of interests [60]. The recommended number of participants according to ITU recommendations is at least 15 [8, 33]. However, we recommend 25–30 participants for the psychoperceptual evaluation to provide good statistical conclusion validity in within-subject designs [61]. For sensory

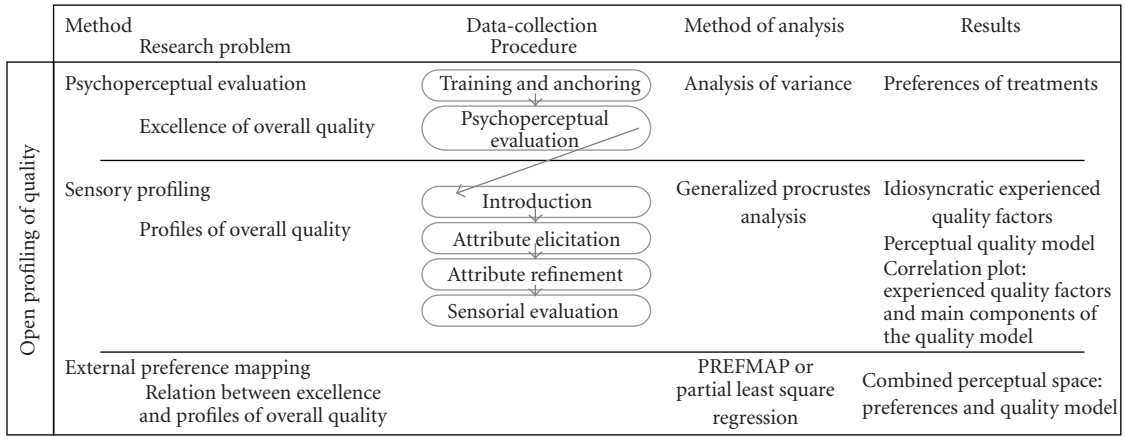


Figure 1: Overview of the subsequent steps in Open Profiling of Quality including respective research question.

profiling and the external preference mapping, a minimum of 12–20 participants is needed [62]. In the optimum case, all assessors participate in both parts of evaluation while the selection of a representative subsample can be considered.

3.1.2. Scheduling the Experiments. The psychoperceptual evaluation task is conducted prior to the sensorial evaluation. Although the order of the tasks may not have an impact on the outcome, as proved in [63], it is recommended to begin with the psychoperceptual evaluation as assessors are "clear of influence" [63]. In addition, the following profiling task can be done more precisely due to the already existing comprehension of the product under test [63].

The experiments are divided into several sessions. Depending on the amount and length of the test stimuli as well as the final design of each part, psychoperceptual evaluation and sensory profiling will take 90–120 minutes, respectively. The length of each part forces the researcher to conduct OPQ in two or three sessions.

3.2. Psychoperceptual Evaluation

3.2.1. Research Problem. The goal of psychoperceptual evaluation is to assess the degree of excellence of the perceived overall quality for multimedia.

3.2.2. Data Collection. Psychoperceptual evaluation is based on the standardized quantitative methodological recommendations [8, 33]. The selection of the appropriate method needs to be based on the goal of the study and the perceptual differences between stimuli. Their provided guidelines to design and conduct the experiments and the quantitative data analysis (for a review see [37]) are recommended to follow. The overall quality of the stimuli is assessed by test participants in the three following ways. (1) It can be used to evaluate heterogeneous stimuli material (e.g., multimedia quality) to build up the global or holistic judgment of quality [1]. This is controversial to the assessment of a certain quality attribute, such as brightness. (2) It assumes that both stimuli-driven sensorial processing and high-level cognitive processing including knowledge, expectations, emotions, and attitudes are integrated into the final quality perception of stimuli [1, 16, 64]. (3) It is a suitable task for consumer- or user-oriented studies in product development conducted with naïve participants [64]. In addition, overall quality evaluations can be complemented with other simple evaluations. Especially for the consumer-oriented studies, the evaluation of an acceptable quality level as an indicator of a minimum useful quality level can be appropriate for quality judgments for novel multimedia services [65].

The test procedure during the data collection contains training and anchoring and the evaluation task. In training and anchoring, participants familiarize themselves with the presented qualities and contents used in the experiment as well as with the data elicitation method in the evaluation task. Often a subset of the actual test set is used, representing the full range of quality in the study. In the following evaluation task the full set of test stimuli is presented according to the selected research method. The stimuli can be evaluated several times.

3.2.3. Method of Analysis. The quantitative data can be analyzed using the Analysis of Variance (ANOVA) or its comparable nonparametric methods if the presumptions of ANOVA are not fulfilled [43].

3.2.4. Results. Fulfilling the first goal of OPQ, psychoperceptual evaluation results in a preference ranking of the excellence of all test stimuli. These results can be translated into preferences of treatments or test parameters under evaluation, respectively.

3.3. Sensory Profiling

3.3.1. Research Problem. The goal of the sensory profiling is to understand the characteristics of quality perception by collecting individual quality attributes.

3.3.2. Data Collection. In sensory profiling, research methods are used to "evoke, measure, analyze, and interpret people's reaction to products based on the senses" [3]. In OPQ, we partly follow the method of Free-Choice Profiling (FCP), originally introduced by Williams and Langron in 1984 [66]. It allows naïve participants to use their own vocabulary, differing sensitivities and idiosyncrasies to describe the characteristics of products in a multistep evaluation procedure [3, 66]. FCP is free of time-consuming panel training but produces comparable results with other methods of descriptive analysis [3, 47, 67, 68]. Furthermore, it is well established in food sciences, acting as a good reference to the multimodal quality evaluation in the other research fields [47, 69].

The test procedure contains four subtasks called (1) introduction, (2) attribute elicitation, (3) attribute refinement, and (4) sensory evaluation task.

(1) Introduction. It aims at training participants to explicitly describe quality with their own quality attributes. These quality attributes are descriptors (preferably adjectives) for the characteristics of the stimuli in terms of perceived sensory quality [3]. The introduction helps participants to understand the nature of the descriptive evaluation task. The descriptive skills of test participants will limit the attribute elicitation [70]. The ability to express quality is an important requirement for the participants to produce strong quality attributes. In training, we start with a small task to describe something familiar to participants, such as apples. "Imagine a basket full of apples. What kind of attributes, properties, or factors can you use to describe similarities and differences of two randomly picked." Thereby, the researcher may help the test participant to find attributes, but he never comes up with suggestions. After the introductory task, participants start to describe the audiovisual quality following the idea presented.

(2) Attribute Elicitation. It aims at identifying individual quality attributes that characterize the participants' quality perception of the different test stimuli. The actual extraction of attributes can be done using different elicitation methods available. In the original Free-Choice Profiling, assessors write down their attributes without limitations [66]. However, it has been reported that it was a hard task for participants to develop their vocabulary and, therefore, supporting elicitation techniques can be applied [71]. In the Repertory Grid Technique, as one of the supporting technique [72] test participants develop attributes in triad stimuli presentations. Attributes are developed as distinguishing factors from two stimuli to the third of the triad. In the second technique, Natural Grouping [73], stimuli are divided into two groups differing in one attribute. Each new group can then be divided again by the use of a second attribute and so on. We have applied the supporting task free technique in our case studies as no additional benefit in term of the attributes' quality has been found for the supporting tasks [71, 74]. Independent of the used elicitation method, stimuli can be replicated several times and people need enough time to watch them, and iteratively develop their attributes, as we have learned over our studies. Overall, the attribute elicitation is a very important step for successful sensory profiling, as only the attributes found in this phase will be taken into account in the later evaluation.

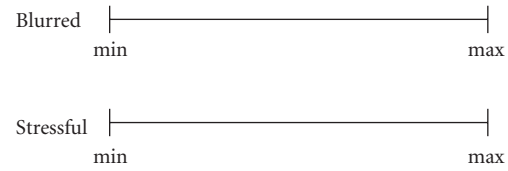


Figure 2: Examples of quality attributes with a related scale in a participant's individual score card.

(3) Attribute Refinement. It aims at separating strong attributes from all developed attributes. In FCP, participants may develop unnecessarily many attributes in their elicitation step whereas strong attributes are needed for accurate profiling. We apply two rules to describe a strong attribute. Firstly, the participants must be able to define the attribute in their own words, that is, they must know very precisely which aspect of quality is covered by the attribute. This is important for the interpretation of the results to understand the individual attributes [3]. Secondly, the attribute must be unique or nonredundant [3]. Each attribute must describe one aspect of quality. Following these rules, test participants are allowed to modify their list of attributes. It has been shown useful to also limit the maximum number of attributes. A larger set of attributes can add larger error than additional information to the sensory data [74]. However, this should be checked in a pilot study. At the end of the refinement, test participants write down a definition of each of the attributes left over for the final evaluation. The attributes are attached with a 10 cm long scale labeled with "min." and "max." in its extremes (see Figure 2). It results in an individual score card which the test participants will use for stimuli evaluation.

(4) Sensory Evaluation Task. It aims at qmid-suantifying the strength of developed attributes per stimuli. Stimuli are presented one by one and the assessment for each attribute is marked on a line with the "min." and "max." in its extremes. "min." means that the attribute is not perceived at all while "max." refers to its maximum sensation.

3.3.3. Method of Analysis. By measuring the distance from the beginning of the 10 cm long line to the mark for the rated intensity, the sensory sensation is transformed into quantitative values. Each test participant produces one configuration, that is, M × N-matrix with M rows = "number of test items" and N columns = "number of individual attributes". To be able to analyze these configurations, they must be matched according to a common basis, a consensus configuration. Generalized Procrustes Analysis (GPA) has been introduced by Gower in 1975 [46]. It rotates and translates all configurations by minimizing the residual distance between the configurations and their consensus

[3, 75]. Kunert and Qannari [76] present an alternative approach to analyze sensory profiling data, claiming this approach to be more applicable for FCP data analysis. The problem of scaling the individual configurations is solved in a way that "all the configurations (are put) on the footing as the sums of squares become equal for all the data sets". The scaled data sets from GPA or Kunert and Quannari's approach [76] can be analyzed using Principal Component Analysis (PCA).

3.3.4. Results. GPA and the alternative Kunert and Quannari approach both create a low-dimensional model of the high-dimensional input matrix. As a value of excellence of the model, the explained variance is the amount of variance of the high-dimensional space that is represented by the model. The results are finally plotted as word charts (correlation plots) showing correlation of the individual attributes with the principle components of the low-dimensional model. In contrast to interview-based evaluation methods [16, 17], no personal data interpretation has been introduced in the analysis. At this stage, the researcher will start to identify the principal components of the perceptual space, the GPA scores of the items and attributes' correlation with the components to understand the rationale behind the model. This fulfills the second goal of the OPQ method.

3.4. External Preference Mapping

3.4.1. Research Problem. The goal of the External Preference Mapping (EPM) is to combine quantitative excellence and sensory profiling data to construct a link between preferences and quality construct.

3.4.2. Research Method. In general, External Preference Mapping maps the participants' preference data into the perceptual space and so enables the understanding of perceptual preferences by sensory explanations [62, 77]. EPM is carried out using methods of multiple polynomial regressions, for example, Partial Least Square Regression [78] or PREFMAP [77].

To show how OPQ can be applied in multimedia quality research, we present three experiments in the field of audiovisual 3D quality. The first experiment explores experienced audiovisual quality when room acoustic audio reproduction and visual presentation mode (2D/3D) on a midsized screen are varied. In the second experiment, experienced audiovisual quality is examined under different audio (mono/stereo) and visual (2D/3D) presentation modes on a small mobile screen size. The third experiment investigates the influence of different 3D video coding methods on experienced quality on small screens. In all experiments, constructed quality level can be considered as moderate, containing perceivable impairments in their presentation.

4. Experiment 1: Experienced Quality of Audiovisual Depth

The goal of the first experiment is to explore the influence of audiovisual depth on perceived quality. In the previous work, bimodal depth experiences are studied for virtual reality systems with large screen sizes and very high-quality multichannel audio, or only one modality is explored at the time [79–82]. In this study, we investigate multimodal quality perception with mixed methods when the depth is varied in both modalities. Our independent variables are mono- and stereoscopic visualizations on midsized screen and audio-related room acoustic simulations for small and large spaces with multichannel loudspeaker reproduction.

(a) (b)

Figure 3: Snapshots of the content used in the study. (a) The virtual classroom with manikin as avatar. (b) The virtual living room with laptop as avatar.

4.1. Research Method

4.1.1. Test Participants. A number of 25 naïve assessors took part in psychoperceptual quality evaluation task (gender: 9 females, 16 males; age: 18–27 years) [1, 57, 58]. Sensory profiling was conducted with a subsample of 19 participants. All participants had normal or corrected-to-normal visual acuity and normal audio acuity.

4.1.2. Stimuli. We varied depth in visual presentation mode (2D/3D) and room acoustic simulations (small/large room) in audio. Two different audiovisual contents, rendered from different sized virtual rooms, were used. Visually, a sharp display offers the possibility to physically switch between 2D and 3D presentation of the content. For the audio part, the IAVAS player offers functions to render different room acoustics [83].

In a large room, visualized as a classroom, an audio is presented by a male speaker and the sound source by a manikin (see Figure 3(a)). In a small room, visualized as a student's living room, the audio plays drum and bass music and the sound source is represented by a laptop (see, Figure 3(b)). The users' movement through the room is automated containing movement straight on and turning right or left. In total, eight 15-second long stimuli were used in the experiment.

The rooms were designed using Maya software. For playback in the IAVAS I3D player [83], the scenes were exported into Binary Format for Scenes (BIFS). The audio was included using Advanced Audio BIFS. The audio files were encoded with AAC at a bit rate of 128 kbps. The room

acoustics were modeled using the perceptual approach that is provided by the player. For each room a suitable room acoustic was modeled taking into account the different sizes and acoustical characteristics of the rooms. To vary depth in audio perception, the room models were exchanged between the rooms.

4.1.3. Stimuli Presentation. The tests were conducted in the Listening Lab at Ilmenau University of Technology, set according to [8, 84]. The videos were presented on a "15" Sharp AL3DU stereoscopic display based on parallax barrier technology. The parallax barrier is built as a secondary LCD layer which can be switched on and off so that the screen can be used for monoscopic and stereoscopic videos. The viewing distance was 55 cm. The sound was played back on a four-speaker surround setup at "30" and "110" and a distance of 1 meter from the assessor [79]. The stimuli were repeated twice in random order for psychoperceptual evaluation.

