Visual Perception of Scalable Video Streaming: Applied to Mobile Scenarios

Pengpeng Ni

December 16, 2015

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1710

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: John Grieg AS, Bergen.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Abstract

Modern video coding techniques provide multidimensional adaptation options for adaptive video streaming over networks. For instance, a video server can adjust the frame-rate, frame-size or signal-to-noise ratio (SNR) of the video being requested to cope with the available bandwidth. However, these adaptation operations give rise to distinct visual artefacts, so it follows that they are not perceived in the same manner. Subjective evaluation shows that we can no longer assume a monotonic rate-distortion function for scalable video. In fact, the perceived video quality that is expressed as the overall viewing experience of the video being delivered, is mutually and interactively affected by many factors ranging from configuration parameters to source material.

Performing subjective evaluation is a tedious task, and to facilitate conducting field experiments of quality assessment, we introduce a practical and economic method, denoted by Randomized Paired Comparison. The performance of this method has been examined by experimental practice and simulations. To help formulating optimal adaptation strategies for streaming services, a sequence of field studies have been conducted to evaluate the perceived video quality, with the focus mainly on mobile streaming scenarios. These studies reveal that dynamic bit-rate variations may bring about the so-called flicker effects, which have negative influence on the perceived quality. Furthermore, the perceptual impacts can be controlled by the intensity of bit-rate changes (amplitude) and the number of bit-rate changes per seconds (frequency). The amplitude of bit-rate fluctuation is the most dominant factor. Thus, the greater amplitude an adaptation scheme has, the lower perceived quality will be brought about. Meanwhile, the frequency factor affects visual quality significantly when the bit-rate changes occurs in the spatial domain. To ensure stability of the perceived video quality, the spatial components (pixel values and frame size) of a video should be kept unchanged for a period more than 2 seconds. Moreover, we have explored the acceptance thresholds of quality degradations in different scaling dimensions, and the general preference order of scaling dimension has been suggested. We made also some preliminary analyses of the effect of content type in all of these studies.

Contents

Preface

This is a doctoral thesis submitted to the Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo in partial fulfillment for the degree of Doctor of Philosophy in Computer Science for the year 2015. The work reported in this thesis has been carried out within the context of the Robust Video Streaming Services (ROMUS) project and Perceptual and Cognitive Quality Evaluation Techniques for Audiovisual Systems (PERCEVAL) project at the Media Performance Group, Simula Research Laboratory from September 2007 to 2011. The projects are funded by The Research Council of Norway (NFR) through Simula Research Laboratory.

During this thesis period, I have also become a mother of two lovely children. Being a mother is the greatest gift in my life, although it postponed the thesis writing. I am grateful for the joys my children bring to me, for the love I feel for them.

This thesis is the result of a collaboration between a number of people. I would like to thank my advisors, Dr. Alexander Eichhorn, Prof. Carsten Griwodz and Prof. Pål Halvorsen for interesting discussions, encouragement, ideas, for being patient and guiding me through the study. I also thank Ragnhild Eg for her assistance in my subjective experiments and feedback on my research papers.

List of Figures

List of Tables

Chapter 1

Introduction

We are witnessing a paradigm shift from a technology-centric to a human-centric view in information science. Over the last decades, technological advances have enabled a rich exchange of multimedia information among people. YouTube, Skype and IP-TV broadcast services are only a few popular examples that show the ubiquity of multimedia communication. With the increased human activities on multimedia information, users' expectation for the next generation of multimedia communication services have been set higher from simple content access to the delivery of "best experiences". In today's competitive world of multimedia services, human-centric system designs that focus on enhancing a user's experience are winning the market. For designing such a system, a better understanding of human perception is needed from the beginning. Observing people in their daily environments and analysing their perceptions constitutes therefore our work in this thesis.

1.1 Motivation

As network technologies are becoming widely applied in our daily life, multimedia streaming services are gaining increased popularity. For instance, nowadays, video streaming services already exist both on the Internet and with mobile technologies. Different from video downloading, streaming stored video to a large number of heterogeneous receivers over various networks introduces several challenges with respect to delivered rate and quality.

To cope with the Internet's varying bandwidth, many video streaming systems use adaptive and scalable video coding techniques to facilitate transmission. Furthermore, transfer over TCP is currently the favored commercial approach for on-demand streaming (Adobe, 2010; Move Networks, 2008; Pantos et al., 2010; Zambelli, 2009) where video is progressively downloaded over HTTP. This approach is not hampered by firewalls, and it provides TCP fairness in the network as well as ordered, lossless delivery. Adaptation to the available bandwidth is controlled entirely by the application. Several feasible technical approaches for performing adaptation exist. One frequently used video adaptation approach is to structure the compressed video bit stream into layers. The based layer is a low-quality representation of the original video stream, while additional layers contribute additional quality. Here, several scalable video codec alternatives exist, including scalable MPEG (SPEG) (Huang et al., 2003), Multiple Description Coding (MDC) (Goyal, 2001)

and the Scalable Video Coding (SVC) extension to H.264 (Schwarz et al., 2007). The other alternative is to use multiple independent versions encoded using, for example, the advanced video coding (AVC) (ITU-T and ISO/IEC JTC 1, 2003), which supports adaptation by switching between streams (Adobe, 2010; Move Networks, 2008; Pantos et al., 2010; Zambelli, 2009). Thus, video streaming systems can adaptively change the size or rate of the streamed video (and thus the quality) to maintain continuous playback and avoid large start-up latency and stalling caused by network congestion.

With the growth of streaming services, a multimedia application will be judged not only by the function it performs, but also by its easy of use and the user's experience in using it. Making adaptation decisions that achieve the best possible user experience has become an open research field. Current video scaling techniques allow adaptation in either the spatial or temporal domain (Schwarz et al., 2007). All of the techniques may lead to visual artefacts every time an adaptation is performed. An algorithm must take this into account and, in addition, it must choose the time, the number of times, and the intensity of such adaptations. There arises the question of how to design streaming algorithms that can dynamically adapts to network conditions in real-time to strive for the best possible viewing experience of video.

The knowledge of user perception and the evaluation of the Quality of Experience (QoE) are prerequisites for successful design and improvement of adaptive multimedia streaming services. As an important measure of the end-to-end performance at the services level from the user's perspective, QoE in multimedia streaming can conceptually be seen as the remaining quality after the distortion introduced during the preparation of the content and the delivery through the network until it reaches the decoder at the end device (Wikipedia, 2015b). Being able to measure the QoE in a controlled manner helps the service provider understand what may be wrong with their services, so that the storage and network resources can be allocated appropriately and sufficiently to maintain expected user satisfaction.

QoE can be evaluated by either objective or subjective methods. Both have their own drawbacks. Objective evaluation methods can provide automatic and fast evaluation results, but usually not accurate enough to match human perception. Subjective quality evaluation, on the other hand, is often difficult to manipulate, time-consuming and expensive due to the human involvement in the evaluation processes. At the beginning of our study, there was little research available on the subject of QoE measurements. We therefore research the challenge of understanding the visual perception of scalable video streaming services.

1.2 Problem statement

Despite the rapid development in the field of video compression, the vast majority of streaming video is still encoded by lossy compression algorithms. In other words, a certain degree of video quality is traded for reduced requirements of storage and network resources in the video compression process. Similar trade-offs can also occur during the video transmission process. For instance, scalable coding techniques offer the possibility of reconstructing lower resolution or lower quality signals from partial bitstreams (Ho and Kim, 2006). Adaptive streaming services delivering scalable video may need to make trade-offs between quality and bandwidth requirement by removing parts of a bitstream

along the end-to-end network path. In consequence, compression artefacts associated with each downscaling operation can be introduced into a video stream after each trade-off is made, and even more visual artefacts may be caused by inappropriate adaptation schemes, which all result in various degrees of video quality degradation.

QoE in the context of video streaming services rests mainly in the perceived video quality, which is a typical subjective measure. Apart from being user dependent, it will be influenced by many factors such as the packet loss, the end device capabilities, the viewing environment and the video content characteristics. It is necessary to investigate the impacts of these (as well as other) factors on QoE, so that an adaptation strategies can be designed accordingly. Currently, the only reliable method to assess the video quality perceived by a human observers is to ask human subjects for their opinion, which is termed subjective video quality assessment (Seshadrinathan et al., 2010). Subjective quality assessment tests have to be carried out using the scientific methods in order to ensure reliability of the test results. However, most of the existing subjective quality assessment methods come with their own limitations. Especially, there was not much research work done on the methodology for field studies, although field study has the advantage of capturing people's real experiences.

How to efficiently estimate the quality of real-life experiences in order to guide the design of multimedia streaming services becomes our main research question. In this dissertation, we therefore present our research efforts in the following two areas:

QoE evaluation in the field: A field study is a collection of data that occurs outside of a laboratory setting. It is often done in natural settings and can therefore deliver the most representative information about users. However, the high cost of its organization usually makes field studies hard to implement. How to carefully plan and prepare a field study in order to ensure accurate and efficient data collection, is the first research step before starting any QoE related studies.

Human perception of visual distortions related to active bandwidth adaptation: Unlike offline video encoders, it is probably not feasible to deploy sophisticated algorithms to optimize the on-line trade-offs between bits and distortions. However, applications should formulate adaptation strategies with the consideration of human visual perception in order to deliver a viewing experience ever closer to user expectation. For video with multi-dimensional scalability and mobile streaming scenarios under dynamic network conditions, it is a challenge to measure and predict user perceived quality of video. There are a number of unsolved problems related to human visual perception of different types of video artefacts, such as whether service users have a common preference of quality degradations in spatial domain over temporal domain, how quickly a streaming service should react to the changes of available bandwidth, and how service users perceive visual artefacts for different video contents etc.

1.3 Limitations

In the scope of this thesis, we have focused our work on mobile devices, where the resource availability changes dynamically due to wireless connections and mobility of users. We

examined only mobile devices with screen resolution up to 480x320 pixels and size up to 3.5 inch (89 mm) in our experiments. The main reason for this limitation is the availability of screen technologies when the experiments were performed. We had not the chance to investigate the perceived video quality on mobile devices with higher screen resolution or larger screen size such as tablets computers. The quality of high-definition video was only examined once on a 42 inch computer monitor. Since high-definition video can be displayed on mobile devices with different screen sizes, we have designed a quality assessment tool with these aspects in mind, helping future work to expand knowledge to this filed of study.

1.4 Research methods

The discipline of computer science is divided into three major paradigms as defined by the ACM Education Board (Comer et al., 1989) in 1989. Each of them have roots in different areas of science, although all can be applied to computing. The board states that all paradigms are so intricately intertwined that it is irrational to say that any one is fundamental in the discipline. The three paradigms or approaches to computer science are:

- The *theory paradigm* is rooted in mathematics. It specifies firstly objects of study and hypothesizes relationships between the objects. Then, the hypothesis is proven logically.

- The *abstraction paradigm* is rooted in experimental scientific method. A scientist forms a hypothesis, constructs a model, makes a prediction before designing an experiment. Finally data is collected and analyzed.

- The *design paradigm* is rooted in engineering. A scientist states requirements and specifications, followed by design and implementation of said system. Finally, the system is tested to see if the stated requirements and specifications were met.

This thesis follows mainly the abstraction and design paradigm. The development of the quality assessment method is based on the design paradigm. We first apply some traditional subjective evaluation methods in some user studies and learn from practices the drawbacks of existing methods as well as the requirements of a new method for performing assessment tasks in the field. Then we design the Randomised Pairwise Comparison method as a cost-efficient tool for field study of multimedia QoE. A prototype implementation is created on iOs platform. The reliability of the new method is examined by experiments and simulations.

All the subjective evaluation experiments follow the abstraction paradigm. Hypotheses about human quality perception are tested by statistical experiments. Before an experiment, the quality impacts of visual artefacts are predicted through the control of the influential variables. Then we collect users' perceptual responses and perform statistical data analysis.

1.5 Contributions

Inspired by the necessity and wide application of QoE evaluations, we have carried out a sequence of user studies with the goal to assess the subjective quality of videos. This work was also motivated by the aforementioned development in scalable coding techniques. We investigated the quality impacts of various visual artefacts associated with different bit-rate scaling operations. As the way in which subjective experiments are performed plays a critical role in the final evaluation results, the methodology for design of experiments is also the focus of this thesis.

This dissertation presents some research work in the field of audiovisual quality assessment (Ni et al., 2011a; Eichhorn et al., 2010; Eichhorn and Ni, 2009; Ni et al., 2010). Several issues were addressed that dealt with visual artefact analysis, experimental design, and subjective evaluation methodology etc. We list our contributions as follows:

Development of a method for audiovisual quality assessment: A field study is the fundamental means for the exploration of a realistic multimedia experience. However, the practicality of subjective studies is often threatened by prohibitive requirements, in particular by the participant's time and the budget for recompensation. We introduced Randomized Paired Comparison (R/PC), i.e., an easy-to-use, flexible, economic and robust tool for conducting field studies. With the use of R/PC, an experimenter can easily obtain stable results with an accuracy close to traditional experiment designs at a much lower cost. We demonstrate the efficiency and practicality of R/PC by simulations. For the first time, we quantify, in a heuristic study, the performance difference between R/PC and classical evaluation methods. We prototyped also a software program on iOS to automate the experimental design.