4.1.4. Test Procedure. The test procedure is described according to the theoretical method description in Section 3.

Psychoperceptual Evaluation. Prior to the actual evaluation, training and anchoring took place. Participants trained for viewing the scenes (i.e., finding a sweet spot) and the evaluation task were shown all contents and the range of constructed quality, including four stimuli. Absolute Category Rating was applied for the psychoperceptual evaluation for the overall quality, rated with an unlabeled 11-point scale [33]. In addition, the acceptance of overall quality was rated on a binary (yes/no) scale [39]. All stimuli were presented twice in a random order. The simulator sickness questionnaire (SSQ) was filled out prior to and after the psychoperceptual evaluation [85, 86].

Sensory Profiling. The Sensory Profiling task was based on a Free-Choice Profiling [47] methodology. The procedure contained four parts and they were carried out in two sessions within three days. (1) An introduction to the task was carried out using the imaginary apple description task. (2) Attribute elicitation—all stimuli were presented three times, one by one. The participants were asked to write down their individual attributes on a white sheet of paper. They were not limited in the amount of attributes nor were they given any limitations to describe sensations. (3) Attribute refinement—the participants were given a task to rethink (add, remove, change) their attributes to define their final list of words. It was transformed into the assessor's individual score card. Finally, four randomly chosen stimuli were presented once and the assessor practiced the evaluation using a score card. In contrast to the following evaluation task, all ratings were done on a one score card. Thus, the test participants were able to compare different intensities of their attributes. (4) Evaluation task—the stimulus was presented three times in a row, and the participants rated it on a score card. If necessary, they were allowed to ask for a fourth repetition.

4.1.5. Methods of Analysis

Psychoperceptual Evaluation. Nonparametric methods of analysis were used (Kolmogorov-Smirnov: $P < .05$). Friedman's test is applicable to the measurement of differences between several ordinal dependent variables and Wilcoxon's test in their pairwise comparisons [43].

Sensory Profiling. The sensory data has been analyzed using Microsoft excel and the GPA routine of XLSTAT 2.9.0. The data was also analyzed using Kunert and Qannari's method [76]. As the GPA produced stronger results in terms of explained variance of the model, the GPA model will be used for further analysis.

4.2. Results

4.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality. All presented stimuli provided a highly acceptable quality level, reaching an acceptance level of 83% at the minimum. The test parameters did not have an impact on the acceptance of overall quality (Cochran's $Q = 0.79$, $P > .05$, ns). All items were rated equally (McNemar: all comparisons $P > .05$).

Overall Quality Satisfaction. Visual presentation modes and room acoustic simulations did not have significant influence on the overall quality satisfaction (Friedman, $\chi^2 = 3.341$, $df = 7$, $P > .05$, ns). All stimuli were equally rated (all pairwise comparisons $P > .05$, ns).

4.2.2. Sensory Profiling. The test participants developed a set of 289 attributes. A total of 216 of the attributes represented between 50% and 100% of the explained variance. These attributes are located between the inner and the outer circle in the correlation plots (see Figures 6, 7, and 8). The assessors used 10 attributes (min. 7, max. 32) on average to describe their sensory perception in the dimensionally reduced data.

Identification of Dimensions and Attributes. Seven components were needed to explain 100% variance in the GPA model. The contribution of each component is described in Table 3. Considering the elbow criteria and the Heyman and Lawless rule of interpretability [3], the first three components are used for further data interpretation. The GPA result with the three principal components forms the GPA model or perceptual space.

To understand the perceptual space, the attributes and test stimuli are plotted into the model, resulting in a three-dimensional space. For better interpretation, component 2 and component 3 are always plotted against component 1 to get two-dimensional slices of the perceptual space shown in Figures 4 and 5. The item names are substituted by the corresponding variables. Comparing variables and the separation of the items in the perceptual space allows for determining the components. Figure 4 shows that Dimension number 1 (PC1) relates to content (classroom or

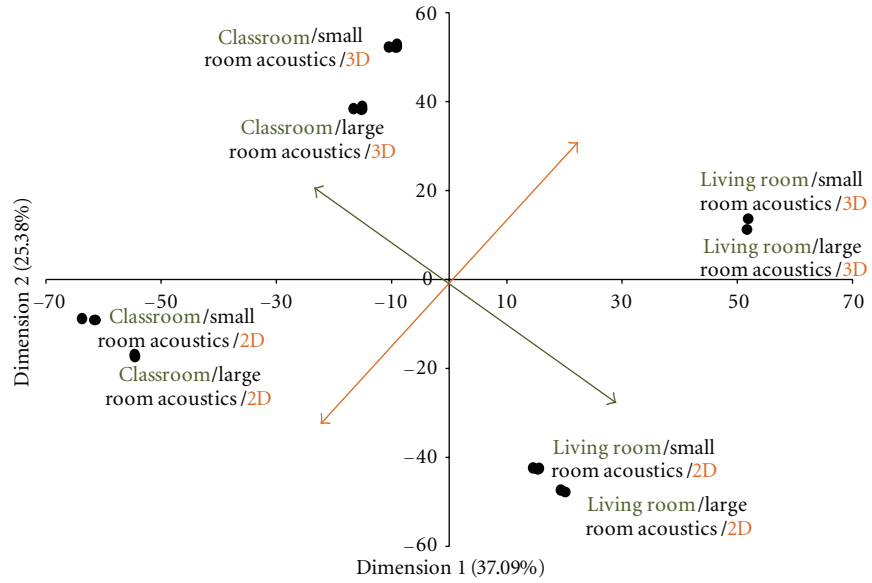


Figure 4: PC1-PC2 slice of the model showing test items plotted into the GPA model. The brown and green arrows clarify the dimensions of content (PC1) and video representation (PC2).

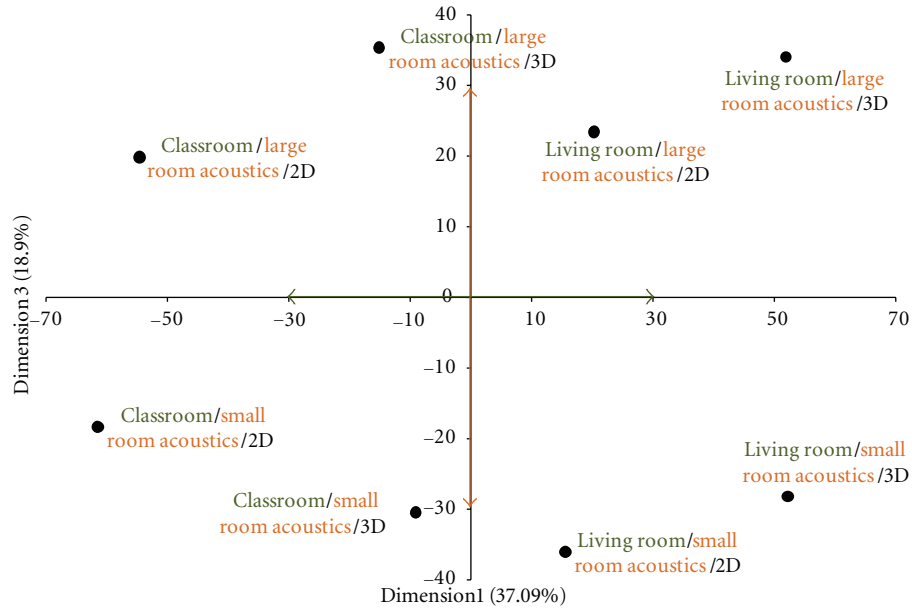


Figure 5: PC1–PC3 slice of the model showing test items plotted into the GPA model. The brown and green arrows clarify the dimensions of content (PC1) and room acoustics (PC3).

Table 3: Seven principal components (also dimensions or factors) of the Generalized Procrustes Analysis, their Eigenvalues, and the percentage of explained variance of the GPA model.

GPA model components	Eigenvalue	Explained variance (%)	Cumulative explained variance (%)
PC1	1646.8	37.09	37.09
PC2	1126.68	25.37	62.46
PC3	839.42	18.9	81.36
PC4	248.9	5.61	86.97
PC5	244.33	5.5	92.47
PC6	205.21	4.62	97.09
PC7	129.15	2.91	100.00

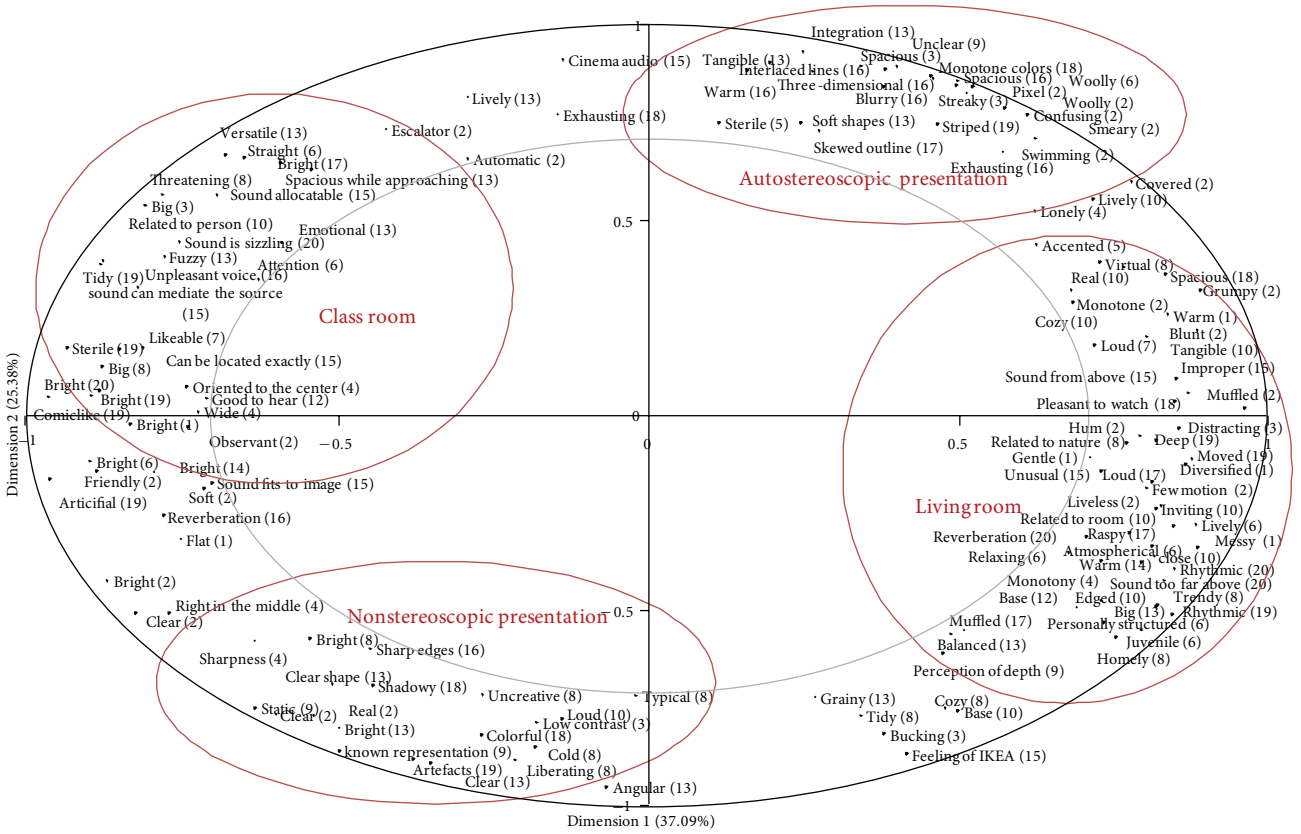


Figure 6: PCA correlation loadings with attributes in the space of PC1 and PC2.

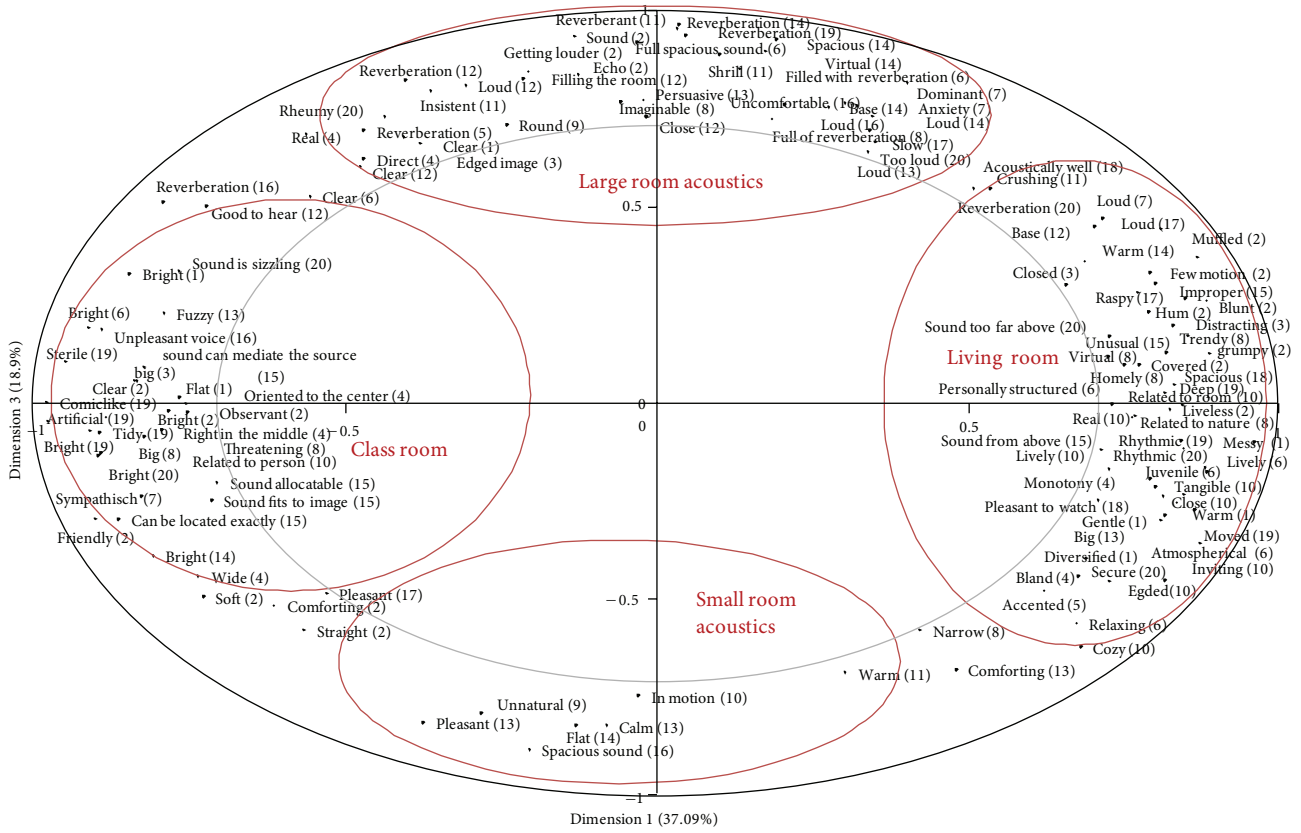


Figure 7: PCA correlation loadings with attributes in the space of PC1 and PC3.

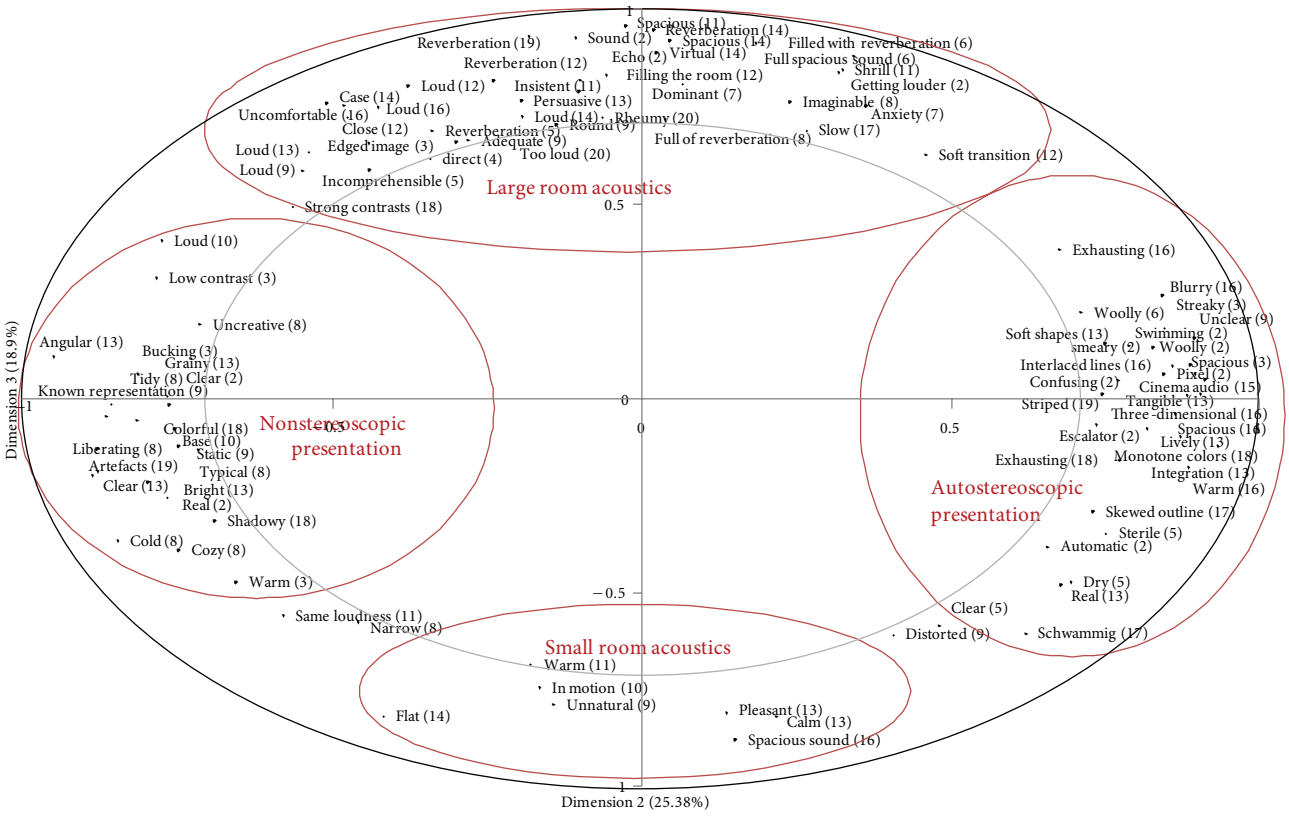


Figure 8: PCA correlation loadings with attributes in the space of PC2 and PC3.

student's room). Dimension number 2 (PC2) separates the test items according to the visual Presentation Mode (2D or 3D presentation). PC2 is identified as "video quality." Dimension number 3 (PC3) divides the items by the room acoustics (simulated small room and simulated large room). It relates to the "audio quality" of the stimuli. Although the interpretation was done based on the test items or their related test parameters, we will refer to the quality aspects of content, video representation, and room acoustics in further interpretation. This first finding confirms that test participants derived their individual quality factors from the chosen test parameters.

The attributes can be classified into two different groups. Technical descriptions directly describe the characteristics of the test variables (like reverberation or grainy). The second group of attributes is characterized by experiences, subjective impressions, and feelings about the test items (e.g., monotone, lively, or obtrusive). This group is called impression descriptions. Following, we will discuss the correlation of the attributes and attribute groups with the GPA model.

Correlation of Attributes and the Perceptual Space. Word charts represent the correlation of the individual attributes with the perceptual space (see Figures 6–8). The closer an attribute is placed to one of the dimensions, the more it correlates with this dimension. Attributes placed between two dimensions correlate with both dimensions equally.

The Dimension "Content" (PC1, 37.09% of Explained Variance). We identify the two polarities of this dimension as classroom on the one side and student's room on the other side by interpreting the attributes. But only a few attributes such as "unpleasant voice", "comiclike" or "messy" describe the content or the layout of the room directly. PC1 is more an impression description of the content or the impression of the content on the individual perception. Descriptions as "liveless", "emotional" and "likeable", or "monotone" and "sterile" highly correlate with one of the two polarities, respectively. The high amount of impression descriptions shows that quality perception is formed on an abstract level by the test participants. The assessors were able to find individual attributes that describe quality on a general level among the test items.

The Dimension "Visual Presentation Mode" (PC2, 25.38% of Explained Variance). The polarities agree with varied visual presentation modes (mono (2D), autostereoscopic (3D)). The 2D polarity shows descriptions of "sharpness" or "sharp edges", "high contrasts", "clear", "light", or "colorful". In contrast, 3D presentation mode is described with a negative description of the visual artifacts, such as "skewed outline", "unclear", or "interlaced lines". It seems that the artifacts and reduced brightness of 3D results from limitations of the display technique (parallax barrier and viewing angle of the display). However, the results also show the participants' ability to experience visual depth. It is

described as "integration", "three-dimensional", "spacious", or "tangible".

The Dimension "Room Acoustic Model" (PC3, 18.9% of Explained Variance). PC3 also corresponds directly to the varying room acoustic models used in the test. The dimension can be divided at the extreme values in the large room and the small room. While the small room acoustics are described poorly, a lot of quality factors can be found for the large room acoustics. In this dimension, technical descriptions are dominating. The large room correlates with a high amount of reverberation, "full spacious sound", and "filling the room". On the level of impression descriptions, PC3 is characterized by "imaginable", "insistent", or "shrill".

Interdimensional Attributes. Attributes that correlate with more than one dimension can be interesting. Especially attributes that correlate with PC2 and PC3, as they describe audiovisual effects. Interdimensional attributes between audio and video dimension are rare (see Table 8). Especially depth-related attributes that we expected to correlate with both dimensions correlate either with the video (e.g., spacious (P3)) or with the audio dimension (e.g., spacious (P14)). These results show that depth was perceived or rated independently either in auditory or visual perception. So, in the next section we will have a closer look at the participants' individual perceptual patterns.

Comparison of the Individual Configurations and the GPA Model. The results show individual differences in the perceptual space. By plotting the assessor's attributes into the perceptual space independently, we identified sensorial preferences between participants. As an example, the word charts illustrates that audiophile assessors (e.g., P14) mainly pay attention to auditory stimuli, while videophile assessors (e.g., P13) emphasize the visual part of stimuli. Just a few assessors (e.g., P25) used the whole perceptual space for characterizing the stimuli with their attributes. These results show that multimodal quality evaluation is also influenced by the participant's sensorial preferences.

4.2.3. External Preference Mapping. The external preference mapping was not applied, as the results of psychoperceptual evaluations did not show any preferences between stimuli.

4.3. Discussion and Conclusions. Our results of psychoperceptual quality evaluation did not show the influence of audiovisual depth on perceived quality. However, the results of sensory profiling gave further understanding of this. Firstly, the nonsignificant difference was not caused by the nondetectable differences between stimuli, as the participants qualitatively differentiated them. Secondly, the perceived depth was highlighted by both modalities contributing to the overall audiovisual perception. Thirdly, when visual 3D presentation mode was used, it was described as spacious and three-dimensional, but more importantly it was attached to several negative terms of inferiority. It is known that the added value induced by the visual depth perception is only acknowledged if the level of visible artifacts is low enough [87–89].

Our results also showed individual preferences towards the quality of one modality. It is known that there are modality-dependent individual differences in human information processing styles. For example, the categorization into visual and verbal information processing styles is common [90]. Our results indicate that these different processing styles can also contribute to final multimodal quality judgments. There are two suggestions for further work. Firstly, the influence of different processing styles on multimodal quality perception under different quality levels and heterogeneous stimulus material needs to be addressed in detail to confirm the phenomenon. Secondly, for the practitioners of audiovisual quality, a well-validated tool is needed for identifying the groups of different information processing styles and reporting these groups to characterize the sample.

5. Experiment 2: Experienced Quality of Audiovisual Depth in Mobile 3D Television and Video

We examined the influences of mono and stereo audio and visual presentation modes on experienced quality for mobile 3D television and video. Visual stereoscopic 3D experience is a multidimensional construct of video quality, depth perception, and visual comfort [88]. Previous work has shown that visual 3D has added value of depth and it can be instrumented in the evaluations of depth perception [89]. In the overall quality evaluations of impaired 3D images, the artifacts dominate over the benefits of depth [91]. Furthermore, viewing comfort is a part of the experienced quality on stereoscopic displays, and on small displays it has a coinfluence on the viewer [92–94]. To date, there are only a limited number of studies published comparing the subjective visual quality between different presentation modes (2D and 3D) on the mobile screen sizes [49, 54, 95]. Although these studies underline the critical aspects of visual 3D, they do not pay attention to overall multimodal where audio can also contribute. Previous studies have shown interaction between audio and video quality on studies for mobile television [41, 96]. The previous studies have not addressed experienced quality when depth has been varied in both audio and visual modalities.

5.1. Research Method

5.1.1. Test Participants. A total of 45 test participants (gender: 13 females, 32 males; age: 15–30, mean = 24 years) took part in the psychoperceptual evaluation task. For sensory profiling, a subsample of 15 participants was randomly selected. All test participants passed a screening for visual acuity, color, and 3D vision and hearing acuity. The majority of the participants were categorized as naïve assessors (87%) [1].

5.1.2. Stimuli

Variables and Their Production. Targeting different depth perception in auditory and visual channels, the videos were varied in video (monoscopic or stereoscopic) and audio (mono or stereo), resulting in 24 videos under test.

The original audio tracks from all videos were exported as mono and stereo tracks from Adobe Premiere in the required length. Audio was normalized. The original videos were resized to a resolution of 856 px × 240 px and exported as stereoscopic videos. To create the monoscopic videos, two original videos were imported into Shake and resized to 856 px × 240 px. The right video was cropped from both videos, resulting in two left videos, each with a resolution of 428 px × 240 px. One of the cropped videos was shifted to the right side. Both videos were added, resulting in two left-side videos next to each other with a resolution of 856 px × 240 px. Finally, the monoscopic videos were exported with mono and stereo audio tracks, respectively. All videos were coded with mp4v codec using Simulcast at 25 fps with high bitrates of at least 10 Mbit/s for the video track.

Contents. Six different contents were used to create the stimuli under test (Table 4). The selection criteria for the videos were spatial details, temporal resolution, amount of depth, and the user requirements for mobile 3D television and video [97].

5.1.3. Stimuli Presentation. The controlled laboratory conditions were similar to experiment 1 [8]. An NEC "autostereoscopic 3.5" display with a resolution of 428 px × 240 px was used to present the videos. This prototype of a mobile 3D display provides equal resolution for monoscopic and autostereoscopic presentation. It is based on lenticular sheet technology [98]. The viewing distance was set to 40 cm. The display was connected to a Dell XPS 1330 laptop via DVI. AKG K-450 headphones were connected to the laptop for audio representation. The laptop served as a playback device and control monitor during the study. The stimuli were presented in a counterbalanced order in both evaluation tasks. All items were repeated once in the psychoperceptual evaluation task. In the sensory evaluation task, stimuli were repeated only when the participant wanted to see the video again.

5.1.4. Test Procedure. A two-part data-collection procedure follows the theoretical method description in Section 3.

Psychoperceptual Evaluation. The procedure was identical to experiment 1. To capture the positive aspects of autostereoscopic presentation, participants also rated for perceived depth on an 11-point unlabeled scale.

Sensory Profiling. A four-part sensory profiling task contained: (1) an introduction to the task—identical to experiment 1; (2) attribute elicitation—participants watched 15 randomly selected items in groups of three items. Triad presentation (as used in the Repertory Grid method (RGM) [99]) was chosen to help the participants with the attribute elicitation through a comparison of different items in unlimited time. The number of generated attributes was not limited per triad. In the end, the participants were given a chance to review and revise their attributes, (3) attribute refinement—the aim of this task was to revise (remove, add, or redefine) all attributes. A number of 15 randomly selected items were presented in triads, and the participants rated each using their score cards. In each triad, the three items were presented one after another without a break and rated on the same scoring card. Triad presentation was chosen to help the participants to compare the items during the rating process. Each triad was repeated once if the participant so needed. At the end of this task, the participants defined their final attributes, (4) evaluation task—in the final evaluation task, all 24 items were rated independently with all attributes. Each item was presented once and the rating time was not limited.