Gathering of subjective evaluations of perceived video quality: We spent a considerable amount of time conducting experiments of subjective evaluation. A large amount of reliable subjective evaluation scores were recorded and can be used as reference when comparing or validating different objective quality metrics. We do not limit ourselves to a single genre of video content, and we therefore collected a rich data set that has wide applicability in video streaming systems.

Subjective evaluation of Scalable Video Coding (SVC): The Scalable Video Coding extension of the H.264-AVC standard provides three different types of scalability for efficient and flexible video adaptation. However, the increased number of scaling options increases also the difficulty of visual quality assessment. We conducted the first study that evaluated the subjective performance of multi-dimensional scalability features in SVC. The study reveals that adaptation decisions for SVC bitstreams should not only be based on bit-rate and layer dependency information alone, as the perceived quality degradation may be non-monotonic to bit-rate reduction and the preferred adaptation paths depend on content and user expectations. The experimental results can help improving the design of objective quality models towards multi-dimensional video scalability, and the evaluation scores from this study can be used to validate the performance of existing and future objective models.

Subjective evaluation of frequent bit-rate adaptation: Optimal bandwidth adaptation is usually achieved via frequent switching between different bit-rate versions

of video segments. To investigate the visual effects and usefulness of frequent bit-rate adaptation, we performed several subjective quality assessment experiments in different scenarios. Our results show that frequent quality variations may create additional visual artefacts denoted flicker effects and it is not worthwhile making quality changes unless the negative impact of flicker on visual quality is eliminated. We associated the clear definition of flicker effect with different types of quality variations. In addition, we found that people can detect slow or irregular frame-rates much easier on large HDTV screens than small screens of mobile devices. Therefore, our suggestions of how to make video adaptation strategies were given with the consideration of screen size of the end devices.

In-depth study on flicker effect: The perception of flicker effects is jointly influenced by multiple factors. To get a better understanding of human quality perception of the flicker effects, we performed a comprehensive set of subjective tests on handheld devices. From the study, we were able to identify the main influential factors on the visibility of flicker effects and determine the threshold quantities of these factors for acceptable visual quality of video. These findings can help improving video adaptation strategies or bit-rate controllers deployed in video streaming services. Since our observations were made about the visual artefacts in general terms, the experimental findings are applicable for both scalable or non-scalable video. This is especially useful for modern HTTP streaming systems which use segmentation to achieve dynamic bandwidth adaptation for non-scalable video. Finally, the flicker effects were explored across different content types of videos. We provided some preliminary analyses of content effects on human quality perception.

1.6 Thesis organization

The rest of this thesis is organized as follows:

Chapter2 presents the related work in the field of adaptive video streaming. It gives an overview of several different but related topics, including video coding techniques, HTTP streaming, audiovisual quality assessments and experimental design.

Chapter3 introduces Randomized Pairwise Comparison (R/PC) as a practical method for audiovisual quality assessments in field. We first present the detailed design of R/PC and then explore the usefulness and the limits of R/PC in this chapter.

Chapter4 presents a field study that evaluated the subjective performance of multi-dimensional scalability supported by H.264-SVC. Quality degradations in spatial domain were compared with temporal quality degradations under the same bit-rate constraints. Some objective video quality metrics were also validated against the subjective evaluations.

Chapter5 first defines flicker effects for visual artefacts caused by frequent bit-rate adaptation with scalable video. We then present three field studies that evaluated the subjective video quality resulting from frequent quality variations in different scenarios, in order to understand whether users consider it beneficial to adapt video quality quickly.

Chapter6 presents a series of user studies that we performed to further investigate the flicker effects. We report our analysis that evaluated the influence of the main factors on the acceptability of frequent bit-rate adaptations.

Chapter7 concludes the thesis by summarising our findings. Finally, it presents ideas for extending our work in the future.

Chapter 2

Background

In this chapter we introduce some background knowledge for a better understanding of the following chapters. The work presented in this thesis touches on several topics in diverse knowledge domains including multimedia compression, networking and psychophysical experiment. Herein, we give a brief overview of these related topics.

2.1 Modern video codecs

Getting digital video from its source (like a camera or a stored clip) to its destination (like a display) involves a chain of processes. Key to this chain are the processes of compression (encoding) and decompression (decoding), in which bandwidth-intensive "raw" digital video is reduced to a manageable size for transmission or storage, and then reconstructed for display. Video compression is necessary since current Internet throughput rates are insufficient to handle uncompressed video in real time. Video compression is also beneficial since more video channels can be sent simultaneously due to a more efficient utilization of bandwidth. From another point of view, the human visual system has its own limitations, which makes the human eyes and brain incapable to perceive the three-dimensional scene flow in full details. It is therefore possible to compress video to a large extent by reducing the requirement imposed on the physical accuracy of the rendered video frames.

Multimedia compression algorithms have been continually developed over the last two decades. Lossy compression, where one permanently trades off some information details for reduced bandwidth requirement, is most commonly used to compress video, audio and image, especially in application such as streaming services. In contrast to lossless compression, lossy compression achieves significantly higher compression ratios at the expense of a certain amount of information loss. For any lossy video codecs, the essential question is then how to achieve the best tradeoff between quality degradation and bit saving. Nowadays, video streaming systems cover a wide range of multimedia applications and need to deliver video in more heterogeneous environments in which the capabilities and working scenarios of end devices vary widely in terms of for example network bandwidth, computing power and screen resolution. New features and functionalities have been added in the development of video codecs in order to address the need for flexibility and customizability of video streaming. For instance, bit-rate scalability of the compressed video is one of the mostly desired features. Modern video codecs have included several compression techniques to enable bit-rate scalability. Bit-rate reduction by frame dropping techniques

is possible with MPEG-2 encoded video. Then, the most popular video codec H.264-AVC supports frame-rate scalability by allowing a more flexible temporal dependence in coded video frames. On the base of layered video coding structure, scalable coding techniques such as SNR scalability and spatial scalability have been defined in the MPEG-2 and MPEG-4 standards (Pereira and Ebrahimi, 2002). Then, with the scalable extension of H.264-AVC, three different types of scalability can be combined so that a multitude of representations with different spatiotemporal resolutions and bit rates can be supported within a single scalable bit stream. An alternative to layered coding techniques is known as multiple description coding, which fragments a single bitstream into several independent sub-streams. Since any sub-stream can be used to decode the original stream, the risk of interrupted playback due to network congestion is reduced. However, none of these techniques can avoid further fidelity losses of the bit-stream signals. Modern video codecs face thus a more complex challenge of how to achieve the balance between perceived quality and compression efficiency under various timing and resource constraints. Currently, H.264-AVC is the most commonly used video coding standard due to its high coding efficiency. In the next subsections, we go through the H.264-AVC standard and its scalable extension, and introduce their implicit rate-distortion optimization question in details.

2.1.1 The H.264-AVC Video Coding standard

H.264/MPEG-4 Advanced Video Coding (H.264-AVC) (ITU-T and ISO/IEC JTC 1, 2003) is a flexible video coding standard that represents current state-of-the-art in the area of versatile highly compressed representation of digital video. Scaling from mobile phone usage to High-Definition television (HDTV) broadcasts, H.264-AVC has been well received in a number of industries such as telecommunications, content production and broadcasting. The dominant factor for the popularity of H.264-AVC is clearly the high rate distortion performance - the ratio *subjective quality/bitrate* (J.Sullivan and Wiegand, 1998), which is achieved through advanced spatial and temporal compression techniques.

The coding structure of H.264-AVC is similar to that of all prior major digital video standards, as shown in figure 2.1. It is based on block-matching prediction (Motion Estimation (ME), Motion Compensation (MC) Intra Prediction) and Discrete Cosine Transform (DCT) coding: Each picture is compressed by partitioning it as one or more groups of macro-blocks, which are blocks of 16x16 luma samples with corresponding chroma samples. Adjacent macro-blocks contains often identical or similar sample values. Given one block as reference, it is possible to predict the other block with its difference to the reference. The difference data value (residual information) has usually low entropy and can therefore be encoded to fewer bits after being transformed and reordered. A macroblock can be predicted by spatially neighboring blocks contained within a video frame or temporally neighboring blocks in other frames. When temporal prediction is used, the compressed video frame is denoted as *Inter-frame*, and its reconstruction replies on its reference pictures. Otherwise, the frame is self-decodable and is denoted as *Intra-frame*. *Inter-frame* takes advantage of temporal similarities between neighboring frames allowing more efficient compression, hence it is the most often used frame type in video streams. According to the reference picture selection, *Inter-frames* are further classified into *P-frames* and *B-frames*, where *P-frames* are forward predicted by earlier decoded

pictures and *B-frames* are bidirectionally predicted by reference pictures in future or past in display order.

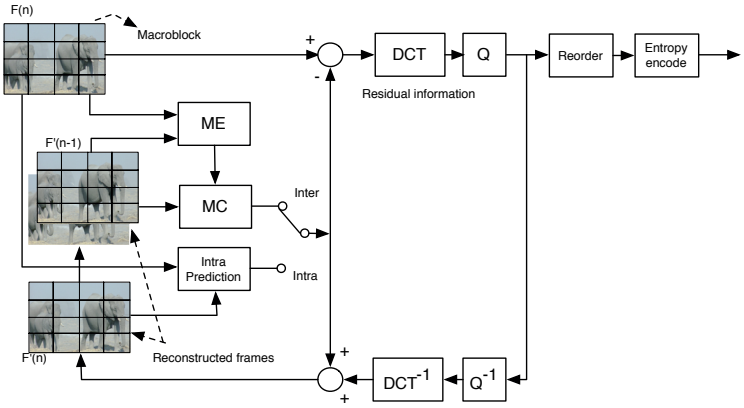


Figure 2.1: H.264-AVC Coding structure

With the concepts of *P-frame* and *B-frame*, H.264-AVC is capable to maintain reasonably high coding efficiency while providing the ability to have multiple frame rates for the same video stream. It is supported with a predetermined temporal prediction structure - hierarchical prediction structure as illustrated in figure 2.2. The hierarchical prediction structure encodes the pictures at lower prediction layers first such that the picture at higher layers can refer to the reconstructed picture at the lower layers. The sequence of firstly encoded pictures are called *key pictures* which represent the coarsest supported temporal resolution. Each key picture is coded as either an *I-frame* that is self-contained or a *P-frame* that uses only previous key pictures as references. Pictures between two key pictures are hierarchically predicted as *B-frame* and can be included layer by layer to refine the temporal resolution. Under low bandwidth situations, *B-frames* can be dropped to save bandwidth without influencing the decoding of the video frames at lower prediction layers, but the bit savings are generally very small.

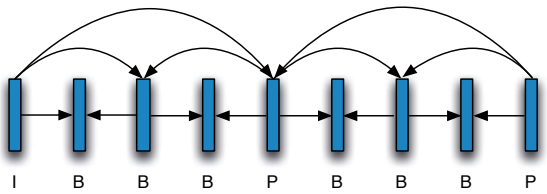


Figure 2.2: Hierarchical prediction structure, the arrows show coding dependency between video frames.

Related to prior video coding methods, H.264-AVC suggests several new features to improve the prediction accuracy. Innovatively, macro-blocks in H.264-AVC can be further

divided into sub-block partitions with seven different sizes - 16x16, 16x8, 8x16, 8x8, 8x4, 4x8 and 4x4. The larger variety of partition shapes provides greater prediction accuracy by taking fine details into consideration. Another feature is the enhanced reference picture selection. In most prior video coding standards like MPEG-2, P-frames use only one picture as reference. In H.264-AVC, *Inter frames* can select, for motion compensation purposes, among a larger number of pictures that have been decoded and stored in the decoder. This gives the encoder a wider range to search for the best match for the blocks to be encoded. Similarly, more features such as fractional pixel search and In-the-loop deblocking filtering etc. are included in the H.264-AVC design (Wiegand et al., 2003). These features enhance the predicting ability and thereby improve the coding efficiency. On the other hand, multiple reference pictures may have the drawback of increasing drift-error propagation among video frames.

As the classical video compression procedure, the residual information after prediction is transformed and then quantized. Quantized coefficients of the transform are scanned and compressed later by entropy coding. Compared to prior video coding standards, integer transform is firstly used in the H.264-AVC instead of a floating point transform, which prevents completely any mismatch between encoder and decoder. Additionally, the H.264-AVC design uses a smaller transform block size. This is perceptually beneficial since it allows the encoder to represent signals in a more locally-adaptive fashion, which reduces artefacts known colloquially as "ringing" (Wiegand et al., 2003). Quantization is the main lossy operation in video compression process. Without quantization, information loss would be very little. But, quantization is still the most efficient way of compressing data. In H.264-AVC, a total of 52 quantization parameters (QP) are supported. It can be noticed that an increase of 1 in quantization parameter means roughly a reduction of bit rate by approximately 12% (Wiegand et al., 2003). In H.264-AVC, the quantization parameter can be specified at the segment, picture or even transform block level. This allows the encoder to flexibly tune of the quantization fidelity according to a model of human sensitivity to different types of error. However, the standard does not include such a model to optimize the quantization process. In addition, H.264-AVC does not have enough flexibility to handle varying streaming conditions in heterogenous environment.