The study was conducted in two sessions of approximately 90 minutes. Psychoperceptual evaluation and the subtasks 1-2 of sensory profiling took place in the first session and the rest in the second session.

Table 4: Snapshots of the six contents under assessment (V_{SD} = visual spatial details, V_{TD}: temporal motion, V_D: amount of depth, V_{DD}: depth dynamism, V_{SC}: amount of scene cuts, and A: audio characteristics).

Screenshot	Genre and their audiovisual characteristics
	*Animation—Knight's Quest 4D* (18 s) V_{SD}: high, V_{TD}: high, V_D: med, V_{DD}: high, V_{SD}: high, A: music, effects
	*Documentary—Cave* (18 s) V_{SD}: high, V_{TD}: med, V_D: high, V_{DD}: low, V_{SC}: low, A: orchestral music
	*Videoconference—Bullinger* (23 s) V_{SD}: med, V_{TD}: low, V_D: med, V_{DD}: low, V_{SC}: low, A: male voice
	*User-created Content—Oldtimers* (16 s) V_{SD}: high, V_{TD}: high, V_D: high, V_{DD}: med, V_{SC}: low, A: train sound
	*Music Video—Mouldpenny* (19 s) V_{SD}: med, V_{TD}: med, V_D: med, V_{DD}: low, V_{SC}: low, A: music Video bitrate: 13 Mbit/s
	*Documentary—Upper Rhine Valley* (18 s) V_{SD}: high, V_{TD}: med, V_D: high, V_{DD}: high, V_{SC}: med, A: ambient music

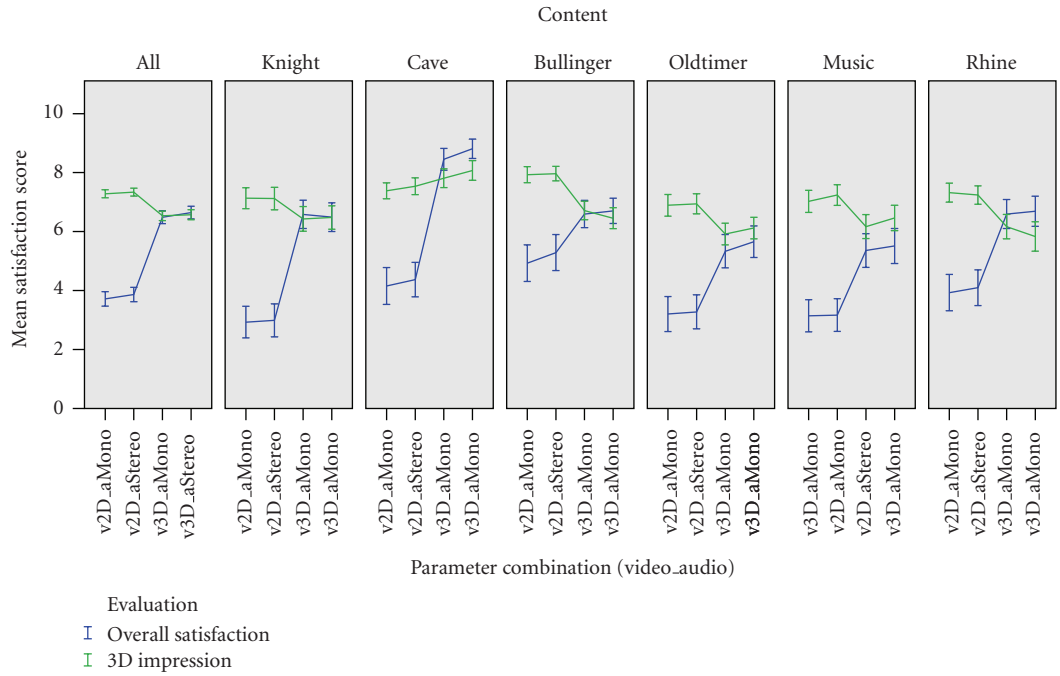


Figure 9: Results for overall quality and 3D impression. The bars show 95% CI of mean.

5.1.5. Method of Analysis

Psychoperceptual Evaluation and Sensory Analysis. The methods were identical to experiment 1.

External Preference Mapping. External Preference Mapping was applied to map the users' preferences into the perceptual space. Two models can be used to describe the participants' preferences: the vector model and the ideal point model [77]. Within the PREFMAP method in XLSTAT, the most suitable model is chosen automatically.

5.2. Results

5.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality. Overall, all presented stimuli provided a highly acceptable quality level. On average, 2D presentation mode reached the acceptance level of 90% and all stimuli reached at least an acceptance of 88%. For 3D visual presentation mode, the average acceptance level of quality was 79%, while none of the stimuli went below 63% of acceptance.

Overall Quality Satisfaction. Parameter combinations influenced overall quality satisfaction when averaged over the content (Fr = 92.2, $df = 3$, $P < .001$). Figure 9 shows the overall quality scores averaged over the content and content by content for the 4 parameter combinations (v2D_aMono, v2D_aStereo, v3D_aMono, and v3D_aStereo). The 2D video presentation mode provided the most satisfying quality compared to the 3D video mode ($P < .001$). The audio presentation mode had no effect on the quality ratings. Mono audio and stereo audio were equally evaluated in both video presentation modes ($P > .05$, ns). The results of content by content analysis follow this main tendency with content cave as an exception. Although there is no overall effect of parameter combinations on satisfaction (Fr = 4.46, $df = 3$, $P = .215$, ns) in this content, detailed pairwise comparisons show that the 3D presentation mode provides higher quality under equal audio conditions (3D versus 2D—mono: $Z = -2.53$, $P < .001$; 3D versus 2D—stereo: $Z = -3.12$, $P < .001$). However, 2D accompanied with stereo audio reaches the quality level equal to 3D with mono audio presentation ($Z = -1.61$, $P = .108$, ns).

3D Impression. The parameter combinations influenced the perceived depth when averaged over the content (Fr = 596.4, $df = 3$, $P < .001$). Figure 9 shows the mean values for 3D impression and overall quality for all content and separately for all six contents. The highest level of depth perception was provided by stimuli in 3D presentation mode ($P < .001$). Under the 3D mode, the used audio presentation mode did not influence depth perception ($Z = -1.45$, $P = .14$, ns), while stereo mode slightly outperformed mono when 2D video mode was used ($Z = -2.91$, $P < .01$).

5.2.2. Sensory Profiling. A number of 15 participants developed 130 individual quality attributes in the OPQ task. For further research, the attributes from one participant were eliminated because this participant had already been eliminated from the quantitative analysis due to outliers. Finally, 116 attributes from 14 participants remained (mean = 8.3, min. = 3, max. = 14).

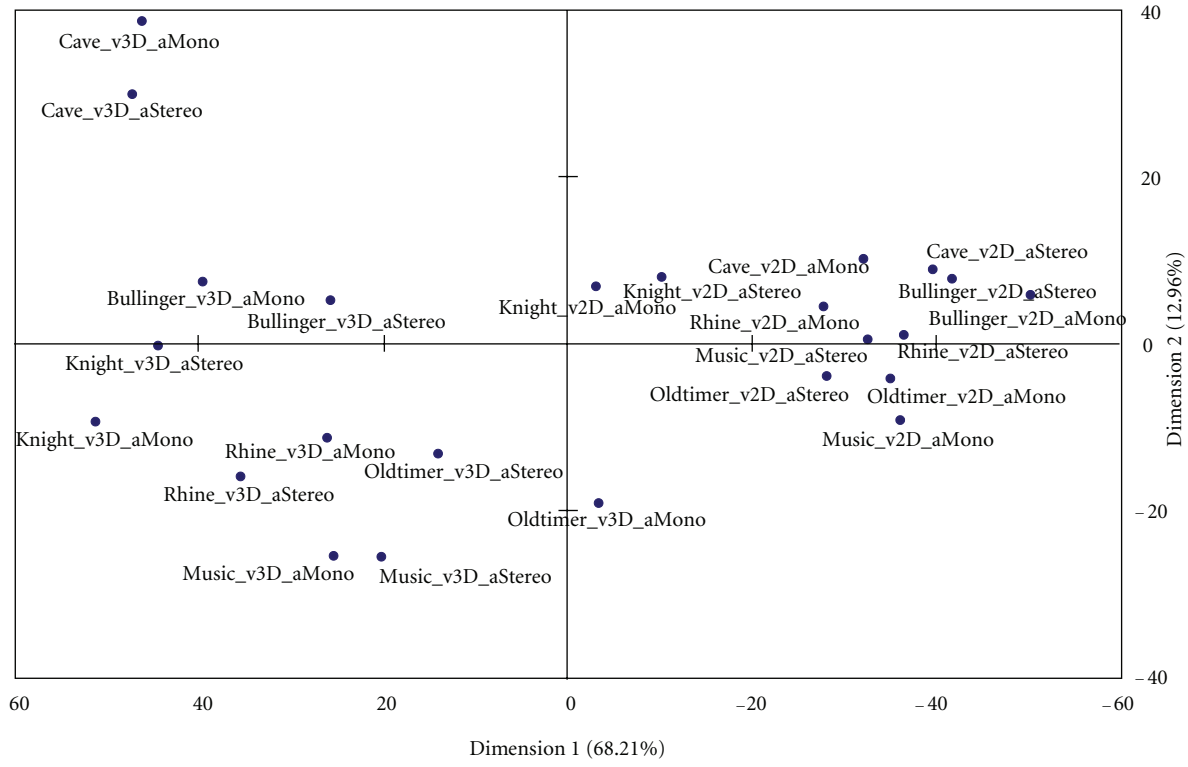


Figure 10: PCA scores of objects in the space of PC1 and PC2.

Table 5: The 14 components of the GPA, their Eigenvalues, and the percentage of explained variance of the GPA model.

GPA model components	Eigenvalue	Explained variance (%)	Cumulative variance (%)
PC1	1157.99	68.21	68.21
PC2	219.97	12.96	81.17
PC3	113.35	6.7	87.84
PC4	66.18	3.9	91.7
PC5	54.8	3.2	94.97
PC6	36.6	2.16	97.13
PC7	18.23	1.1	98.2
PC8	10.17	0.6	98.8
PC9	8.3	0.49	99.29
PC10	5.23	0.31	99.6
PC11	2.83	0.17	99.76
PC12	2.5	0.15	99.91
PC13	1.36	0.08	99.99
PC14	0.15	0.009	100.00

Identification of Dimensions and Attributes. A total of 14 components were needed to explain 100% variance in the GPA model (Table 5). The first two components are used for further data interpretation according to the elbow criteria and the rule of interpretability [3]. These two components form the perceptual space.

Figure 10 shows the test stimuli in the perceptual space of PC1 and PC2. Correspondingly, Figure 11 shows the attributes in the space of PC1 and PC2. Attributes with an explained variance between 100% and 50% are emphasized and considered for further interpretations. As can be seen in both plots (Figures 10 and 11) PC1 divides the items by video presentation mode (2D and 3D). PC2 relates to positive and negative descriptions of overall quality. Interestingly, participants concentrated on the video quality description and their impressions and did not find attributes related to the audio presentation mode.

Dimension 1 ("visual presentation mode," 68.21% explained variance). The polarities agree with varied visual presentation mode (monoscopic (2D) and stereoscopic (3D)). Monoscopic videos are described by attributes like "normal," "natural," "flat," "sharp," and "focusable". In contrast, attributes like "3D benefit," "depth impression," "3D feeling," "three-dimensional," "tangible," or "sharp" describe videos in stereoscopic presentation mode. The assessors were able to distinguish between 2D and 3D video presentation mode and described the quality on a general level.

Dimension 2 ("impression descriptions," 12.96% explained variance). It divides the perceptual space into negative and positive impression descriptions. Videos in monoscopic presentation mode are described only by positive descriptions like "exciting," "pleasant," "beautiful," "focusable image," or "stress-free." The participants described the stereoscopic videos with positive and negative attributes. On the one

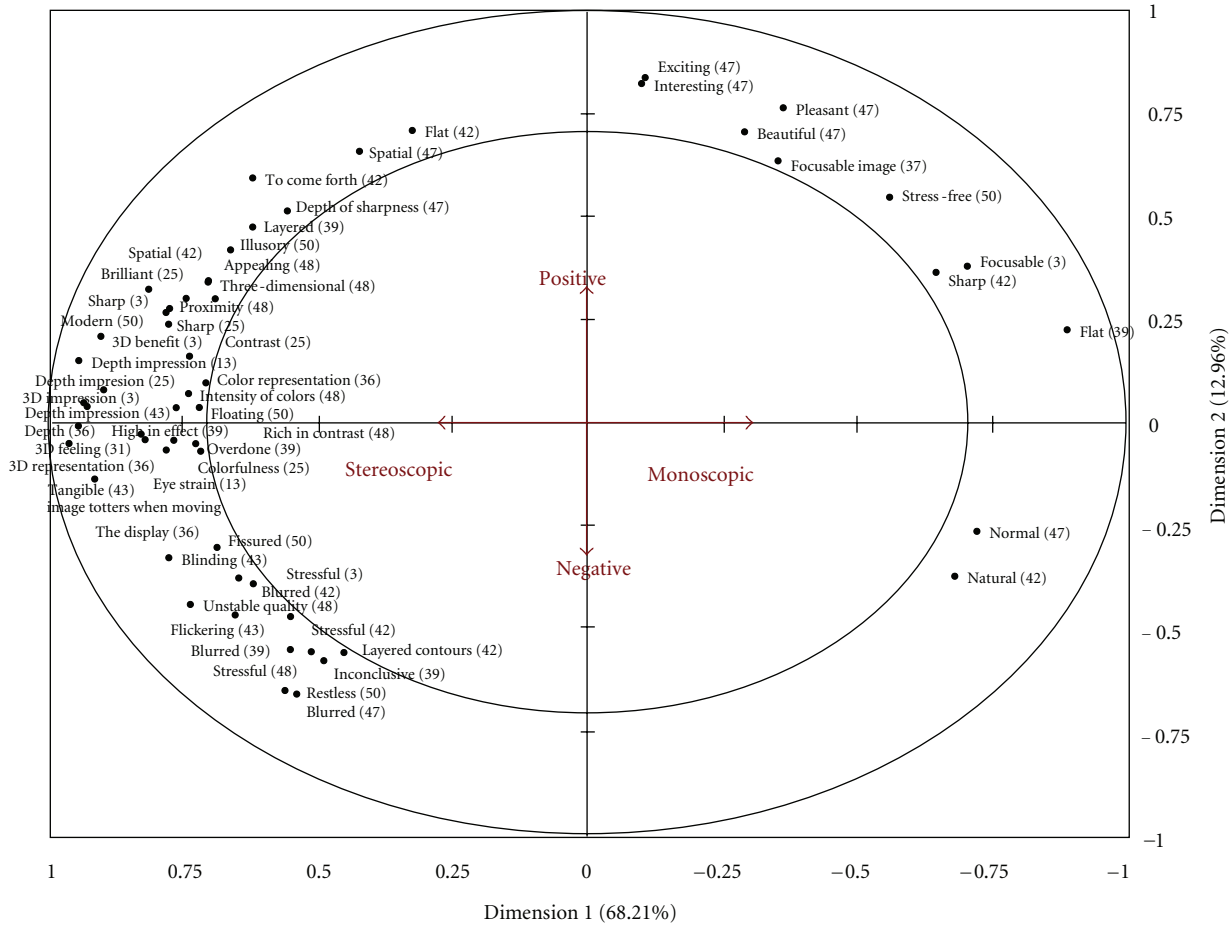


Figure 11: PCA correlation loadings with attributes in the space of PC1 and PC2.

hand, side negative descriptions are artifact related and concern the display technology like "image totters when moving the display", "flickering," "inconclusive," "stressful," "restless," or "blurred." On the other hand, 3D videos are described as "spatial," "brilliant," "appealing," "illusory," or "layered." The effects of content on quality perception can be seen in the participants' impression descriptions, like "interesting" or "exciting."