2.1.2 Scalable extension of H.264-AVC

In 2007, the scalable extension of the H.264 standard (H.264-SVC) (Schwarz et al., 2007) was released to support video streaming in more heterogeneous scenarios. Multi-dimensional scalability is designed in add-on fashion on the basis of H.264-AVC. The term "scalability" in this context refers to the removal of parts of the video stream in order to adapt it to varying network conditions or terminal capabilities while still remaining a valid decodable partial stream. The reconstruction quality of the partial streams may have lower temporal, spatial resolution or reduced quantization fidelity compared to the video stream in full scale.

Compared to the video coding structure of H.264-AVC, H.264-SVC encodes the video into a layered structure. As figure 2.3 shows, a fully scalable bit stream consists of several enhancement layers as well as a base layer, where the base layer is an H.264-AVC compliant subset bit stream. H.264-SVC supports combined scalability. Each enhancement layer is a certain combination of spatial, temporal and SNR layers depending on the supported

frame-rate, picture size and Signal-Noise-Ratio(SNR) levels. The total number of layers is thus the number of combinations of the supported spatial, temporal and SNR layers.

Temporal scalability has already been included in H.264-AVC and remains unchanged in H.264-SVC. H.264-SVC proposes some new coding features for spatial and SNR scalability. In the figure 2.3, each horizontal layer corresponds to a supported spatial resolution or a SNR level and is referred to as a spatial or SNR layer. Each spatial or SNR layer is essentially encoded by separate encoders, and the coding process is in principle the same as single-layer encoding. The original video is spatially down-sampled or cropped to several spatial resolutions before sending as input videos to the corresponding encoders. Motion-compensated prediction and intra-prediction are employed to reduce the entropy. In addition, the predictions can come from spatially up-sampled lower layer pictures. Since the information of different layers contains correlations, this so-called inter-layer prediction mechanism reuses the texture, motion and residual information of the lower layers to improve the coding efficiency at the enhancement layer. After prediction and entropy encoding, the bitstreams from all spatial or SNR layers are then multiplexed to form the final SVC bitstream.

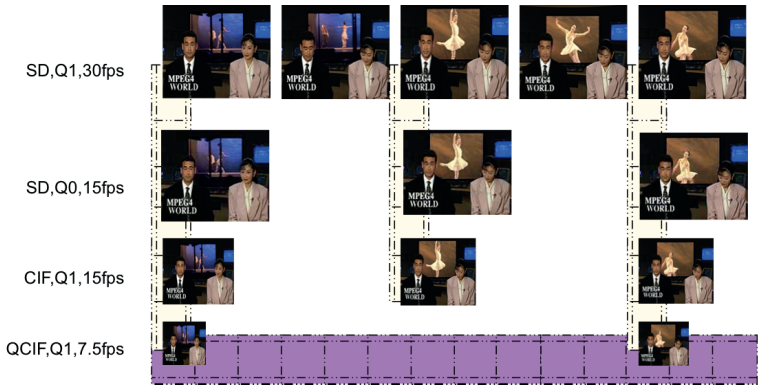


Figure 2.3: Layered video structure of scalable video

The SNR layer is considered a special case of spatial layer with the identical picture sizes for base and enhancement layer. The same inter-layer prediction mechanism is employed but without up-sampling operations. The purpose of including SNR layers in scalable bit streams is to provide coarse-grained SNR scalability (CGS). Differentiation between bit-rates of CGS layers is achieved by quantizing the enhancement layer with a smaller quantization step size relative to that used for the preceding CGS layer. However, the number of CGS layers determines the number of *operation points*[1] for bit-rate downscaling. Just as its name indicates, only a few selected bit rates can be supported by CGS and the relative rate difference between successive CGS layers can not be sufficiently small without negative impact to the coding efficiency. In order to increase the flexibility of bit-rate adaptation, a variation of the CGS approach, which is referred to as medium-grain quality scalability (MGS), is included in the SVC design (Schwarz et al., 2007).

[1] A subset of a scalable video stream that is identified by a set of particular values of scaling parameters. A bitstream corresponding to an operation point can be decoded to offer a representation of the original video at a certain fidelity in terms of spatial resolution, frame-rate and reconstruction accuracy.

The MGS coding approach splits the zigzag-ordered transform coefficients of a quality refinement picture into several fragments corresponding to several MGS layers, as shown in figure 2.4. Any MGS fragment can be discarded during bit-rate adaptation, and by this means, MGS is supposed to provide more finely granular SNR scalability than CGS. The drift propagation due to any loss of an MGS fragment is limited to the neighboring *key pictures*, since only base-quality reconstruction is used for decoding key pictures. The more key pictures are placed in a video sequence, the less quality impact of any possible drift error. On the other hand, the enhancement layer coding efficiency will be decreased when encoding more key pictures. There is again a tradeoff between the coding efficiency and error concealment ability.

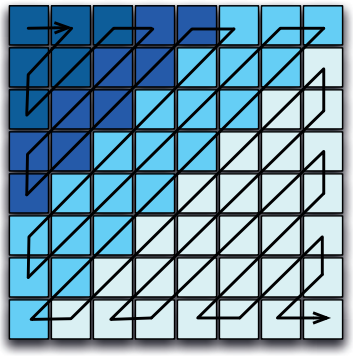


Figure 2.4: Zigzag ordered transform coefficients of a picture which is splited into four fragments to represent four MGS layers.

Compared with previous scalable technologies in MPEG-2 and MPEG-4, H.264-SVC does not make theoretical advances leading to significant improvements in scalable video coding. The streaming flexibility provided by SVC comes at the price of decreased rate-distortion performance. A very complicated encoding controller with multi-layer optimization is required to achieve 10% bit rate increase relative to H.264-AVC single layer encoding (Schwarz et al., 2007). Such an optimized encoder control is, however, not feasible for practical encoders because of the computational complexity.

2.2 Dynamic Adaptive Streaming over HTTP

With the development of video coding techniques, the consumption of digital video has experienced a massive growth. Meanwhile, the popularity of wireless network and mobile devices is also expanding dramatically. Providing ubiquitous access to multimedia content through heterogeneous networks and terminals becomes the goal of Internet streaming services. In recent years, the HTTP protocol has been used extensively for the delivery of multimedia content over Internet. The popularity of HTTP streaming is mainly due to its convenience for end users and content providers. First, as HTTP is a widely used protocol on the Internet, existing low cost Internet infrastructure (HTTP servers, proxies, and caches, etc) can be reused to efficiently support multimedia streaming for millions

of users. Additionally, HTTP is stateless. When an HTTP client sends a data request, the server responds by sending the data, and then the transaction is terminated. In this way, each HTTP request is handled as an independent transaction that is unrelated to any previous request. The HTTP server does not need to retain session information or status about each user for the duration of multiple requests. Therefore, streaming content to a large number of users does not impose any additional load on the server beyond standard web use of HTTP. However, current Internet services offer only best-effort delivery while multimedia streaming is not a best-effort type of traffic. Streaming video over Internet, especially over wireless networks, has its own challenges for supporting video streaming due to its fluctuating capacity. There is a demand for adaptation solutions to improve the robustness and quality of service (QoS) for video transmission. To address this need, several HTTP based adaptive streaming solutions, such as Microsoft Smooth Streaming (Zambelli, 2009), Apple HTTP Live Streaming (Pantos et al., 2010) and Adobe Dynamic Streaming for Flash (Adobe, 2010), have been implemented in the streaming media industry. Later, Dynamic Adaptive Streaming over HTTP (DASH) (Sodagar, 2011; Stockhammer, 2011), also known as MPEG-DASH, has been developed by MPEG since 2010 and was published as an international standard in 2012. The standardization of DASH promotes universal deployment of such an adaptive streaming solution over HTTP.

In DASH, each multimedia content is partitioned into one or more segments. A Media Presentation description (MPD) is associated with each content and describes media information such as segment timing, location on the network, resolution and other content characteristics. Segments contain the actual media data in the form of chunks, stored in single or multiple files. Typically, more than one version of the media data segment at different resolutions or bit-rates is available for adaptive streaming. If a DASH client requests to play a media content, the MPD will first be delivered to the client. Using the MPD information, the DASH client selects appropriate segment representations by fetching the segments using HTTP GET requests. The DASH client fully controls the streaming session, for instance, it can dynamically switch between different media segments encoded at different bit-rates to cope with the varying network condition. Thus, non-scalable video can be also streamed in an adaptive manner at some storage cost for storing multiple representations of video segments at the sever side. In the case of streaming scalable content, a DASH client decides then on-the-fly when and how to switch between different layers. Alternatively, highly scalable content is often delivered by other streaming solution such as priority progress multicast streaming method. However, considering that the coding overhead increases with the number of SVC layers and the storage prices have dropped massively in recently years, streaming multiple independent non-scalable video representation or scalable video that only contains a small number of layers in a DASH-based system becomes a practical and efficient solution for multimedia streaming over the Internet.

The DASH specification only defines the MPD and segment formats to let the DASH client control the streaming session. The delivery of the MPD and the media-encoding formats containing the segments, as well as the client behavior for fetching, adaptation heuristics, and playing content, are however outside of DASH's scope. In fact, multimedia consumers expect a high-quality viewing experience in terms of visual quality, start-up time, responsibility and trick-mode support etc. The question that remains for researchers: *how to make the right segment selection to guarantee the quality of experience for video*

streaming services over Internet, especially over wireless networks, given fluctuations in network dynamics and device conditions? First and foremost, the QoS provided by a DASH client is dynamic as it can be frequently modified over time. For robust and satisfactory streaming services, it is very important to perform investigations that help to measure end-user acceptance, compare alternative adaptation options and find optimal streaming schemas.

2.3 Quality of Experience (QoE)

In the multimedia communities, there are various definitions of the QoE concept. QoE is defined by International Telecommunications Union (2007, 2008b) as *the overall acceptability of a product or service, as perceived subjectively by the end-user.* A more comprehensive definition of QoE is given by Wu et al. (2009) as *"a multi-dimensional construct of perceptions and behaviors of a user, which represents his/her emotional, cognitive, and behavioral responses, both subjective and objective, while using a system".* It shows that the QoE metric can be refined into multiple concrete metrics in order to measure different kinds of cognitive perceptions and the resulting behavioral consequences. For example, user experience of a streaming service can be characterized from different aspects, using words such as "Enjoyment", "Usefulness" and "Easy of use".

A QoE metric can be related to one or more QoS metrics that measure the quantifiable or tunable aspects of service performance in terms of bit-rate, loss and delay etc. As different applications have their own features and therefore put particular emphasis on some QoE metrics, it is essential to understand the influences of related QoS metrics for a well-specified QoE metric. Research activities have been conducted to map the relationships between different quality metrics. For example, Gottron et al. (2009) performed a study to identify the major factors affecting voice communication. Considering objective measurement of QoE metrics does not always provide reliable results due to the difference between individuals' cognitive abilities, Wu et al. (2009) suggested an empirical mapping methodology to correlate QoS metrics with QoE. Furthermore, empirical studies can translate the cognitive abilities of users to the preference or perceptive thresholds on different dimensions of sensory information. Therefore, user studies are often involved in the research of QoE optimization. For example, Huang and Nahrstedt (2012) conducted subjective tests to determine user preferences for different streaming opinions. These subjective preference results were then used to design an adaptive streaming scheme for improving QoE. Xu and Wah (2013) performed subjective tests to find the Just-Noticeable Difference (JND) (Wikipedia, 2015a; Sat and Wah, 2009) of the send-receiver delay on audio signals. Knowing the JND threshold, the delay can then be increased to an extent in order to smooth out network congestions, without incurring noticeable degradation in interactivity.

2.4 Audiovisual quality assessment

For most multimedia streaming services, user-perceived quality of the delivered audio and video content is the major factor that contributes to the QoE. However, measuring user perceived audiovisual quality is difficult, because the perceived quality is inherently

subjective and under the influence of many different factors. The final audiovisual quality encompasses the complete end-to-end system effects (client, terminal, network, service infrastructure, etc) and may be influenced by user perceptions, expectations, attitude and context etc.

Currently, objective quality metrics for video are still in the early stage of development. Most existing objective metrics fail to estimate the subjective experience of a human observer watching a video display, especially for scalable video that can vary its own properties during the streaming sessions. Audiovisual quality assessment relies on carefully controlled subjective tests to capture the subjectiveness associated with human perception and understanding.

2.4.1 Objective quality evaluation

Objective video quality metrics are generally classified into full-reference (FR), no-reference (NR) and reduced-reference (RR) categories according to the availability of the reference video (Martinez-Rach et al., 2006). FR metrics perform a frame-by-frame comparison between a reference video and the test video so that they require the entire reference video to be available. This type of metrics are most suitable for offline video quality evaluation. For example, FR metrics can be used for designing and optimizing video processing algorithms as replacements of subjective tests. NR and RR metrics evaluate the test video with none or limited information from the reference video. They are better suited for in-service quality monitoring where adaptive streaming and transcoding strategies are needed. When RR metrics are used to provide realtime measurement, a back-channel is often required to fetch some information about the reference such as the amount of motion or spatial details.

$$PSNR_{dB} = 10log_{10}\frac{(MAX)^2}{MSE} \tag{2.1a}$$

$$MAX = 2^B - 1 \tag{2.1b}$$

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{2.1c}$$

FR metrics are most intensively studied among the three categories. The Peak Signal to Noise Ratio (PSNR) (Equation 2.1a) metric is measured on a logarithmic scale and depends on the mean squared error (MSE) relative to the square of the maximum possible pixel value of the image (MAX). The equation 2.1b gives the estimation of MAX when the image samples are represented with B bits per sample. The MSE is the sum over all $m \times n$ squared pixel value differences between an original video frame I and its impaired approximation K. PSNR can be quickly and easily calculated and is therefore widely used as a quality metric. However, it does not reflect well the video quality perceived by human observers since it simply performs byte-by-byte comparison of the pixel values without considering their actual meaning and spatial relationship (Winkler and Mohandas, 2008). Huynh-Thu and Ghanbari (2008) have explored the narrow scope in which PSNR is representative for video quality assessment. Instead of comparing single pixel values, Wang et al. (2004) proposed the Structural Similarity (SSIM) index that measures the similarity between images based on the mean, variance and covariance of small patches inside a

image. Although it is claimed that the SSIM index can provide good estimation of the overall reaction of the human visual system, it is not a true perceptual quality metric. Dosselmann and Yang (2011) have shown that SSIM is only a mathematical transformation of MSE (and PSNR). Objective metrics that perform extraction and analysis of content features and artefacts in the video data usually have better performance. Pinson and Wolf (2004) introduced the National Telecommunications and Information Administration General Model (NTIA GM) for combining measures of the perceptual effects of different types of artefacts such as blurring, blockiness, jerkiness, etc. However, despite some reported superiority of the objective model over PSNR, the evaluations performed by Martinez-Rach et al. (2006) and Loke et al. (2006) indicate that NTIA GM does not work well on multimedia video with low bitrates, various frame-rates, and small frames size. A few objective metrics that use psycho-visual models to predict human visual responses were validated by International Telecommunications Union (2008a, 2011). Although these models also significantly outperform PSNR, they are not accurate enough to fully replace subjective testing.

Most NR and RR metrics focus also on measuring artefacts and use the artefact metric to predict perceptual quality score. Most existing NR metrics are based on estimating the most prominent compression artefacts such as blockiness and blurriness (Sheikh et al., 2002; Wang et al., 2002). Without reference, it is hard to accurately quantify a certain artefact since there is a high risk of confusing actual content with artefacts. As a compromise between FR and NR metrics, RR metrics make use of a set of side information which is comprised of important features from the reference and/or test video (Gunawan and Ghanbari, 2008; Callet et al., 2006). To be practical, theses features have usually much lower data rate than the uncompressed video stream. Examples of features are the amount of motion or spatial detail (Winkler and Mohandas, 2008).

The scaling options of H.264-SVC increase the perceptual uncertainty dramatically. It is therefore highly desired to count the quality influence of variations in multi-dimensional adaptation space into an objective quality metric. An objective metric multiplicatively combining the SNR distortion and frame loss was proposed by Wu et al. (2006). It was claimed that SNR scaling worked better than temporal scaling under most circumstances. Kim et al. (2008) proposed a scalability-aware video quality metric (SVQM), which incorporated the spatial resolution together with frame-rate and spatial distortion into a single quality metric. However, none of these objective metrics have considered the temporal variation of different impairments.

To determine the accuracy of an objective quality metric, the calculated objective quality values are usually compared with the evaluation scores from a carefully conducted subjective test. It is expected that a reliable objective quality metric would provide values that are strongly correlated with subjective assessments. Therefore, the performance of the above-mentioned objective metrics PSNR, SSIM, NTIA GM and SVQM have been validated against subjective assessment scores in this thesis (see chapter 4.7). Again, none of these metrics provided satisfying estimation of the perceptual quality of scalable video. As a matter of fact, our work shows negative correlation between subjective studies and all these objective metrics.

2.4.2 Subjective quality evaluation

Audiovisual quality assessment fundamentally relies on subjective quality studies to capture the perceived quality experience of human observers. Such studies are known to be expensive and difficult to administer. They require time, a detailed knowledge of experimental designs and a level of control which can often only be achieved in a laboratory setting. Hence, few research efforts have been dedicated to subjective studies.

There are a few published user studies investigating the quality influence of different scaling options. Cranley et al. (2006) carried out a set of experiments to discover the Optimal Adaptation Trajectory that maximizes the user-perceived quality in the adaptation space defined by frame-rate and spatial resolution. It was shown that a two-dimensional adaptation strategy outperformed one-dimensional adaptation. The effects of fidelity degradation and frame-rate downscaling were evaluated by subjective tests (McCarthy et al., 2004). It was shown that high frame-rate is not always more preferable than high image fidelity for high motion video. According to results of the user studies introduced in this thesis (see chapter 5), the impact of frame-rate loss on perceived video quality is also related to the device being used. The perceived video quality in the course of video adaptation process is harder to estimate due to the potential large quality changes. Zink et al. (2003) performed systematically a series of subjective evaluation tests to investigate quality degradation caused by different variations in the transmitted layers during streaming sessions. One interesting finding of their experiment is that the perceived quality of streamed video is influenced by the amplitude and the frequency of layer switchings. Hence, it is recommended to keep the frequency and amplitude as small as possible. However, more knowledge is required for developing a transparent video adaptation strategy. To handle sharp bandwidth degradation, the streaming service designer needs often to determine threshold quantities of different layer adaptation options for the acceptability of the end QoE. Additionally, only SNR layer variations have been examined in Zink et al.'s experiment. Modern scalable video coding techniques enable layer switching occurring in both spatial and temporal dimensions, which have different impact on perceived video quality. More investigations are needed to explore the entire adaptation space of layer-structured video. To avoid confounding the effect of scaling dimension with other factors, (e.g., frequency and amplitude), quality variations in different dimensions should also be treated separately. Another limitation of Zink et al.'s experiment is that interaction effects between video content characteristics and layer variations have not been investigated although human perception are often strongly influenced by the video content.

Codec performance is critical for decoded video quality. Essentially, all video compression techniques are trade-offs between bit savings and distortions. Subjective evaluation is often used to compare the rate-distortion performances of different video compression algorithms. In these experiments, video compressed by different methods should be compared under the same bit-rate constraints. On the other hand, bit-rate is not the only factor that influences a user's overall impression of viewing a video. Given the characteristics of the human visual system, there are many other factors that may affect the perceived video quality. The source of quality impairment is not only limited to the lossy video compression methods. Subjective evaluation experiments may also have the purpose of factor screening and interaction detection. In these experiments, it is practical to compare videos with the same coding parameters instead of bit-rates. Using the same coding parameter, we assume the fidelity losses caused by video compression are the same

between different videos, thus making the comparison between different content materials possible. Examples of these experiments are given in chapter 6 in this thesis. Although we use H.264-SVC encoded video to simulate the visual artefacts caused by online video adaptation, we believe that the experimental result is also applicable to other coding algorithms such as H.264-AVC due to the use of the same coding parameters and the similar compression procedures of H.264-AVC and H.264-SVC.

2.5 Experimental design for subjective tests

Subjective tests have to be carried out rigorously in order to ensure reliability of the test results. This requires the experimenter to design their experiments according to the scientific method so that appropriate data that can be analyzed by statistical methods will be collected, resulting in valid and objective conclusions (Montgomery, 2006). The design of an experiment is a detailed plan of the entire experimental process. It covers many different issues ranging from the choice of process variables (factors) to the allocation of experimental units (test stimuli) to observation units (subjects[2]). A good experimental design need not only to be robust, but must also be efficient in terms of gathering statistical information under given time and economic constraints.

To have a comprehensive overview of experimental designs, we go through the basic steps of the design of a statistical experiment in this section.

2.5.1 Formulation of statistical hypotheses

The starting point of any experimental design is the definition of research question and statistical hypotheses. The research question is a broad formulation of the problems to be studied. An example is the general question we had in the ROMUS[3] research project:

> "How to scale video streams for the best user experience?".

This question can be further divided into a number of more detailed questions referring to some specific influential factors that may have impact on the visual quality of a video. For example,

- "What is the preference order among the available scaling options in scalable video?"
- "Does the perfect utilization of available bandwidth always improve the user experience?".

The former question is targeting the comparison of different scaling dimensions, while in the latter question, the influential factors are aimed at the adaptation pattern and frequency. To examine the effect of these factors, these questions should be then narrowed to an experimentally manageable hypotheses. Similar to many scientific experiments, the so called "hypothetico-deductive" method (Bech and Zacharov, 2006) is applied in

[2]We use the terms of subject and assessor interchangeably in this thesis.
[3]The ROMUS project was a four years research effort funded by the Norwegian Research Council

the design of most perceptual evaluation tests. According to the hypothetico-deductive method, the experimenters proceed through their investigation by formulating one or several hypotheses in a form that could be falsified by a test of observed data. These hypotheses, also known as "null hypotheses", usually state that there are no treatment effects in the population from which samples of response data are drawn. If the statements are found to be false then it is deduced that there may be some effects not to be ignored and this will lead to refinements of the hypotheses or new research ideas.

Formulating a hypothesis is important for the experiment since the statistical analysis of the response data depends on its form. An experiment is commonly affected by random variation, such as observational errors, random sampling etc. The aim and purpose of data analysis is not only to summarize the features of a sample of responses by some basic descriptive statistics such as mean, median or standard deviation, but also to infer statistically that the findings in the experiment are not obtained by chance. For example, the size of standard deviation must be large enough for a given sample size to demonstrate a significant treatment effect. To make such a statistical inference, a hypothesis must be testable and it is often formulated precisely under the premise of some initial test conditions. For example, a hypothesis related to preference of scaling options can be as below:

Hypothesis:
Jerky playback (low frame-rate) is more annoying than coarse picture fidelity.

Initial condition:
Videos with temporally and spatially downscaled quality have similar bit rates and are compared on the same device in the same environment.

Given the hypothesis above, we have a well-defined experimental task that is to examine whether the Mean Opinion Score (MOS) (International Telecommunications Union, 2006) for the annoyance as reported by the subjects is statistically significantly different or not. By adhering to the initial conditions, interferences from nuisance or confounding variables can be eliminated as much as possible and this is helpful to achieve real experimental progress.

2.5.2 Experimental variables

An experimenter usually defines an experiment to examine the correlation between a number of variables. These variables to be investigated are divided into two groups: dependent and independent variables. The experimenter deliberately changes the independent variables to observe the effects the changes have on the dependent variables.

2.5.2.1 Dependent variables

The dependent variables in subjective tests are the response data provided by the subjects. Subjective audiovisual quality assessment is one specific variant of subjective test. The typical way of measuring the human perception of a visual signal is to ask them to quantify their experience. Dependent on the purpose of the experiment, the question to the subject can be related to the sensorial strength of individual visual attributes of

the test stimulus such as blurriness and fogginess, text readability, blockiness etc, or an overall impression of the stimulus expressed as acceptance, annoyance or preference etc. When asking questions relative to visual sensation, extra attention should be given to verbalization as people may use different vocabularies to describe their sensation of the same attribute. Interviews and trainings are often required in order to elicit a list of words for the subjects to represent an attribute. Comparatively speaking, there is less ambiguity in formulating questions related to an overall impression compared to one's sensation to an individual visual attribute. An overall impression is a single impression that combines the sensory influence of all visual attributes and personal cognitive factors such as mood, expectation, previous experience and so on. Based on the single impression, subjects tell their desire of buying a product, or how happy or annoyed they are. No formal training or selection criteria are required for the subjects.

2.5.2.2 Independent variables

The independent variables are experimental factors controlled by the experimenter to affect the dependent variables. There are usually many factors that can influence the response of subjects. The experimenter has to define the factors that may be of importance to the research question, the ranges over which these factors will be varied, and how many levels of each factor to use. To aid the selection of variables, we classify the experimental factors that may considerably affect humans' evaluation of visual quality into a small number of groups.

Signal fidelity Signal fidelity factors are related to the digitalization, compression and transmission of video source materials. With current multimedia techniques, the quality loss during the course of video processing is irreversible, but the effects of many different kinds of artefacts can often be adjusted by the corresponding parameter settings. For example, the blockiness artefact in a video is highly correlated with the quantized transformation. Another example is the bit allocation scheme. It may be beneficial to allocate more bits to some parts of a video program because video artefacts outside the region of interest could be negligible or less annoying. Much research has been devoted to seeking optimal encoder/network configuration for better user satisfaction and greater robustness of streaming services. These studies should be supported and verified by experiments that take the coding/networking parameters as main factors that affect user perception. Knowing the correlation between the signal-fidelity related factors and user perception is particularly useful for a middleware designer.