5.2.3. External Preference Mapping. Psychoperceptual data was combined with the sensory profiling results using external preference mapping (Figure 12). It can be seen that many participants prefer content cave in stereoscopic mode and the other contents in monoscopic video mode. Three preference clusters could be obtained from the preference map. The participants in cluster 1 prefer stereoscopic items, especially participants 10 and 25. The participants in cluster 2 prefer item cave in stereoscopic mode. Monoscopic video items, especially items knight and cave, are preferred by cluster 3. It is also possible to combine the users' preferences with the attributes from the profiling task. Therefore, both plots, the external preference map and the word plot, should be observed next to each other. The preferences from the preference map can be correlated with the attributes from the word plot. The participants who prefer items cave in 3D video mode seem to like items that can be described as "spatial," "to come forth," or "three-dimensional." Furthermore, 2D videos that are preferred are described as "interesting," "stress-free," or "pleasant." On the contrary, contents music, rhine and oldtimer correlate with attributes like "blurred," "stressful," or "inconclusive" and are disliked.

5.3. Discussion and Conclusion. Our results underlined the dominance of visual quality factors over the audio factors and their interaction in the experienced quality. This result was confirmed by three different evaluation tasks used (psychoperceptual quality satisfaction and depth impression and sensory profiling). Similar to our results, nonsignificant influences of audio on audiovisual quality have been concluded in the context of large displays and surround systems in a good quality level [100, 101]. Neuman et al. [100] have shown that naïve participants have difficulties in differentiating between mono and stereo audio under the video viewing task. Furthermore, Lessiter and Freeman [101] underlined that the feeling of presence is not enhanced by audio mode. It is also possible that the visual variable acted as the most changing variable in the experiment and captured the greatest attention as suggested by peak-end theory [102].

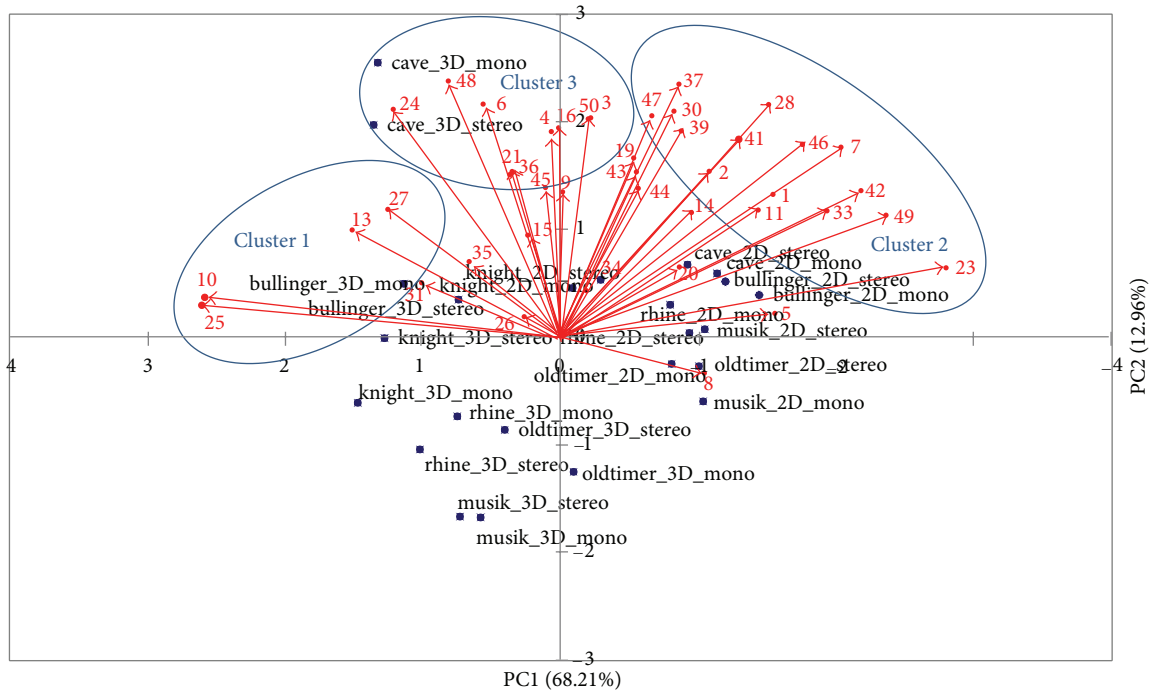


Figure 12: Objects and participants' preferences in the preference map of PC1 and PC2.

The results showed also a controversial impact of 3D presentation mode on overall quality and depth impression. While the use of 3D mode increased the depth impression, it decreased the overall satisfaction. The descriptive GPA results gave further explanations to these results by underlining the inferiority (spatial, stressfulness, flickering, eye-strain) in the case of 3D. However, our results also showed that in artifact-free cases, 3D can reach higher perceived quality compared to 2D. In that case the perceived depth and the exciting 3D sensation make the stereoscopic videos subjectively better. This result indicates that the added value induced by the depth perception in stereoscopic presentation is only valid when the level of visible artifacts is low. These findings support results from further studies [88, 89, 91].

Further work needs to address the most annoying artifacts to improve 3D presentation to the sufficient level of technical resources for portable devices.

6. Experiment 3: Experienced Quality of Video Coding Methods for Mobile 3D Television

The third case study targeted the selection of an optimum stereo video coding method for mobile 3D television and video applications. Different approaches of coding algorithms have currently been optimized for mobile 3D video [103]. No previous work evaluated these approaches in a large-scale study. Previous work on stereo video coding was mainly done on still images [65, 88, 91]. These studies showed that the added value of stereoscopic stimuli given for the uncompressed case [89] is not valid for MPEG2 or JPEG compressed material [65, 91, 104]. In these cases, the depth perception did not increase the perceived overall quality of the stimuli.

6.1. Research Method

6.1.1. Test Participants. A total of 47 naïve assessors (gender: 23 females, 24 males; age: 16–37, mean: 24) took part in the psychoperceptual evaluation task [1]. A total of 15 of them were randomly selected from this sample for the sensory profiling task. All assessors passed a screening for visual acuity, color, and 3D vision and were also among the potential users for mobile 3D television [97]. Parents' consent was required for the participation of underage assessors.

6.1.2. Stimuli

Variables and Their Production. We varied four coding methods and two quality levels in this study. Four coding methods, especially adapted for mobile stereo video [103], were chosen for evaluation. As Video + Video approaches H.264/AVC Simulcast [105], a straight-forward coding solution was chosen. As an advanced approach H.264/AVC MVC [106], Mixed Resolution Stereo Coding (MRSC) [107] was chosen. In addition, Video + Depth [108] as an alternative approach to the Video + Video coding methods was selected. As a coding profile, the Baseline profile, that is, IPPP structure and CAVLC (Context Adaptive Variable Length Coding), was used. The GOP size was set to 1. A low and a high quality level was defined for each test sequence. To guarantee comparable low and high quality for all sequences, individual bit rate points had to be determined for each sequence. For the definition of low quality for all sequences, the quantization parameters (QPs) for simulcast coding were set to 30. The resulting bit rates for each sequence are given in Table 6. These bit rates were used as target rates for the other three approaches.

Table 6: Target bit rates of the final test sequences.

Quality level	Bullinger	Butterfly	Car	Horse	Mountain	Soccer2
Low	74	143	130	160	104	159
High	160	318	378	450	367	452

Two different codecs were used for video encoding. H.264/AVC Reference Software JM 14.2 was used for the Simulcast, Mixed Resolution, and Video + Depth. MVC was performed using H.264/MVC reference Software JMVC 5.0.5. The test stimulus production for Simulcast and MVC-encoded sequences was straightforward according to the target bit rates in Table 6. To achieve these target bit rates, the quantization parameters for the left and the right were changed together. Thus, the left and the right views were of the same quality. The depth for the Video + Depth approach has been estimated from the left and the right view using a Hybrid Recursive Matching algorithm [99]. The view synthesis was performed using Merkle et al.'s algorithm [109]. For the generation of Mixed Resolution sequences, the right view was decimated by a factor of two in both the horizontal and vertical direction. For up- and down-sampling, tools provided with the JSVM reference software for Scalable Video Coding have been utilized. The applied optimization approach is described in [110]. The frame rate of all sequences was set to 15 fps.

Contents. Six different contents were chosen to create the test stimuli (Table 7) according to similar criteria to experiment 2. None of the contents contained scene cuts.

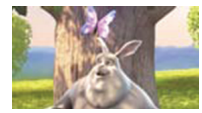
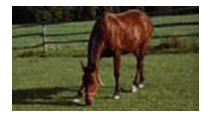
6.1.3. Stimuli Presentation. The conditions in the controlled environment were similar to experiment 1. The same setup as in experiment 2 was used, but without headphones. All items were presented twice in psychoperceptual evaluation. Each item was presented three times in a row in the sensory profiling task.

6.1.4. Test Procedure. According to the theoretical model in Section 3, the study contained two parts. A psychoperceptual evaluation and a subsequent sensory profiling were conducted.

Psychoperceptual Evaluation. The psychoperceptual evaluation followed the same method as described in experiments 1 and 2. Test participants evaluated overall quality acceptance and satisfaction with overall quality in this study. The session took about 90 minutes.

Sensory Profiling. Sensory profiling was conducted in the second session, lasting 75 minutes. A Free-Choice profiling approach was applied with the following subtasks (1) introduction to task—identical to experiments 1 and 2; (2) attribute elicitation—the test participants watched a subset of 24 randomly chosen test items. While watching, they wrote down their idiosyncratic quality attributes. No limit for the number of attributes was given in this step. During the last clips, the test participants were encouraged to review their attributes by checking if all quality aspects were covered with these; (3) attribute refinement—at the beginning of the attribute refinement, the assessors were asked to select a maximum of 15 attributes to their score card. After the selection, 12 test items were presented and the test participants evaluated these on their score cards. Still, the possibility of revising the score card (add, remove, redefine) was given. The score card was then finalized, and each assessor defined his quality attributes; (4) evaluation task—in the final evaluation task, all 48 items were rated independently. Each item was shown three times in a row to allow for enough time to apply all attributes. The rating time was not limited.

Table 7: Screenshots and characteristics of the test stimuli used in case study 3.

Screenshot	Genre and their audiovisual characteristics
	*Videoconference—Bullinger* V_{SD}: med, V_{TD}: low, V_D: med, V_{DD}: low Length: 7.7 sec Size in pixels: 432 × 240
	*Animation—Butterfly* V_{SD}: high, V_{TD}: med, V_D: med, V_{DD}: low Length: 12 sec Size in pixels: 432 × 240
	*Action/Movie—Car* V_{SD}: high, V_{TD}: high, V_D: med, V_{DD}: med Length: 7.8 sec Size in pixels: 432 × 240
	*Nature/Documentary—Horse* V_{SD}: high, V_{TD}: low, V_D: high, V_{DD}: low, Length: 9.3 sec Size in pixels: 432 × 240
	*Nature/Documentary—Mountain* V_{SD}: high, V_{TD}: low, V_D: high, V_{DD}: high Length: 8 sec Size in pixels: 320 × 240
	*Sports—Soccer2* V_{SD}: med, V_{TD}: high, V_D: high, V_{DD}: high Length: 13.3 sec Size in pixels: 320 × 240

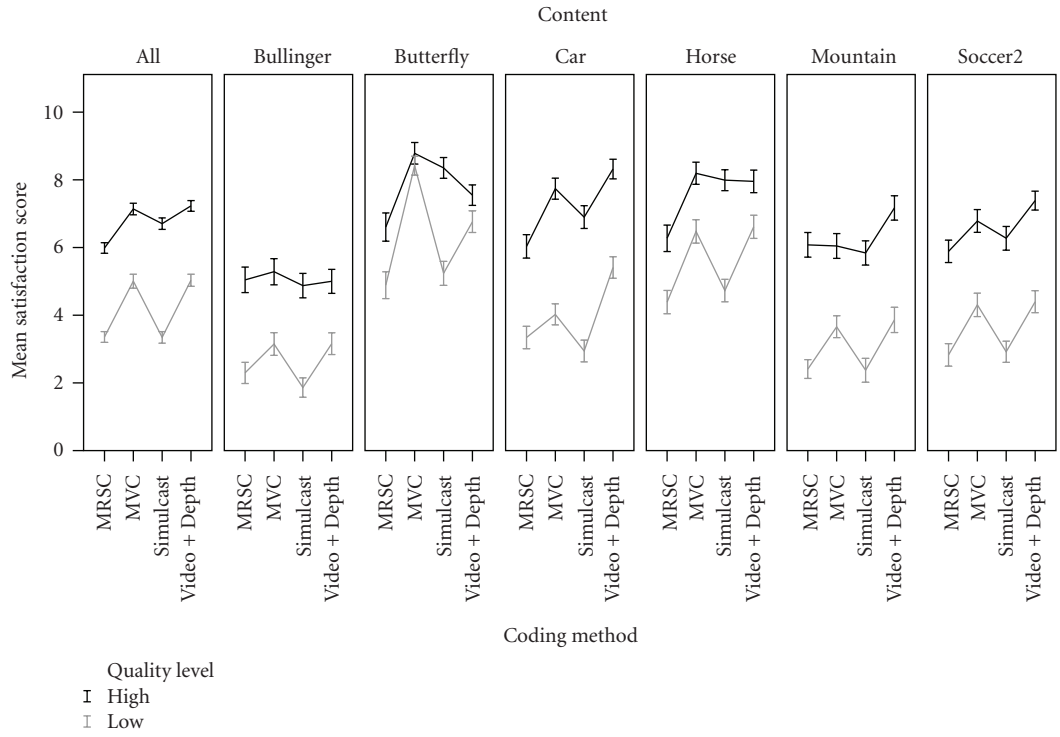


Figure 13: Mean satisfaction scores for the coding methods given at high and low quality levels. Error bars show 95% CI of mean.

6.1.5. Method of Analysis

Psychoperceptual Evaluation, Sensory Profiling, and External Preference Mapping. The analysis was identical to experiment 2.

6.2. Results

6.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality. All coding methods provide highly acceptable quality at the high quality level, 80% at the minimum. At the low quality level, MVC and V + D still reached 60% of acceptance, while the acceptance for MRSC and Simulcast was below 40%.

The distributions of acceptable and unacceptable ratings on the satisfaction scale differ significantly ($\chi^2(10) = 2368$, $P < .001$). The scores for nonaccepted overall quality are found to be between 1.4 and 4.2 (Mean: 2.8, SD: 1.4). Accepted quality was expressed with ratings between 4.5 and 8.5 (Mean: 6.5, SD: 2.0). The Acceptance Threshold can be determined as being between 4.2 and 4.5.

Overall Quality Satisfaction. At the high quality level, coding methods had an influence on quality satisfaction (Fr = 241.83, $df = 3$, $P < .001$; Figure 13). MVC and Video + Depth provided the highest overall quality satisfaction scores when averaging over the content (MVC versus V + D: $Z = -.828$; $P > .05$; ns), outperforming MRSC and Simulcast (all pairwise comparisons: $P < .001$). The results were confirmed for low quality level (Fr = 648.97, $df = 3$, $P < .001$), where MVC and Video + Depth outperform MRSC and Simulcast (all pairwise comparisons $P < .05$).

Content-by-content analysis showed that Video + Depth outperformed all other methods at the high and low quality levels (all comparisons $P < .01$). For Butterfly content, MVC had the best satisfaction scores for both quality levels (all comparisons: $P < .01$). Coding methods did not have an influence on Bullinger content at the high quality level (Fr = 2.942; $df = 3$; $P > .05$; ns).

6.2.2. Sensory Profiling. A number of 15 assessors in the sensory profiling session developed a total of 102 individual quality attributes.

Identification of Dimensions and Attributes. Considering Lawless and Heymann's rule of interpretability [3], two dimensions were identified to be important for the GPA model. The first two components of the GPA model had 88.36% explained variance, where PC1 covered the majority of explained variance (83.32%). Figure 14 shows the item plot and Figure 15 the correlation plot of the GPA model. The analysis emphasizes attributes explaining more than 50% of the variance. As can be identified from the plots, PC1 is mainly determined by video quality. PC2 is discriminating the items (Figure 14) into items with high amount of motion (soccer) and low amount of motion (bullinger).

Dimension 1 ("video quality", 83.32% explained variance). PC1 shows a high correlation of its negative polarity with attributes like "blurry," "blocky," or "grainy." On its

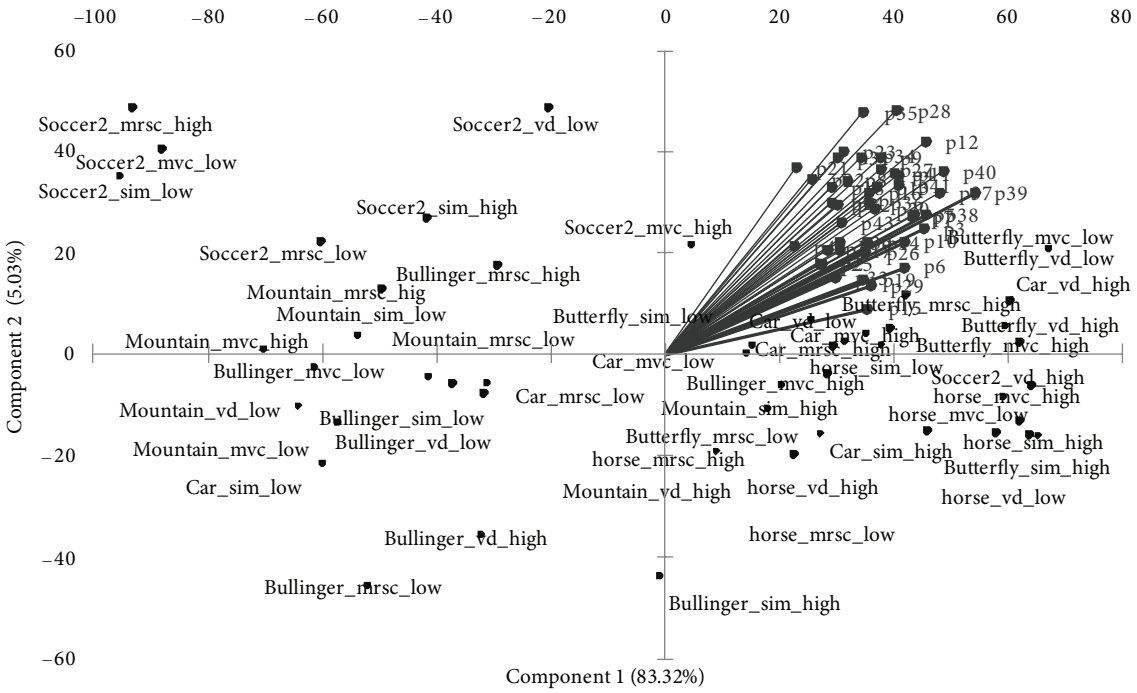


Figure 14: The item plot of the GPA model showing the first two principal components and the test items within the space. Gray arrows mark users' preferences which were mapped into the model using PREFMAP.

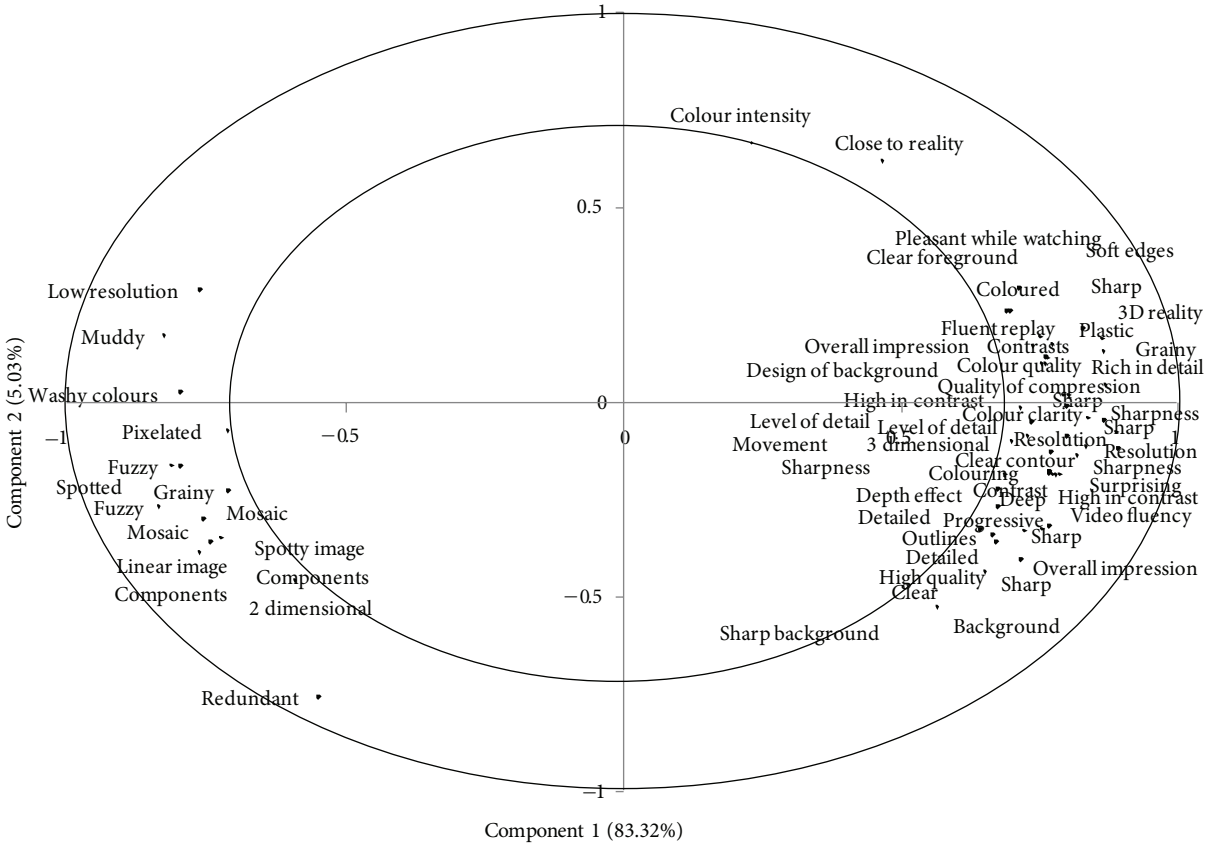


Figure 15: Correlation plot of the experienced quality factors. The figure shows the first two principal components of the GPA model and the correlation of the attributes with these components. Inner and outer circles show 50% and 100% explained variance, respectively.

positive polarity it correlates with attributes like "sharp," "detailed," and "resolution." This component describes the video quality. It separates the model into good and bad quality. The bad quality mainly contains descriptions of artifacts.

Dimension 2 ("amount of motion", 5.03% explained variance). Along PC2, static test content (Bullinger, Mountain, Horse) and content containing motion (Butterfly, Soccer2, Car) are separated (Figure 14). It is remarkable that the explained variance of PC2 is very small compared to the first dimension. However, it is reasonable that the amount of motion is impacting on perceived quality due to the applied coding methods. No attributes were identified to describe the perception of motion.

A separate depth component was not identified in the GPA model. The correlation plot shows that 3D-related attributes like "spacious," "3D reality," or "background depth" correlate with the positive polarity of PC 1. The results show that depth descriptions seem to be part of good quality. If video quality is low due to coding artifacts, this quality degradation will exceed the additional value provided by the stereoscopic video presentation. Depth will not be taken into account to describe quality.

6.2.3. External Preference Mapping. The results show a preference for artifact-free content (Figure 12). The content with the highest user preference is identified along PC1. The least preferred items are all Bullinger clips at the opposite side of the marks. It can also be seen that the Bullinger clips correlate with an attribute called "redundant." Although this attribute only appeared once, it may explain the quantitative results of Bullinger clips. Quantitative analysis has shown that the differences between coding methods were rather small for Bullinger content. The "redundancy" of the Bullinger items may show that the participants evaluated the content on a more affective level, not on its provided quality.

6.3. Discussion and Conclusion. Our results of psychoperceptual evaluation showed that Multiview Coding and Video + Depth provide the highest experienced quality among the tested coding methods. They also represent contrary methods in the coding of 3D video. While MVC uses inter- and intraview dependences of the two video streams (left and right eye), the Video + Depth approach renders virtual videos from a given view and its depth map [103]. In addition, the provided quality level was highly acceptable compared to previous studies [40].

The results of sensory profiling showed that artifacts are still the determining quality factor for 3D. The expected added value through depth perception was rarely mentioned by the test participants. When mentioned, it was connected to the artifact-free video. These results are in line with previous studies concluding that depth perception and artifacts both determine 3D quality perception [88, 104]. In contrast to Seuntïens' model [88], our profiles showed a hierarchical dependency between depth perception and artifacts. When the visibility of artifacts is low, depth perception seems to contribute to the added value of 3D. To conclude, this experiment confirms the findings of experiment 2. With respect to stereo video coding methods, we can see that the compression of the depth map in Video + Depth approaches directly impacts depth quality. In contrast, depth is not affected in Video + Video approaches by related coding methods. Further work needs to investigate more deeply the interaction between artifacts and depth to improve coding methods for mobile stereo video.

7. Discussion and Conclusions

The aim of this paper was to present a novel method, OPQ for multimedia quality evaluation with naïve participants. As a mixed method, OPQ combines a conventional quantitative psychoperceptual evaluation and qualitative descriptive quality evaluation to gain deeper understanding of the quality factors. We applied the method presented in three audiovisual quality evaluation experiments.

7.1. Convergence and Complementation. Our three studies highlighted the complementation and convergence between the results acquired, with different methods underlining the positive features of mixed method research [24]. The results are summarized in Table 8. They complemented each other in all studies, and even more importantly quantitative quality preferences were explained by qualitative descriptions. For example, when quantitative excellence between stimuli was not identified, the qualitative results showed the detectable differences between the used variables, inferiority nullified the positive influence of quality (audiovisual depth), and the participants' sensorial preferences can contribute to final multimodal quality evaluations.

Furthermore, we were able to explain the excellence between parameter comparisons by understanding the relationship between quality and depth using sensory profiling. The descriptions of depth and error-freeness were attached to good quality when visual presentation mode and coding factors were varied. Without qualitative data, the reasons beyond the quantitative data had been based on assumptions, while sensorial data as a single method is not capable to show preferences.