Content and usage Content and usage related factors belong to the high level design of a multimedia system and should be taken into consideration earlier during the experiment design procedure. Video content can be classified into different genres such as News, Sports, Cartoon and Commercials. Different video genres demonstrate different characteristics in terms of motion speed, scene complexity, variation etc, which result in different interaction effects with other types of factors. For example, motion jerkiness due to the change of frame-rate in a Cartoon movie may not be as noticeable as in Sports movies. Usage related factors refer to the features and functionalities of a multimedia application. A single-purpose application usually aims at some specific content character-

istics. For example, a video surveillance system may often display crowded natural scenes while a video chatting program typically displays a single object and usually has a predictable region of interest. In addition, users expectation to video quality may vary when they use an application for education instead of entertainment. Their perception can also be influenced by interactive functionalities such as channel switching, quick browsing or remote communication.

Device and environment With ubiquitous video streaming, a variety of end devices and their surrounding environments compose heterogeneous streaming scenarios. The end devices range from large HDTV screens to small smart phones and the environments range from living rooms to aircraft cabins. Obviously, this brings different experiences to the user, and the effects of viewing environment and device are worthy of investigation.

2.5.3 Evaluation method

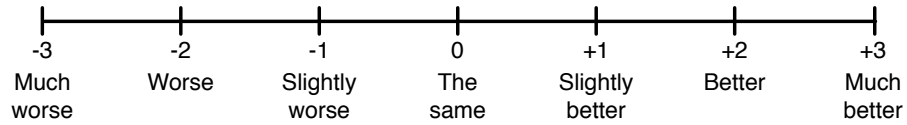
When the influential factors have been identified and questions to the subject have been formulated, the next step is to decide how to measure the responses of interest. The task involves specifying a measurement scale and an evaluation method that the subject can apply when reporting the visual impression.

A good measurement scale should be reliable, easy to use and provided powerful discrimination. It is also highly recommended to apply standardized scales when it is used as general practical measure of human characteristics so that experiments conducted at different time can be comparable. Figure 2.5 shows several examples of standardized scales. These scales can be classified into three categories, namely nominal (figures 2.5a, and 2.5b), ordinal (figure 2.5c) and interval (figures 2.5d, 2.5e, and 2.5f). The selection of scale type depends on the dependent variables in the experiment plan. On the other hand, it determines also to a large extent the type and degree of statistical analysis allowed.

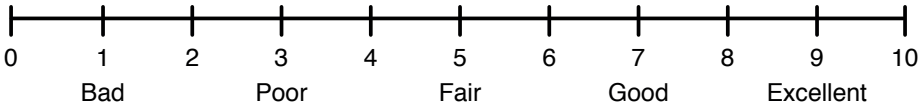
International recommendations such as ITU BT.500-11 (International Telecommunications, 2002) and ITU-T P.910 (International Telecommunications Union, 1999) suggest several evaluation methods for visual quality assessment. In summary, there are three recommended basic forms of evaluation, Degradation Category Rating (DCR), Absolute Category Rating (ACR) and Paired Comparison (PC). Different modified versions of these evaluation methods exist. In the DCR method, subjects evaluate the quality of test stimuli related to their unimpaired source reference while the ACR method presents only the test stimuli and asks subjects to evaluate their absolute quality. Both the DCR and ACR methods are based on direct scaling. Each subject gives a score to grade the perceived quality or perceived impairment in a test stimulus. MOS scores are obtained by averaging the scores from a number of individual subjects for each test stimuli. The PC method compares pairs of test stimuli with each other. In contrast to DCR and ACR, the PC method is an indirect scaling method that asks subjects to evaluate their preference instead of perception. The evaluation results can be MOS scores or frequency tables depending on the levels of measurement. The measurement scale used for comparison can simply contain two points (e.g., “Same”,“Different”) for checking just the existence of differences, or include multiple points as shown in 2.5c for evaluating the direction and degree of the differences.

(a) 5-grade categorical quality scale

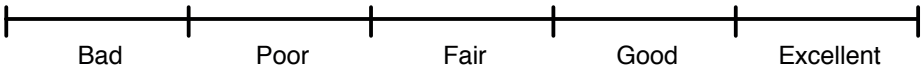
(b) 5-grade categorical impairment scale



(c) 7-grade comparison scale



(d) 11-grade quasi-continuous quality scale



(e) Continuous quality scale

(f) Double-stimulus continuous quality scales

Figure 2.5: Standardized rating scales.

All of the three types of evaluation method have claimed advantages. The fundamental difference between the DCR and ACR methods is the explicit use of a reference, which makes the DCR method most suitable for testing the fidelity of transmission or super quality system evaluation. Relatively, it is easier and quicker to perform subjective tests based on ACR methods. The presentation of the test stimuli is similar to that of the common use of the systems, thus, ACR is well-suited for qualification tests. In addition, continuous quality evaluation is possible via the ACR method. A comparison of the DCR and ACR method by Pinson and S.Wolf (2003) shows that the ACR method can generate quality estimates comparable to DCR methods, but humans consider only the last 9 to 15 seconds of video when forming their quality estimate. When the ACR method is used for continuous quality evaluation, subjects are required to vote at least once every 10 seconds. A handset-voting device should be used to facilitate the rating procedure in this case.

The PC method is a prominent assessment method due to its high discriminatory power (International Telecommunications Union, 1999). In addition, the rating procedure of the PC method is simpler than that of DCR and ACR. Making a comparative judgment based on preference is a commonly encountered practice in our daily life, which makes the PC method suitable for many unsupervised field studies with untrained subjects. Results obtained with the PC experiments are often robust and known to closely reflect perceived sensations on a psychological scale (Thurstone, 1994). Another advantage of the PC methods is that the comparative judgement can be easily verified by examining the transitivity of the ratings, which makes PC methods well-suited for inspecting the agreement between different users. However, the PC method recommended in the standards applies a full-factorial repeated-measures design that includes all possible test conditions in one experimental run (see section 2.5.4). With this design, the number of stimuli pairs in a test session grows exponentially with the number of factors and their levels under investigation. This makes the standardized PC method counterproductive for audiovisual quality studies that are known to contain a large number of both factors and levels.

The standardized evaluation methods are mainly targeting the performance of television systems in environments where the viewing conditions (room luminance, background colour, viewing angle and distance etc) are strictly controlled and more critical than realistic situations. Field experiments which examine the system performance in the real world rather than in the laboratory are gaining the interest of multimedia system designers, especially when mobile devices are commonly used to display digital media. However, running subjective tests in a natural environment adds difficulties to the experimental design. Thus, new thoughts and amendments for the current test methodology are highly demanded.

2.5.4 Statistical experimental designs

The standard recommendations for subjective quality evaluation focus on common aspects of subjective assessment such as stimuli presentation, viewing conditions, measurement scales and basic statistics for data analysis, while the treatment [4] design and allocation are left to the experimenter. However, the treatment related design determines to a great

[4]The term of treatment is referred to a specific level of an experimental factor

extent whether or not an experiment can be performed correctly and efficiently.

Factorial design is an experimental design in which the effects of more than one factor are investigated simultaneously. With this design, the effect of one factor can be estimated at several levels of the other factors, which yields conclusions that are valid over a wide range of experimental conditions. Due to the efficiency, factorial design is most commonly used for planning audiovisual perception evaluation that involves a large number of experimental factors.

The simplest factorial experiment is the *full factorial design* (Montgomery, 2006; Kirk, 1982) based on *repeated measures* (Coolican, 2004; Bech and Zacharov, 2006). In such a design, all possible treatment combinations are investigated in each complete run of the experiment. Observations of the various treatments are made in random order. The full factorial design is an optimal design from the standpoint of data analysis, as all treatments are assigned randomly to all subjects and all experimental factors can be analyzed statistically. The major disadvantage of this design is that a large number of treatment combinations need to be tested in the same experiment run, which makes the experiment too long to accomplish for most of subjects.

To stay within time and resource limits, a *fractional factorial design* or *blocking* strategies (Lewis and Tuck, 1985; Bech and Zacharov, 2006) may be used, which reduces the number of treatment combinations systematically. These approaches reduce the statistical information that can be extracted because relations between treatments in one or the other partition cannot be determined, but they are repeated-measures designs and retain all other benefits of these approaches. The opposite approach to reduction is taken by the *independent-measures design* (also known as between-subjects design) (Coolican, 2004; Bech and Zacharov, 2006). Here, each subject is tested under one and only one condition. This implies large numbers of subjects and complex management, and can lead to very long experimental periods for the experimenter due to the overhead of interacting with a huge number of subjects. For many research projects, the time and resources required can make between-subjects design impractical. Designs that fall neither in the repeated-measures nor the independent-measures category are not typical. Their drawback is that they cannot classify and group individuals like repeated-measures designs, and do not have the protection from misbehaving individuals of independent-measures designs. They avoid, however, the extensive time required for subjects of repeated-measures designs and for experimenters of independent-measures designs.

2.5.5 Data processing

Once a test has been carried out, the reliability of subjects should be checked from the collected rating scores. Outliers whose ratings appear to deviate significantly from the rest of subjects are usually excluded from the data sample. Then, according to the type of rating scores (categorical, ordinal or interval), appropriate statistical analysis of the test results should be reported. The MOS and the standard deviation of the statistical distribution are commonly used methods of analysis and are often presented in graphics to summarize the performance of the system under test. When we suspect that we have detected a real effect, significance tests are used to help us decide whether or not rejecting the null hypothesis.

2.6 Summary

Streaming audiovisual content to heterogeneous environments needs to apply adaptable techniques and policies to handle dynamic resource availability. The latest developments in video coding techniques and streaming protocols have provided a great variety of adaptation options for designing adaptive streaming services. For instance, the H.264-AVC video coding standard, which is widely accepted by industry, has included the amendment for scalable video coding (H.264-SVC), allowing bitstream scalability in multiple dimensions. As a more general streaming solution, MPEG-DASH enables HTTP-based adaptive streaming with even non-scalable videos. To provide a highly qualified multimedia experience, it has become increasingly important to devise audiovisual quality/distortion measurements that help to measure end-user acceptance, compare different adaptation options and find optimal configurations. However, audiovisual quality is a subjective measure, and none of the current objective quality metrics have the required accuracy for monitoring the overall quality in video communication. Audiovisual quality assessment still relies on subjective quality studies, which are known to be expensive, time consuming and hard to administrate.

In this thesis, we both examine subjective quality assessment procedures and start looking at gaining knowledge of how to perform useful adaptation. In chapter 3, we therefore present an approach based on the PC method to facilitate subjective quality studies in the field. Then, we introduce several user studies that investigate the quality impacts associated with different types of video adaptation options in chapters 4, 5, and 6.

Chapter 3

Randomized Pairwise Comparison

Subjective quality perception studies with human observers are essential for multimedia system design, because objective metrics are still in the early stage of development. Along with the widespread use of mobile devices and ubiquitous multimedia application, it is much useful to test realistic multimedia experiences directly and frequently. This requires a new method for quality assessment in the field.

3.1 Introduction

A field study is often desired for the evaluation of quality of experience (QoE), because it captures people's true experience in a real world setting, without artificially creating the environmental context. This is especially useful for evaluating multimedia experiences on portable devices, e.g., situations where one streams video to mobile devices located in public places with complex surroundings that are hard to simulate in laboratories. Typical designs for subjective studies with non-expert subjects are the full-factorial pairwise comparison and a variety of randomized block designs (Lewis and Tuck, 1985). However, compared to laboratory studies, field studies are more difficult to design, and the classical designs become highly impractical. In the field, extraneous variables such as noise, interruption and distraction may not be under the control of the experimenter, which threatens the validity of the experimental findings, although it is perfectly representative for real-world applications. Moreover, additional problems exist in field studies related to human perception. Screen-based tasks are especially susceptible to fatigue effects, even for durations as short as 15 minutes (Chi and Lin, 1998). When assessing video quality in the field, assessors can easily become tired, bored and uncooperative, or just run out of available time. Their responses will therefore be increasingly unreliable, leading to greater unexplained variance. Furthermore, audiovisual quality studies are known to contain a large number of factors and levels to be investigated. The number of all possible test stimuli grows exponentially with the number of factors and factor levels under investigation. It is therefore often impractical for the assessors to evaluate all test stimuli of a field experiment due to time and location constraints.

This chapter extends my co-authored paper "Randomised Pair Comparison: An Economic and Robust Method for Audiovisual Quality Assessment"(Eichhorn et al., 2010), included in Appendix B, with more studies and conclusions.

To resolve the above problems, we introduce *Randomised Pairwise Comparison* (R/PC) (Eichhorn et al., 2010) as a convenient and practical method for multimedia QoE evaluations in field. The method is designed with realistic assumptions of the time and effort that an assessor will have to spend. As in conventional pairwise comparison methods, a paired preference test is used as the basic form of evaluation because it involves a simple cognitive task, comparing two stimuli in a pair against each other. The assessors are only required to make a choice based on their own preference. No formal training is needed, which makes it possible to recruit random people with different background and age. The novelty of R/PC is that, in contrast to conventional pairwise comparison methods that require that every assessor provides responses to all possible paired stimuli or a pre-defined number of paired stimuli that is identical for all assessors, R/PC selects small subsets of pairs randomly and creates a unique experiment session for each assessor. Assessors are therefore not tied up with the entire experiment, which provides the experimenter a flexible way to administrate large studies with many test stimuli. This feature is especially useful for observed field studies and self-controlled web-based studies that are often conducted among volunteers who participate in the tests during their leisure time. To facilitate field studies on mobile devices, we prototyped a software tool on the iOS platform to automate the R/PC experimental design.

In R/PC, the evaluation tasks are shared by assessors. There is a trade-off between the number of required assessors and the time every assessor needs to contribute. To get enough response data, R/PC requires more assessors than other repeated-measures designs. In addition, as the responses to different test conditions are from different groups of people, individual variability and assignment bias are the most suspected sources of nuisance factors that may skew the data and make the R/PC experiments unstable. To verify that the R/PC method is robust and practical, we examine the performance of R/PC method by experiments and simulations.

3.2 Method design

We designed R/PC with realistic expectations about the time assessors are willing to spend in a study and practical assumptions about the ability of experimenters to control environmental and content-related factors. In contrast to the standardized PC method, R/PC does not request assessors to evaluate *all* pair stimuli. It allocates subsets of stimuli to assessors randomly. By this, session duration is separated from factorial complexity of an experiment, and an experimenter can balance between experiment costs and the information obtained. A session can have an arbitrary duration (down to a single pair) and assessors can quit their session anytime, e.g., when they get distracted by phone calls or have to exit a bus or train. With the aid of software-based randomization, multiple factors can be investigated together in a flexible and economic manner. This is particularly efficient for the exploratory quality assessment of multimedia experience.

3.2.1 Independent randomized design

In a typical factorial experiment using the PC method, any two clips in a pair may differ in one or multiple factors, as defined by the experimenter. We denote such pairs *contrast pairs*. They are used for actual exploration and hypothesis testing. An experimenter may,

for example, base his research hypothesis on assumptions about the visibility and effect size of contrasting factors. An experimenter should first identify factors, which will be controlled in the study and the number of levels for each factor. Factors can be discrete or continuous and the number of levels may differ between factors.

Pair stimuli are created for all treatment combinations in R/PC. However, R/PC creates only a short list including an *unique random subset of pairs* for each assessor and then randomizes the presentation order for these pairs. The size of the subsets is decided by the experimenter according to the preferred session duration in the target test scenarios. We restrict all pairs in R/PC experiment to equal duration at 8∼10 seconds with the consideration of human memory effects, the subset size is thus given by $s = d_s/d_p$, where d_s is an estimation of the length of time that most of the assessors can spend on the test without tiredness and d_p is the duration of a pair presentation including an estimated time for voting.

3.2.2 Quality measurement

R/PC presents test stimuli as pairs of clips. Each clip in a pair is introduced by a 2 seconds long announcement of the clip name and the letter A or B, displayed as a 50% grey image with black text. This results in the time pattern as shown in figure 3.1.

After the presentation of one pair, an assessor is expected to *compare the overall quality* of the presented pairs and to report their preference using a self-report scale. We do not impose any time limit that would force a decision for the rating. Instead, we propose to measure the time it takes an assessor to respond and use this in later data analysis. The session continues with the next pair immediately after the response has been entered. With the aid of software-based random selection, the order of pair presentation is completely randomized.

R/PC experiments do not include training sessions. A brief introduction about the purpose of their study and the scale to be used may be given at the beginning of a test session. Assessors can also be reminded to pay close attention, but an experimenter should avoid specific restrictions or hints which might guide an assessor's focus.

Assessors report their preference on one of the following comparison scales:

- a *binary preference scale* which allows to express either a preference for clip A or a preference for clip B (Equal reference pairs, as defined in section 3.2.3, are not used together with this scale.)
- a *3-point Likert scale*, which contains a neutral element in addition to a preference for A or B (promotes indecisiveness, but more suitable for the detection of just noticeable difference (JND) thresholds (Wikipedia, 2015a; Sat and Wah, 2009))
- a *4-point Likert scale* which provides items for weak and strong preference for either clip, but lacks a neutral element (has a higher resolution than the binary scale and forces preference selection, excluding equal reference pair as well)
- a *5-point Likert scale*, which contains a neutral element as well as items to express weak and strong preference (high resolution, may promote indecisiveness, but fits better for the detection of JND)

Figure 3.1: Presentation pattern for a single clip pair.

3.2.3 Consistency Checking

Field studies using randomly selected subjects encounter often erroneous or fraudulent user responses. To filter out unreliable assessors, R/PC introduces some *reference conditions* into the test stimuli set.

R/PC uses two types of reference conditions: (1) *equal reference pairs* that pair up each single test sequence with itself, and (2) *matched contrast pairs* that just differ in the presentation order of all the existing contrast pairs. The equal reference pairs can also help to understand the perceptual limits of individual assessors. In R/PC, the reference conditions are randomly distributed and hidden among contrast pairs when presented to an assessor to avoid their detection. The random pair allocation procedure ensures that (1) each selected contrast pair is contained in both possible presentation orders (both matched contrast pairs AB and BA are present), and (2) equal reference pairs are selected according to treatment options in previously selected contrast pairs.

By packing contrast pairs together with their corresponding reference pairs, the generated test session possesses self-checking capability whilst remaining independent and flexible. R/PC checks only the consistency of input scores from each individual assessor. Errors are counted if the preference scores for a contrast pair and its matched contrast pair have the same direction, i.e., if the same score is given to paired treatments in both AB and BA order; or an equal reference pair receives non-neutral or extremely prejudiced score, dependent on the selected comparison scale. The error rate for each assessor is calculated according to the counted number of error scores and the size of the playlist. An assessor is rejected if his/her error rate exceeds a predefined threshold.

3.2.4 Data Analysis

The nature of the binomial and Likert scales suggests non-parametric statistics for data analysis. The table 3.1 shows some commonly used non-parametric statistical methods. The choice of statistical test tool is dependent on the type of significance tests, the number and the nature of experimental variables etc. The Binomial tests or χ^2 tests are useful for majority analysis on frequency tables (to check whether a majority of preference ratings for one factor-level is significant). As non-parametric counterparts to t-tests and ANOVA that test differences between two or more conditions, the Mann-Whitney U, Kruskal-Wallis and Friedman tests (Coolican, 2004) exist. Rank-order analysis for finding total transitive preference orders between pairs are provided by the zeta method (International Telecommunications Union, 1990) and Thurstone's law of comparative judgement (Thurstone, 1994). For more thorough investigations on the underlying structure of the data and to find a linear combination of variables that explains how factors contribute to effects, an exploratory factor analysis using generalised linear models or logit models should

Methods	Type of relationship tested	Data type	Measure
Binomial	Difference between two conditions	Binary / Categorical	proportions
χ^2	Association between two variables	Nominal data / Categorical	proportions
Mann-Whitney U	Difference between two conditions	Ordinal data	medians
Kruskal-Wallis	Difference between three or more conditions	Ordinal data	medians
Friedman	Difference between two or more conditions	Ordinal data	medians

Table 3.1: Non-parametric statistical methods for data analysis

be considered.

The random pair allocation in R/PC leads to unbalanced data. Without balance, common statistical tools like ANOVA or GLMs become unstable (Shaw and Mitchel-Olds, 1993). Thus, although assessors repeatedly give responses to multiple test stimuli in an R/PC experiment, for analysis we regard these response data as independent. Subject-specific data such as age, profession, culture background etc are not involved in the analysis.

3.3 Method Validation

To validate the usability and reliability of our R/PC method, we performed a simple quality assessment study. The purpose of the study was to obtain two data sets, one with a conventional pairwise comparison method based on a full factorial experiment design (F/PC) and a second data set with R/PC. A correlational analysis of the two data sets shows that R/PC can provide accurate experimental results comparable to the F/PC design. To prove that the strong correlation between R/PC and F/PC data is not a coincidence, we implement a software simulator to generate data with distribution similar to data samples from realistic R/PC experiments. Based on the simulated R/PC samples, we quantify, in a heuristic study, the performance difference between full factorial and randomized pairwise comparison.

We first explain the design of our example study and present an initial analysis of our findings afterwards. Then, we introduce the implementation of the simulator and report our analysis of the average performance of R/PC experiments based on simulation.

3.3.1 Example Study

As a simple example of an audiovisual study, we examined the visibility of different video quality reductions in relation to already existing impairments. Our quality impairments originate in a loss of image fidelity between five different operation points, which have been created using different fixed quantization parameters (QP) for encoding.

In this study, we focus on three factors that are assumed to have *main effects*, namely the original quality level, the amplitude of a quality change and the content type. These factors can mutually influence the user's perception. For example, the same amplitude in quality changes may be perceived differently depending on the original quality and

The effect of a factor on a dependent variable averaging across the levels of any other factors.

(a) Animation

(b) Cartoon

(c) Documentary

(d) Movie

(e) News

(f) Sports

Figure 3.2: Selected Video Sequences.

Genre	Content	Detail	Motion	Audio	Thumbnail
Animation	BigBuckBunny HD	3.65	1.83	sound	
Cartoon	South Park HD	2.75	0.90	speech	
Documentary	Earth HD	3.64	1.61	sound	
Short Movie	Dunkler See	1.85	0.58	sound	
News	BBC Newsnight	2.92	0.69	speech	
Sports	Free Ride	3.32	1.90	music	

Table 3.2: Selected sequences and their properties. Detail is the average of MPEG-7 edge histogram values over all frames (Park et al., 2000) and Motion is the MPEG-7 Motion Activity (Jeannin and Divakaran, 2001), i.e., the standard deviation of all motion vector magnitudes.

the direction of the quality change. Interactions may also exist between the change's amplitude and the content.

To test different kinds of content with varying detail and motion, we selected video materials from six different genres (see table 3.2 and figure 3.2). The length of a test sequence was restricted to 8 seconds (200 frames), taking human memory effects into consideration (International Telecommunications Union, 1999). To maintain as far as possible similar content characteristics throughout a video sequence, all video clips were extracted from one single video shot and scene cuts were filtered out deliberately. All clips were downscaled and eventually cropped from their original resolution to 480x320 pixel in order to fit the screen size of our display devices. We used x264 to encode the original clip in constant quantiser mode so that the same amount of signal distortion was added to all frames in a test sequence. Since the visibility of quality impairments is not linearly related to the size of QP, we selected a set of five QPs with logarithmically distributed values. In a pilot study, the corresponding QPs (10, 25, 34, 38, and 41) have been verified to yield perceptual differences. Using the pairwise comparison, quality change can be presented as a pair of different quality levels. Thus, we do not specifically generate test sequences for each amplitude level. The experimental design of this study is simplified to two factors with variations at 5∼6 levels. With five quality levels we can create $\binom{5}{2} = 10$ unique combinations of contrast pairs that have quality change amplitudes between 1 to 4 and five equal reference pairs per content type. In total, we created 120 contrast pairs in both orders and 30 (25%) equal reference pairs. In our example study, an F/PC test session lasted for 60 min while an R/PC test session lasted for only 12 min.

The clip pairs were displayed to assessors on an iPod touch, which has a 3.5-inch widescreen display and 480x320 pixel resolution at 163 pixels per inch. Display and voting were performed on the same device using a custom quality assessment application. The experiment was carried out in a test room at Oslo university. Overall, 49 participants (45%

a) Full factorial Pairwise Comparison	
Unique Subjects:	34
Unique Pairs:	150
Unique Responses:	5100
Resp/Subj (min/mean/max):	150 / 150 / 150
Resp/Pair (min/mean/max):	34 / 34 / 34

b) Randomised Pairwise Comparison	
Unique Subjects:	49
Unique Pairs:	150
Unique Responses:	1470
Resp/Subj (min/mean/max):	30 / 30 / 30
Resp/Pair (min/mean/max):	4 / 9.8 / 19

Table 3.3: Grand totals and statistics for the two data sets in our example study.

female) at an age between 19 and 39 performed the experiment. Among the participants, 34 people (50% female) were paid assessors who performed both the F/PC test and R/PC test while 15 participants (40% female) were volunteers who performed only the R/PC test. Half of the participants who did both tests, performed the R/PC method first, while the other half did the F/PC test first. During all test sessions, the participants were free to choose a comfortable watching position and to adjust the watching distance. They were also free to decide when and for how long they needed a break.

Based on the two data sets we gathered using F/PC and R/PC, we did some initial comparative analysis. We were interested whether an investigator using different statistical procedures on either data set would be able to find similar results. Hence, we first looked at the correlation between both data sets and second, we tried to fit a linear model to the data in order to find factors which influence main effects.