The convergence between the results was represented in the whole affective dimension. Poorly rated quality was attached to badness, inferiority, and erroneousness. In the neutral case, the influences were not visible in any of the measures (e.g., the variables of constructed audio quality factors did not influence quality ratings nor were they described in sensory profiling). Similarly, the most satisfying variable was visible in both measures, consistently showing the high quality ratings and goodness of layered visual depth. Finally, one of our case studies also showed a slightly contradictory aspect between the results. The results showed that participants were able to identify the preferences between visual stimuli while these differences were not visible as such in the results of sensory profiling. This may indicate that for the stimuli with small differences, naïve participants are able to express overall quality preferences quantitatively,

Table 8: Summary of results of three experiments using the open profiling of quality.

Experiment 1
Goal: influence of audiovisual depth on perceived quality
Psychoperceptual evaluation: nonsignificant influence
Sensorial profiling: impressions of content, visual presentation mode and room acoustic
Visual presentation mode: 2D described as error-free while 3D mainly as erroneousness, some positive mentions (descripstions of 3D, spacious, tangible).
Audio: impressions of room space, divided according to room acoustic models
Participants' sensorial preferences towards audio, video, or both modalities
External preference mapping: N/A
Experiment 2
Goal: influence of audiovisual depth on experienced quality of mobile 3D television and video
Psychoperceptual evaluation: dominance of visual quality factors over audio and audiovisual quality
Sensorial profiling: impressions of visual presentation mode and affective factors
Visual presentation mode: 2D—neutral/positive quality descriptions (pleasant, beautiful, focusable);
3D—positive (spatial, illusory, layered, depth impression), negative (artifacts, flickering, stressful, blurred)
Experienced added value of depth is visible only if the level of artifacts is low
External preference mapping: one content is clearly preferred in 3D described as spatial, to come forth, and three-dimensional while others in 2D described as interesting, stress-free, pleasant
Experiment 3
Goal: influence of video coding methods on perceived quality of mobile 3D television and video
Psychoperceptual evaluation: the most satisfying quality provided by multiview coding and video + depth coding methods when averaged over the contents; some content-dependencies exist
Sensorial profiling: added value of depth conveyed when level of artifacts is low
Visual 3D—positive quality (sharp, detailed, resolution) versus negative quality (blurry, blocky grainy)
External preference mapping: artifact-free videos preferred; exceptional contents identified

while their ability to express them or sensibility in sensory profiling can be limited [111, 112] or may be guided by the most changing variables according to peak end theory [102]. To sum up, the benefits of using OPQ as a mixed method for multimedia quality evaluation were expressed as the ability to provide the complementation, and as mainly the ability to explain quantitative results with qualitative descriptions. Further work needs to systematically probe the method with small and detectable differences in multimodal stimuli with naïve participants to expand the consciousness of the limitations of its use.

7.2. Further Work. The other aspects of the further development of the OPQ method are mainly targeted on the sample and on conducting and analyzing the results. For multimodal quality evaluation studies with naïve participants, according to our results, it is worth considering a well-validated tool for identifying the groups of different information processing styles (e.g., [90]) and reporting these groups to characterize the sample. Our experience using OPQ has repeatedly highlighted the importance of training and careful attribute development in the sensorial studies. Individual differences in the ability to describe properties accurately are not only a typically reported challenge in food science [52], but it also seems to be present in multimedia quality studies. Based on our informal observations, we have noticed that the apple description task in the training, as something concrete and familiar, helps participants to start to create their descriptions. We have also observed the importance of giving enough time for attribute elicitation and refinement tasks which can contribute to the success of the final sensory evaluation. As the last remark, the use of OPQ requires participation in multiple sessions. In general, drop-out rates can be a problem of construct validity in multi-session studies [60]. Although we did not face this problem in our studies, it is good for practitioners to keep this limitation in mind if considering small sample sizes for the sensory profiling task.

Based on our experience with the analysis of interview-based descriptive data (e.g., [16, 42]), the analysis of sensorial data seems to be comparably quick and straightforward. However, there are four main suggestions to be considered in further work. Firstly, the guidelines for detecting outliers in this data are needed. While in quantitative results, outliers can be detected and removed, sensory evaluation methods do not provide robust methods that can be applied. However, the residuals given for each configuration after GPA [46] show large differences between the most important (low residual) and the least important configuration (high

residual). These residuals may provide the possibility for outlier detection [113]. Secondly, the issue of dominance of components needs to be addressed. The PC1 of the perceptual models in study 2 and study 3 is very dominating, that is, larger than 60% of the explained variance. This may lead to a loss of information in the components of lower explained variance and eventually to an incomplete understanding of perceptual mechanisms. Thirdly, the reliability aspects of the interpretation of the perceptual spaces in sensory profiling and external preference mapping need to be further considered. Currently, the results of GPA and EPM charts can be constructed based on one researcher's interpretation. For example, in the interview-based data-driven analysis (e.g., Grounded theory, content analysis), the reliability aspect is considered using interrater reliability estimations and reviews of multiple independent researchers [114]. A similar type of procedure needs to be considered to improve the reliability of interpretations of results in GPA and EPM charts. Fourthly, the impact of different methods of analysis needs to be investigated. A comparison of GPA and Kunert and Quannari's approach [76] returned a stronger GPA model in terms of explained variance for Generalized Procrustes Analysis. The same investigations need to be done for the methods of External Preference Mapping (PREFMAP, PLS) to understand similarities and differences and to minimize the impact of different methods on the results.

Finally, systematical comparisons between OPQ and existing methods are needed to provide guidelines for an effective use of these methods for the practitioners. To probe aspects in the comparisons, OPQ can provide a relatively easy data-collection and analysis procedure, but, on the other hand requires multiple evaluation sessions. In contrast, interview-based methods can require good interviewing skills of personnel, a relatively slow procedure in the analysis while they can complete the whole study with one-session participation. The systematical comparisons need to verify performance-related aspects (e.g., accuracy in different quality range, validity, reliability, and costs), complexity (e.g., ease of planning, conducting and analyzing, and interpreting results), evaluation factors (e.g., number of stimuli, knowledge of research personnel) (e.g., [115–117]). The long-term goal is to support the idea of safe development of these instruments by understanding their benefits and limitations when capturing deeper understanding of experienced multimedia quality.

Acknowledgments

MOBILE3DTV project has received funding from the ICT programme of the European Community in the context of the Seventh Framework Programme (FP7/2007–2011) under Grant agreement no. 216503. The paper reflects only the authors views, and the European Community or other project partners are not liable for any use that may be made of the information contained herein. The work of the second author is supported by the Graduate School in User-Centered Information Technology (UCIT).

References

[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation—Theory, Method and Application*, John Wiley & Sons, Chichester, UK, 2006.

[2] P. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, Mass, USA, 2000.

[3] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food: Principles and Practices*, Chapman & Hall, New York, NY, USA, 1999.

[4] U. Neisser, *Cognition and Reality: Principles and Implications of Cognitive Psychology*, W.H. Freeman, San Francisco, Calif, USA, 1976.

[5] J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, Mass, USA; Lawrence Eribaum, Mahwah, NJ, USA, 1979.

[6] E. B. Goldstein, *Sensation and Perception*, Thomson Wadsworth, Belmont, Calif, USA, 7th edition, 2007.

[7] S. T. Fiske and S. E. Taylor, *Social Cognition. Singapore*, McGraw-Hill, New York, NY, USA, 1991.

[8] Recommendation ITU-R BT.500-11, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Recommendation ITU-R BT.500-11. ITU Telecom. Standardization Sector of ITU, 2002.

[9] S. Jumisko-Pyykkö and J. Häkkinen, "Profiles of the evaluators—impact of psychographic variables on the consumer-oriented quality assessment of mobile television," in *Multimedia on Mobile Devices*, vol. 6821 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2008.

[10] C. Cui, "Do experts and naive observers judge printing quality differently?" in *Imaging Quality and System Performance*, vol. 5294 of *Proceedings of SPIE*, pp. 132–145, San Jose, Calif, USA, January 2004.

[11] S. R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 4, pp. 241–257, 2006.

[12] S. R. Gulliver and G. Ghinea, "Stars in their eyes: what eye-tracking reveals about multimedia perceptual quality," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 34, no. 4, pp. 472–482, 2004.

[13] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, and A. Tirel, "Task impact on the visual attention in subjective image quality assessment," in *Proceedings of the 14th European Signal Processing Conference (EUSIPCO '06)*, Florence, Italy, September 2006.

[14] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?" in *Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '08)*, pp. 63–72, ACM, September 2008.

[15] H. Knoche and M. A. Sasse, "The big picture on small screens delivering acceptable video quality in mobile TV," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 5, no. 3, article 20, 2009.

[16] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman, "Experienced quality factors—qualitative evaluation approach to audiovisual quality," in *Multimedia on Mobile Devices 2007*, vol. 6507 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2007, convention paper 6507-21.

[17] J. Radun, T. Leisti, J. Häkkinen, et al., "Content and quality: interpretation-based estimation of image quality," *Transactions on Applied Perception*, vol. 4, no. 4, pp. 1–5, 2008.

[18] S. Bech, R. Hamberg, M. Nijenhuis et al., "Rapid perceptual image description (RaPID) method," in *Human Vision and Electronic Imaging*, vol. 2657 of *Proceedings of SPIE*, pp. 317–328, San Jose, Calif, USA, January-February 1996.

[19] G. Lorho, "Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction," in *Proceedings of Audio Engineering Society 119th Convention*, New York, NY, USA, 2005, convention paper 6629.

[20] G. Lorho, "Perceptual evaluation of mobile multimedia loudspeakers," in *Proceedings of Audio Engineering Society 119th Convention*, Vienna, Austria, 2007.

[21] G. Nyman, J. Radun, T. Leisti et al., "What do users really perceive—probing the subjective image quality," in *Image Quality and System Performance III*, vol. 6059 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2006.

[22] R. B. Johnson and A. J. Onwuegbuzie, "Mixed methods research: a research paradigm whose time has come," *Educational Researcher*, vol. 33, no. 7, pp. 14–26, 2004.

[23] A. Tashakkori and C. Teddlie, "Quality of inferences in mixed methods research," in *Advances in Mixed Methods Research*, M. Bergman, Ed., Sage, London, UK, 2008.

[24] N. K. Denzin, *The Research Act: An Introduction to Sociological Methods*, McGraw-Hill, New York, NY, USA, 1978.

[25] K. Nahrstedt and R. Steinmetz, "Resource management in networked multimedia systems," *Computer*, vol. 28, no. 5, pp. 52–63, 1995.

[26] G. Wikstrand, "Improving user comprehension and entertainment in wireless streaming media," Tech. Rep., Introducing Cognitive Quality of Service, Department of Computer Science, Umeå University, Umea, Sweden, 2003.

[27] ITU-T Recommendation P.10 Amendment 1, *Vocabulary for performance and quality of service. New Appendix I Definition of Quality of Experience (QoE)*, International Telecommunication Union, Geneva, Switzerland, 2008.

[28] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, "Quality of experience in distributed interactive multimedia environments: toward a theoretical framework," in *Proceedings of the ACM Multimedia Conference, with Co-Located Workshops and Symposiums (MM '09)*, pp. 481–490, ACM, Beijing, China, October 2009.

[29] D. S. Hands, "A basic multimedia quality model," *IEEE Transactions on Multimedia*, vol. 6, no. 6, pp. 806–816, 2004.

[30] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[31] G. Ghinea and J. P. Thomas, "QoS impact user perception and understanding of multimedia video clips," in *Proceedings of the 6th ACM International Conference on Multimedia (ACM Multimedia '98)*, pp. 49–54, Bristol, UK, September 1998.

[32] S. Jumisko-Pyykkö, ""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*, vol. 36, no. 1-2, pp. 167–184, 2008.

[33] Recommendation ITU-T P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," Recommendation ITU-T P.910. ITU Telecom. Standardization Sector of ITU, 1999.

[34] F. Kozamernik, P. Sunna, E. Wyckens, and D. I. Pettersen, "Subjective quality of internet video codecs—phase 2 evaluations using SAMVIQ," *EBU Technical Review*, no. 301, 2005.

[35] M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, and K. Brunnström, "Subjective multimedia quality assessment," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 11, pp. 2920–2932, 2006.

[36] D. M. Rouse, R. Pépion, P. Le Callet, and S. S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment," in *Human Vision and Electronic Imaging XV*, vol. 7527 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2010.

[37] S. Jumisko-Pyykkö and D. Strohmeier, "Report on research methodologies for the experiments," Technical Report Project Mobile 3DTV, November 2008.

[38] S. R. Gulliver, T. Serif, and G. Ghinea, "Pervasive and standalone computing: the perceptual effects of variable multimedia quality," *International Journal of Human Computer Studies*, vol. 60, no. 5-6, pp. 640–665, 2004.

[39] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, and M. M. Hannuksela, "Acceptance threshold: a bidimensional research method for user-oriented quality evaluation studies," *International Journal of Digital Multimedia Broadcasting*, vol. 2008, Article ID 712380, 20 pages, 2008.

[40] S. Jumisko-Pyykkö and T. Utriainen, "User-centered quality of experience of mobile 3DTV: how to evaluate quality in the context of use?" in *Multimedia on Mobile Devices*, vol. 7542 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2010.

[41] H. Knoche, J. D. McCarthy, and M. A. Sasse, "Can small be beautiful? Assessing image size requirements for mobile TV," in *Proceedings of the 13th ACM International Conference on Multimedia (ACM Multimedia '05)*, Singapore, November 2005.

[42] S. Jumisko-Pyykkö, U. Reiter, and C. Weigel, "Produced quality is not perceived quality—a qualitative approach to overall audiovisual quality," in *Proceedings of the 1st International Conference on 3DTV (3DTV-CON '07)*, Kos, Greece, May 2007.

[43] H. Coolican, *Research Methods and Statistics in Psychology*, J. W. Arrowsmith, London, UK, 4th edition, 2004.

[44] H. Stone and J. L. Sidel, *Sensory Evaluation Practices*, Academic Press, San Diego, Calif, USA, 3rd edition, 2004.