For the correlation analysis, we first calculated the arithmetic mean and the median of all responses per pair. Then, we calculated Pearson, Spearman and Kendall correlation coefficients as displayed in table 3.4. In statistics, the Pearson coefficient is a measure of the degree of linear dependence between two variables, while the Spearman and Kendall coefficients are nonparametric measures of rank correlation between two variables, i.e., the similarity of the orderings of the data when ranked by each of the quantities. The value of these coefficients are always within the range $[-1, 1]$, where 1 and -1 indicate total positive and negative correlation, and 0 shows that the two variables are totally independent. In our experiment, we found that all the three correlation coefficients were significantly different from zero with $p < 0.01$.

Despite the fact that responses in the R/PC data set are very unbalanced (min = 4, max = 19 responses for some pairs, see table 3.3) and that the total unique responses collected with our R/PC method are only $< 1/3$ of the total F/PC responses, there is still a very strong correlation between both data sets. This supports the assumption that random pair selection may become a useful and robust alternative to full factorial designs for audiovisual quality assessment.

Secondly, we compared the results of fitting a generalized linear model (GLM) to both data sets. We used a binomial distribution with a logit link function and modeled the

Metric	CC	SROCC	$\mathcal{T}$
mean	0.974	0.970	0.857
median	0.961	0.965	0.931

Table 3.4: Correlation between R/PC and F/PC data sets. CC - Pearson Product-Moment Correlation Coefficient, SROCC - Spearman Rank-Order Correlation Coefficient, $\mathcal{T}$ - Kendall's Rank Correlation Coefficient.

main effects of original quality level (Q-max), amplitude of quality change (Q-diff) and content type (content), but no interaction effects. As table 3.5 shows, all main effects are significant, although the significance is lower in the R/PC case, which was to be expected. Again, it is plausible to argue for a sufficient reliability of the R/PC method.

	Factor	Df	Dev	R.Df	R.Dev	P($> \chi^2$)
F/PC	Q-diff	4	718.43	5095	3505.5	$< 2.2e-16$
	Q-max	4	54.31	5091	3451.2	4.525e-11
	content	5	34.39	5086	3416.8	1.995e-06
R/PC	Q-diff	4	236.18	1465	1085.2	$< 2.2e-16$
	Q-max	4	20.48	1461	1064.8	0.0004007
	content	5	16.94	1456	1047.8	0.0046084

Table 3.5: Deviance analysis for a simple GLM considering main factor effects.

3.3.2 Simulation

From the example study, we obtained the first evidence of the reliability of R/PC, showing an R/PC experiment that generated data significantly correlated with data from a full factorial experiment based on repeated pairwise comparison (F/PC). To prove that the strong correlation between R/PC and F/PC data is not a coincidence, more correlational studies are needed. Additionally, it is highly desired to find the minimal time and resource requirements of R/PC in order to know how large experiments one must design.

To examine the general performance of the R/PC method, we built a simulator using the R language to a simulate test procedures using R/PC. The simulator takes a complete data matrix as input and draws R/PC data samples from it according to the random pair selection rules specified in the R/PC method design (3.2.1, 3.2.3). An example of R/PC data sample is given in table 3.6. Simulations of R/PC data samples are made for various session durations, which correspond to D_n, the number of randomly selected responses from subject S_n. For simplicity, we assume that each subject spends an equal amount of time on an R/PC experiment, thus $D_1 = D_2 = ... = D_n$ is set in our simulations. A treatment C_m is a specific test condition referring to a paired test stimuli. To measure the effect of a treatment, statistic y_m is calculated based on all the available responses to that treatment. Depending on the chosen data analytical tool, y_m can be a descriptive statistic such as mean, median, or a test statistic such as frequency distributions etc.

	Treatments					
Subject	C_1	C_2	C_3	...	C_m	**Duration**
S_1	x		x	...	x	D_1
S_2	x	x		...	x	D_2
...	...	...	...	...	...	...
S_n	x	x	x	...		D_n
	y_1	y_2	y_3	...	y_m	

Table 3.6: A R/PC data sample, x - rating available, y_m - statistic under a treatment condition, D_n - the number of ratings provided by subject S_n.

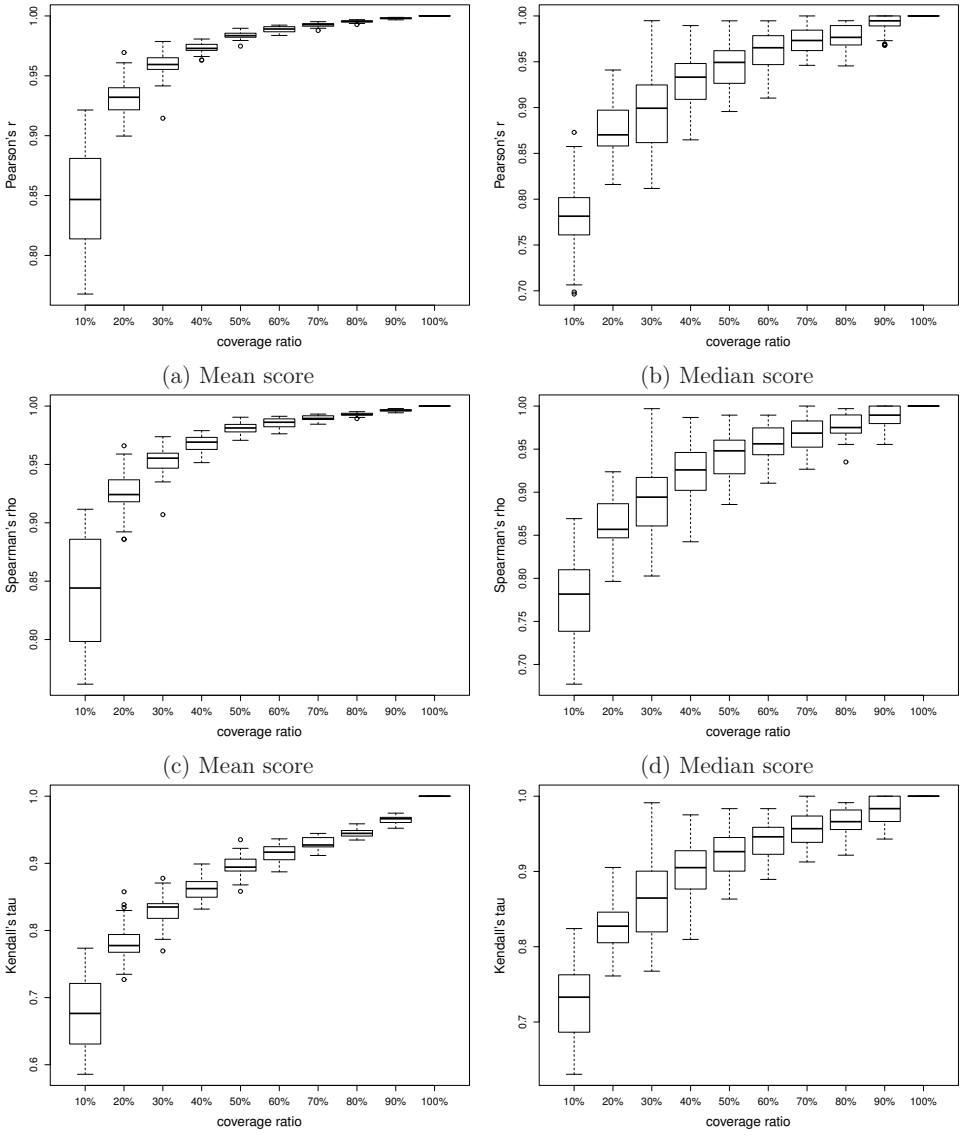
To simulate real life conditions, our simulations are based on real experimental data. The F/PC data set from the example study introduced in section 3.3.1 is used as the input data of our simulator. It is also used as the criterion for testing the average performance of the simulated R/PC data samples.

3.3.3 Reliability

We examine the reliability of R/PC method by comparing simulated R/PC data samples with the F/PC data. A simulated R/PC data sample is a random subset of a complete F/PC data set and the parameter D determines its coverage ratio, which also reflects the experimental time contributed by each subject (session duration). The random pair allocation in R/PC method may create different variations in an R/PC sample. It is doubtful that large variations may exists between different R/PC samples and threaten the validity of the experimental findings of R/PC method. Thus, we did some correlation analysis to check how close an simulated R/PC data sample is related to its complete variant.

We first looked at the correlations between each R/PC sample and the F/PC data set from our earlier experiment. Pearson, Spearman and Kendall correlation coefficients are calculated based on the arithmetic mean and the median value of all responses per treatment. We illustrate these relationships in figure 3.3 and report the average coefficients associated with R/PC data sampled at different coverage ratios in table 3.7. As expected, positive correlations exists between all R/PC data samples and the F/PC data set and the strength of correlation increases along with the coverage ratio. Despite the fact that responses in the R/PC data samples are randomly distributed and unbalanced, very strong correlations (> 0.9) can already be found for R/PC data samples with 20% coverage ratio. This indicates that it is possible to predict the variations in the F/PC data set with an R/PC data sample with more than 20% coverage ratio.

In the next step, we take a closer look at the significance test of the R/PC data samples. A binomial test is used to analyze this data. For each sample, we compare the test results with the findings based on F/PC. Table 3.8 summarizes the number of treatment effects that have been found significant by either R/PC simulation or the original experiment. Figure 3.4 shows the hit rates of effect findings. It is clear that higher coverage ratio helps to find more significant effects. Further, more than 50% effects can be detected successfully by R/PC data samples with only 20%∼30% coverage ratio (each



(a) Mean score

(b) Median score

(c) Mean score

(d) Median score

(e) Mean score

(f) Median score

Figure 3.3: Correlations between the R/PC and F/PC data sets

a) Mean scores per treatment

	Coverage ratio									
Correlations	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
CC	0.844	0.931	0.958	0.973	0.984	0.989	0.993	0.995	0.998	1
SROCC	0.840	0.925	0.953	0.968	0.981	0.986	0.990	0.993	0.996	1
$\mathcal{T}$	0.679	0.782	0.830	0.863	0.896	0.915	0.930	0.945	0.965	1

b) Median scores per treatment

	Coverage ratio									
Correlations	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
CC	0.782	0.872	0.897	0.930	0.945	0.960	0.973	0.978	0.990	1
SROCC	0.778	0.863	0.890	0.922	0.942	0.956	0.967	0.977	0.988	1
$\mathcal{T}$	0.727	0.827	0.861	0.902	0.924	0.942	0.957	0.967	0.981	1

Table 3.7: Correlations between R/PC and F/PC data sets (averages over 30 samples for each coverage ratio), all coefficients were significant below the $p < 0.01$ level. The coefficients indicate the strength of a correlation ranging from -1 through 0 to +1. CC - Pearson Product-Moment Correlation Coefficient, SROCC - Spearman Rank-Order Correlation Coefficient,$\mathcal{T}$ - Kendall's Rank Correlation Coefficient.

subject spends 12∼18 minutes), meanwhile the number of wrong findings (Type I error) is in general very low (<1) for all R/PC data samples. However, the type I errors are not reduced by the use of a higher confidence level at 0.99 for coverage ratios larger than 40%, as shown in 3.8. We believe this reflects the disturbance of random errors. In audiovisual quality assessment tests, each subject is an observation unit and individual variation in perception of quality is the major source of random errors. R/PC relies on randomization to smooth out the individual variations. With complete randomization, each subject covers one and only one random test condition to exclude any related factor, which makes the experimental design a *between-subjects design* that is often accompanied by a large number of participants and high administrative cost. On the other hand, large coverage ratios restrict the randomization, which may result in biased treatment allocation. Thus, to have a feasible experimental design and sufficiently low random error, a small coverage ratio should be considered for an R/PC design.

3.3.4 Practicality

So far, the R/PC as well as the F/PC data are obtained from the same number of subjects. Due to the incompleteness of the R/PC data, the power of R/PC-based testing is lower than that of F/PC for the same size of subject group. This results in failures to find some effects that are actually there (Type II error). As an example, see the number of missed treatment effects in table 3.8. Increasing the sample size is one solution to guard against such an error. The larger a size of an R/PC sample, the greater is its power to detect all potential effects. However, specifying the minimum required subject number is a tricky question. In this section, we try to estimate by simulation the required number of subjects for detecting all significant treatment effects found in a real F/PC experiment.

a) 95% Confidence level

	Coverage ratio									
Count	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Hit	6.1	21.2	28.3	32.6	35.5	36.4	37.9	38.8	39.3	41
Neglect	34.9	19.8	12.7	8.4	5.5	4.6	3.1	2.2	1.7	0
Error	0.17	0.27	0.17	0.27	0.13	0.17	0.03	0	0	0

b) 99% Confidence level

	Coverage ratio									
Count	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Hit	1.1	9.0	19.3	25.0	29.4	32.3	34.3	34.9	35.5	36
Neglect	34.9	27.0	16.7	11.0	6.6	3.7	1.7	1.1	0.5	0
Error	0	0.13	0.07	0.37	0.37	0.17	0.6	0.8	0.67	0

Table 3.8: Comparison between the significance test results of R/PC and F/PC data sets. Hits - treatments show significant effects in both R/PC and F/PC data samples, Neglect - treatments show significant effects only in the F/PC data sample, Error - treatments show significant effects only in the R/PC data samples, all the counted numbers are averaged over 30 R/PC samples for each coverage ratio).