[45] N. Zacharov and K. Koivuniemi, "Audio descriptive analysis & mapping of spatial sound displays," in *Proceedings of the International Conference on Auditory Displays*, 2001.

[46] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[47] F. R. Jack and J. R. Piggott, "Free choice profiling in consumer research," *Food Quality and Preference*, vol. 3, no. 3, pp. 129–134, 1991.

[48] J. Delarue and J.-M. Sieffermann, "Sensory mapping using flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products," *Food Quality and Preference*, vol. 15, no. 4, pp. 383–392, 2004.

[49] J. Häkkinen, T. Kawai, J. Takatalo et al., "Measuring stereoscopic image quality experience with interpretation based quality methodology," in *Image Quality and System Performance V*, vol. 6808 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2008.

[50] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, Sage, Thousand Oaks, Calif, USA, 2006.

[51] A. Strauss and J. Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, Sage, Thousand Oaks, Calif, USA, 4th edition, 1998.

[52] P. Faye, D. Brémaud, M. Durand Daubin, P. Courcoux, A. Giboreau, and H. Nicod, "Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings," *Food Quality and Preference*, vol. 15, no. 7-8, pp. 781–791, 2004, 5th Rose Marie Pangborn Sensory Science Symposium.

[53] D. Picard, C. Dacremont, D. Valentin, and A. Giboreau, "Perceptual dimensions of tactile textures," *Acta Psychologica*, vol. 114, no. 2, pp. 165–184, 2003.

[54] T. Shibata, S. Kurihara, T. Kawai et al., "Evaluation of stereoscopic image quality for mobile devices using interpretation based quality methodology," in *Stereoscopic Displays and Applications XX*, vol. 7237 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2009.

[55] Thesaurus, 2010, http://thesaurus.com/.

[56] MOT Oxford Dictionary of English 1.0. Oxford University Press Great Clarendon Street, Oxford OX2 6DP—The Oxford Dictionary of English, Revised Edition © Oxford University Press, 2005. Retrieved, 2010.

[57] ISO 8586-1, *Sensory analysis—general guidance for the selection, training and monitoring of assessors—part 1: selected assessors*, International Standardization Organization, 1993.

[58] ISO 8586-2, *Sensory analysis—general guidance for the selection, training and monitoring of assessors—part 2: experts*, International Standardization Organization, 1994.

[59] ISO 7029, *Statistical distribution of hearing threshold as a function of age*, International Standardization Organization, 2000.

[60] W. Shadish, T. Cook, and D. Campbell, *Experimental and Quasi-Experimental Designs*, Houghton Mifflin, Boston, Mass, USA, 2002.

[61] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Erlbaum, Hillsdale, NJ, USA, 1988.

[62] J. A. McEwan, "Preference mapping for product optimization," in *Multivariate Analysis of Data in Sensory Science*, T. Naes and E. Risvik, Eds., Elsevier, Amsterdam, The Netherlands, 1996.

[63] P. Faye, D. Brémaud, E. Teillet, P. Courcoux, A. Giboreau, and H. Nicod, "An alternative to external preference mapping based on consumer perceptive mapping," *Food Quality and Preference*, vol. 17, no. 7-8, pp. 604–614, 2006.

[64] S. Lê, J. Josse, and F. Husson, "FactoMineR: an R package for multivariate analysis," *Journal of Statistical Software*, vol. 25, no. 1, pp. 1–18, 2008.

[65] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 188–193, 2000.

[66] A. A. Williams and S. P. Langron, "The use of free-choice profiling for the evaluation of commercial ports," *Journal of the Science of Food and Agriculture*, vol. 35, pp. 558–568, 1984.

[67] P. N. Jones, H. J. H. McFie, and S. L. Beilken, "Use of preference mapping to relate consumer preference to the sensory properties of a processed meat product (tinned cat food)," *Journal of the Science of Food and Agriculture*, vol. 47, pp. 113–123, 1989.

[68] A. A. Williams and G. M. Arnold, "Comparison of the aromas of six coffees characterized by conventional profiling, free-choice profiling and similarity scaling methods," *Journal of the Science of Food and Agriculture*, vol. 36, pp. 204–214, 1985.

[69] J. M. Murray, C. M. Delahunty, and I. A. Baxter, "Descriptive sensory analysis: past, present and future," *Food Research International*, vol. 34, no. 6, pp. 461–471, 2001.

[70] D. C. Oreskovich, B. P. Klein, and J. W. Sutherland, "Procrustes analysis and its applications to free choice and other sensory profiling," in *Sensory Science Theory and Applications in Foods*, H. T. Lawless and B. P. Klein, Eds., pp. 353–394, Marcel Dekker, New York, NY, USA, 1991.

[71] J. A. McEwan, J. S. Colwill, and D. M. H. Thomson, "The application of two freechoice profile methods to investigate the sensory characteristics of chocolate," *Journal of Sensory Studies*, vol. 3, pp. 271–286, 1989.

[72] G. A. Kelly, *The Psychology of Personal Constructs*, Norton, New York, NY, USA, 1955.

[73] J. B. E. M. Steenkamp and H. C. M. Van-Trijp, "Free-choice profiling in cognitive food acceptance research," in *Food Acceptability*, D. H. M. Thompson, Ed., pp. 363–378, Elsevier Applied Science, London, UK, 1988.

[74] J. R. Piggott and M. P. Watson, "Comparison of freechoice profiling and the repertory grid method in the flavor profiling of cider," *Journal of Sensory Studies*, vol. 7, pp. 133–145, 1992.

[75] D. Dijksterhuis, "Procrustes analysis in sensory research," in *Multivariate Analysis of Data in Sensory Science*, T. Naes and E. Risvik, Eds., pp. 185–220, Elsevier, Amsterdam, The Netherlands, 1996.

[76] J. Kunert and E. M. Qannari, "A simple alternative to generalized Procrustes analysis: application to sensory profiling data," *Journal of Sensory Studies*, vol. 14, no. 2, pp. 197–208, 1999.

[77] P. Schlich, "Preference mapping: relating consumer preferences to sensory or instrumental measurements," in *Bioflavour 95*, P. Etievant and P. Schreiner, Eds., INRA, Versailles, France, 1995.

[78] H. Abdi, "Partial least squares regression (PLS-regression)," in *Encyclopedia for Research Methods for the Social Sciences*, M. Lewis-Beck, A. Bryman, and T. Futing, Eds., pp. 792–795, Sage, Thousand Oaks, Calif, USA, 2003.

[79] D. Strohmeier and S. Jumisko-Pyykkö, "How does my 3D video sound like?—Impact of loudspeaker set-ups on audiovisual quality on mid-sized autostereoscopic display," in *Proceedings of the True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON '08)*, pp. 73–76, Istanbul, Turkey, May 2008.

[80] W. Ijsselsteijn, J. Freeman, D. Bouwhuis, and H. de Ridder, "Presence as an experiential metric for 3-D display evaluation," in *Proceedings of the Society for Information Display International Symposium*, pp. 19–24, Boston, Mass, USA, May 2002.

[81] U. Reiter, *Bimodal audiovisual perception in interactive application systems of moderate complexity*, Ph.D. thesis, TU Ilmenau, 2009.

[82] R. Storms, *Auditory-visual cross-modal perception phenomena*, doctoral dissertation, Naval Postgraduate School, Monterey, Calif, USA, 1998.

[83] U. Reiter and U. Kühhirt, "Object-based A/V application systems: IAVAS I3D status and overview," in *Proceedings of IEEE International Symposium on Consumer Electronics (ISCE '07)*, Irving, Tex, USA, June 2007.

[84] Recommendation ITU-R BS.1116-1, "Method for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," Recommendation ITU-R BS.1116-1. ITU Telecom. Standardization Sector of ITU, 1997.

[85] R. Kennedy, N. Lane, K. Berbaum, and M. Lilienthal, "Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness," *International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[86] S. Jumisko-Pyykkö, T. Utriainen, D. Strohmeier, A. Boev, and K. Kunze, "Simulator sickness—five experiments using autostereoscopic mid-sezed or small mobile screens," in *Proceedings of the True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON '10)*, Tampere, Finland, 2010.

[87] S. Jumisko-Pyykkö and T. Utriainen, "User-centered quality of experience: is mobile 3D video good enough in the actual context of use?" in *Proceedings of the 5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM '10)*, Scottsdale, Ariz, USA, January 2010.

[88] P. J. H. Seuntïens, *Visual experience of 3D TV*, Ph.D. thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2006.

[89] W. A. Ijsselsteijn, H. de Ridder, and J. Vliegen, "Subjective evaluation of stereoscopic images: effects of camera parameters and display duration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 225–233, 2000.

[90] T. L. Childers, M. J. Houston, and S. E. Heckler, "Measurement of individual differences in visual versus verbal information processing," *Journal of Consumer Research*, vol. 12, pp. 125–134, 1985.

[91] L. B. Stelmach, W. J. Tam, D. V. Meegan, A. Vincent, and P. Corriveau, "Human perception of mismatched stereoscopic 3D inputs," in *Proceedings of International Conference on Image Processing (ICIP '00)*, vol. 1, pp. 5–8, September 2000.

[92] J. Häkkinen, M. Pölönen, J. Takatalo, and G. Nyman, "Simulator sickness in virtual display gaming: a comparison of stereoscopic and non-stereoscopic situations," in *Proceedings of the 8th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '06)*, vol. 159 of *ACM International Conference Proceeding Series*, pp. 227–229, ACM, Espoo, Finland, September 2006.

[93] M. Lambooij, M. Fortuin, W. A. IJsselsteijn, and I. Heynderickx, "Measuring visual discomfort associated with 3D displays," in *Stereoscopic Displays and Applications XX*, Proceedings of SPIE, San Jose, Calif, USA, January 2009.

[94] M. Lambooij, W. Ijsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: a review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, Article ID 0302011, 4 pages, 2009.

[95] S. Jumisko-Pyykkö and T. Utriainen, "User-centered quality of experience: is mobile 3D video good enough in the actual context of use?" in *Proceedings of the 5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM '10)*, Scottsdale, Ariz, USA, January 2010.

[96] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 973–980, 2006.

[97] S. Jumisko-Pyykkö, M. Weitzel, and D. Strohmeier, "Designing for user experience: what to expect from mobile 3D TV and video?" in *Proceedings of the 1st International Conference on Designing Interactive User Experiences for TV and Video (UXTV '08)*, pp. 183–192, Mountain View, Calif, USA, October 2008.

[98] S.-I. Uehara, T. Hiroya, H. Kusanagi, K. Shigemura, and H. Asada, "1-inch diagonal transflective 2D and 3D LCD with HDDP arrangement," in *Stereoscopic Displays and Applications XIX*, vol. 6803 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2008.

[99] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, pp. 51–66, 1999.

[100] W. R. Neuman, A. N. Cringler, and V. M. Bove, "Television sound and viewer perceptions," in *Proceedings of the 9th International Conference on Television Sound Today and Tomorrow*, February 1991.

[101] J. Lessiter and J. Freeman, "Really hear? The effect of audio quality on presence," in *Proceedings of the 4th Annual International Workshop on Presence*, 2001.

[102] B. L. Fredrickson, "Extracting meaning from past affective experiences: the importance of peaks, ends, and specific emotions," *Cognition and Emotion*, vol. 14, no. 4, pp. 577–606, 2000.

[103] G. Tech, A. Smolic, H. Brust et al., "Optimization and comparision of coding algorithms for mobile 3DTV," in *Proceedings of the True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON '09)*, Potsdam, Germany, May 2009.

[104] W. J. Tam, L. B. Stelmach, and P. Corriveau, "Psychovisual aspects of viewing stereoscopic video sequences," in *Stereoscopic Displays and Virtual Reality Systems V*, Proceedings of SPIE, pp. 226–235, San Jose, Calif, USA, January 1998.

[105] ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC JTC 1, Advanced Video Coding for Generic Audiovisual Services, November 2007.

[106] ISO/IEC JTC1/SC29/WG11, Text of ISO/IEC 14496-10:200X/FDAM 1 Multiview Video Coding. Doc. N9978, Hannover, Germany, July 2008.

[107] H. Brust, A. Smolic, K. Mueller, G. Tech, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," in *Proceedings of the True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON '09)*, Potsdam, Germany, May 2009.

[108] ISO/IEC JTC1/SC29/WG11, ISO/IEC CD 23002-3: representation of auxiliary video and supplemental information. Doc. N8259, Klagenfurt, Austria, July 2007.

[109] P. Merkle, Y. Wang, K. Müller, A. Smolic, and T. Wiegand, "Video plus depth compression for mobile 3D services," in *Proceedings of the True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON '09)*, Potsdam, Germany, May 2009.

[110] G. Tech, H. Brust, K. Müller, A. Aksay, and D. Bugdayci, "Development and optimization of coding algorithms for mobile 3DTV," Tech. Rep. D2.5 Mobile3DTV, 2009.

[111] G. E. A. Solomon, "Psychology of novice and expert wine talk," *American Journal of Psychology*, vol. 103, no. 4, pp. 495–517, 1990.

[112] H. T. Lawless, "Flavor description of white wine by "expert" and non-expert wine consumers," *Journal of Food Science*, vol. 49, pp. 120–123, 1984.

[113] T. Dahl and T. Næs, "Outlier and group detection in sensory panels using hierarchical cluster analysis with the Procrustes distance," *Food Quality and Preference*, vol. 15, no. 3, pp. 195–208, 2004.

[114] M. Q. Patton, *Qualitative Research & Evaluation Methods*, Sage, Thousand Oaks, Calif, USA, 3rd edition, 2002.

[115] M. McTigue, H. Koehler, and M. Silbernagel, "Comparison of four sensory evaluation methods assessing cooked dry bean flavour," *Journal of Food Science*, vol. 54, no. 5, 1989.

[116] H. R. Hartson, T. S. Andre, and R. C. Williges, "Criteria for evaluating usability evaluation methods," *International Journal of Human-Computer Interaction*, vol. 15, no. 1, pp. 145–181, 2003.

[117] J. T. Yokuma and J. S. Armstrong, "Beyond accuracy: comparison of criteria used to select forecasting methods," *International Journal of Forecasting*, vol. 11, no. 4, pp. 591–597, 1995.