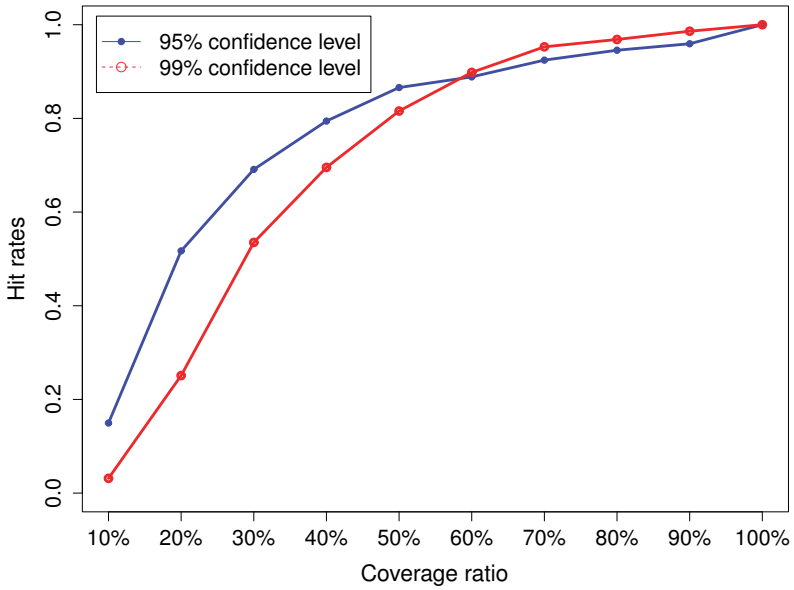


Figure 3.4: Hit rates of significance test.

	Session duration per assessor									
Confidence level	6m	12m	18m	24m	30m	36m	42m	48m	54m	60m
95%	116	74	56	47	41	38	36	35	33	31
99%	137	78	60	49	41	37	34	31	31	31

Table 3.9: Estimated number of subjects required by R/PC experimental design.

To simulate R/PC data with larger sample size, scores from artificial subjects are appended to the F/PC data set. The scores of these artificial subjects are randomly selected scaling options but we assure the original (F/PC) data distribution remains in the expanded data set. R/PC data samples are then generated by randomly selecting a small number of scores from each subject. The minimum required number of subjects is estimated by iteratively increasing the number of subjects until all target effects can be detected with the R/PC data samples. In table 3.9, we report our estimation. For the same experiment, a between-subject design will require $60 * 10 = 600$ subjects to obtain at least 10 repeated observation for each treatment condition.

3.4 Conclusion

Field studies have the advantage of discovering user needs in complex environments. However, successfully conducting an audiovisual field experiment is demanding. In this chapter, we compared our R/PC approach, which is both a practical and economic assessment method for the design of field experiments, with the classical F/PC approach. R/PC is designed as a simple preference test, thus it can easily be used to test people with different background. Reasonable user screening is included in R/PC, which guards it from spurious user input. More important, R/PC provides the possibility of investigating numerous factors, while maintaining the freedom of both experimenters and assessors. An R/PC-based experiment that involved 72 different factor level combinations and recruited response data from random passengers or passers-by in real life locations such as bus terminus, can be found in chapter 5. Compared to F/PC, R/PC requires much less time from the sum of all subjects, and each subject contributes only a small amount of time. This makes R/PC a feasible and low cost design option for experimenters.

R/PC is an adjustable design, which uses randomization to eliminate bias from subject selection and treatment allocation. Although it generates unbalanced data, our simulation results still provide evidence that the robustness of R/PC is comparable to other repeated-measures designs. With R/PC, an experimenter adjusts the test session duration to balance between the costs and accuracy of an experiment. Small session duration helps to lower the risk of confounding , but requires more subjects to prevent type II errors. According to our validation, session durations of 12∼18 minutes that cover 20%∼30% of all possible pairs achieve the best balance between the cost and accuracy of an experiment. Given the small number of evaluation tasks and reasonable number of subjects needed, we have reasons to believe that R/PC provides a feasible and easy design option

A confounding variable is an extraneous variable that is statistically related to the independent variable, which leads to an error in the interpretation of what may be an accurate measurement.

for many field studies. In addition, due to the larger number of subjects, R/PC has higher population validity than repeated F/PC designs. However, R/PC was developed mainly for investigating the influence of non-personal factors on audiovisual quality. Subject-specific data (age, sex, sociocultural background, capabilities, etc.) is ignored by R/PC. If a study is aimed at a specific group of subjects, assessors should be selected beforehand to be representative for that group. Finally, the explanatory power of R/PC suffers from the requirement to use short clips to avoid memory effects. The future development of new test methods that are suited for longer durations without increase in memory effects and fatigue is desirable.

Chapter 4

Pick your layers wisely

In this chapter, we will present our first study of video quality assessment. By this study, we investigated how the subjective video quality is influenced by the multi-dimensional scalability as defined in H.264 Scalable Video Coding (H.264-SVC).

4.1 Introduction

H.264-SVC defines multi-dimensional scalability for efficient and network-friendly operations. In a multi-dimensional adaptation space, it is also possible to find adaptation strategies that provide better user-perceived quality than one-dimensional adaptation space (Cranley et al., 2006). However, the scaling options of H.264-SVC increase the perceptual uncertainty dramatically. SVC techniques inherently rely on objective video quality metrics (VQM) (Winkler and Mohandas, 2008) for optimal performance, but current objective VQMs fail to estimate human perception at low frame-rates or in mobile environments (Loke et al., 2006; Martinez-Rach et al., 2006). Therefore, in order to understand human quality perception of H.264-SVC scalability, we performed a subjective field study (Eichhorn and Ni, 2009) with a special focus on mobile devices. Our goals were to (1) identify when quality degradations become noticeable, (2) find optimal adaptation paths along multiple scaling dimensions and (3) examine whether objective VQMs can predict subjective observations with reasonable accuracy.

To our knowledge, our study was the first study that investigated the subjective performance of multi-dimensional scalability features in H.264-SVC. Due to the lack of encoders capable of full scalability, previous studies could not investigate the influence of spatial, temporal and SNR scaling on quality perception in the same experiment. Although codec performance is critical for decoded video quality, SVC performance was only measured using PSNR metric by Wien et al. (2007). Additionally, many existing subjective tests (Cranley et al., 2006; McCarthy et al., 2004; Zink et al., 2003) were conducted on desktop monitors in a controlled laboratory environment. This differs from our testing scenario defined for mobile video applications.

Our results reveal that adaptation decisions for SVC bitstreams should not only be based on bit-rate and layer dependency information alone. We found that quality degradation may be non-monotonic to bit-rate reduction and that preferred adaptation paths depend on content and user expectations. Confirming previous studies, we also found that common objective VQM like Peak signal-to-noise ratio (PSNR) and structural sim-

ilarity (SSIM) index fail for scalable content and even scalability-aware models perform poor. Our results are supposed to help improving the design of objective quality models towards multi-dimensional video scalability. Enhanced objective models will be useful for several applications and network-level mechanisms, such as bandwidth allocation for wireless broadcasting networks, streaming servers, packet scheduling, unequal error protection and packet classification schemes and quality monitoring.

In this chapter, we present this study in more details than our original paper (Eichhorn and Ni, 2009) included in Appendix C.

4.2 Experimental Design

In this study, we formulated two research questions with regard to the rate distortion performance of the standardized SVC coding algorithm:

- Can we predict the visual quality of a scalable video solely by its bit-rate?
- If not, how does a human perceive the multi-dimensional scalability features in H.264-SVC?

To investigate the general subjective performance of H.264-SVC, we applied SVC to different types of video content in order to find any difference in human quality perception of H.264-SVC scalability. As the first subjective evaluation of H.264-SVC, this experiment restricts itself to on-demand and broadcast delivery of pre-encoded content at bit-rates offered by the available wireless networks in 2008. We focus only on static relations between SVC scaling dimensions. Dynamic aspects like SVC's loss resilience or the impact of layer switching and timing issues on quality perception are not included in this experimental design. In summary, we selected three factors as the independent variables to be controlled in this experiment: bit-rate reduction, scaling dimension and content type factors. We let the three factors change between 2, 3 and 6 levels, respectively. This study was then designed as a full factorial experiment with 36 treatment combinations.

We performed this experiment with the Double Stimulus Continuous Quality Scale (DSCQS) method as defined by International Telecommunications (2002). DSCQS is a hidden reference method where the original and a distorted video sequence (one of the operation points) are displayed twice in A-B-A-B order without disclosing the randomized position of the original. The assessors are asked to score the quality of both sequences on a continuous five-grade scale (see figure 2.5f). The differences of scores between a reference condition and a test condition are used to evaluate the perceived quality. According to the recommendation of International Telecommunications (2002), each video clip have the length of 8 seconds. We interspaced the A-B clips with 4 second breaks, displaying a mid-grey image with black text that announced the following clip or called for voting. Thus, the presentation time of one test sequence together with its reference is about 40 seconds. With the DSCQS method, a complete test session that covers all 36 treatment combination lasts for about 24 minutes, which is doable for a within-subjects design.

Because we are interested in realistic QoE perception on mobile devices, this experiment was conducted as a field study in indoor natural environments using iPods (generation 5.5) as mobile display device. This allows us to study natural user experience under

familiar viewing conditions rather than quality perception in a single synthetic laboratory environment.

4.3 Content Selection and Encoding

We selected six sequences from popular genres and with different characteristics (figure 3.2 and table 3.2) to assess effects of motion type (smooth, jerky) and speed (low, high), scene complexity (low, high) and natural vs. artificial content. All sequences were downscaled and eventually cropped from their original resolution to QVGA (320x240). From each sequence, we extracted an 8 second clip (200 frames) without scene cuts.

These video materials were then encoded by an H.264-SVC encoder, the SVC reference software (JSVM 9.12.2). The encoder was configured to generate bit streams in the scalable baseline profile with a GOP-size of 4 frames, one I-picture at the beginning of the sequence, one reference frame, inter-layer prediction and CABAC encoding. Due to the lack of rate-control for enhancement layers in JSVM, we determined optimal quantisation parameters (QP) for each layer with the JSVM Fixed-QP encoder.

Since we are interested in quality perception along and between different scaling dimensions, we defined a full scalability cube including the following downscaling paths.

- Spatial scaling: QVGA (320x240) $\mapsto$ QQVGA (160x120)
- Temporal scaling: 25 fps $\mapsto$ 12.5 fps $\mapsto$ 6.25 fps
- SNR scaling: 1536 Kbit $\mapsto$ 1024 Kbps, 256 Kbps $\mapsto$ 128 Kbps

The target bit-rates were chosen according to standard bit-rates of radio access bearers in wireless networking technologies such as HSDPA and DVB-H. For SNR scalability, we used SVC's mid-grain scalability (MGS) due to its adaptation flexibility that supports discarding enhancement layer data almost at the packet level (Schwarz et al., 2007).

From the scalable bitstreams, we extracted six scalable operation points (OP) which cover almost the total bit-rate operation range (see table 4.1). The selection lets us separately assess (a) the QoE drop for temporal scaling at the highest spatial layer (OP1, OP3, OP4), (b) the QoE drop of spatial scalability at two extreme quality points with highest frame-rate (OP1 vs. OP5 and OP2 vs. OP6), and (c) the QoE drop of quality scalability at two resolutions with highest frame-rate (OP1 vs. OP2 and OP5 vs. OP6).

4.4 Test Procedure

We displayed the test sequences in fullscreen on an iPod classic (80GB model, generation 5.5) as a typical mobile video player. Our iPod models contain a 2.5-inch display with 163 ppi and a QVGA resolution. Low spatial resolutions were upscaled to QVGA using JSVM normative upsampling and low frame-rates were upscaled by frame copy to the original 25 fps.

We applied within-subjects design in this study. All assessors received all test sequences in random order. The test session lasted for half an hour. Thirty non-expert

Available at http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm.

Operation Point	Spatial Resolution	Frame Rate	Quality Level	Layer ID	Average Bit-rate	Bit-rate Percentage
OP1	320x240	25.00	highest	23	1198.7	100%
OP2	320x240	25.00	lowest	14	1099.2	92%
OP3	320x240	12.50	highest	20	991.5	83%
OP4	320x240	6.25	highest	17	804.0	68%
OP5	160x120	25.00	highest	11	247.7	21%
OP6	160x120	25.00	lowest	2	123.1	10%

Table 4.1: Selected Operation Points. Frame-rates are given in frames per second (FPS) and bit-rates are given in kilobits per second (kbps).

assessors (33% female) in age classes between 18 and 59 with different education participated in the test. They watched the test sequences in a quiet lounge or office room.

4.5 Results

4.5.1 Data analysis

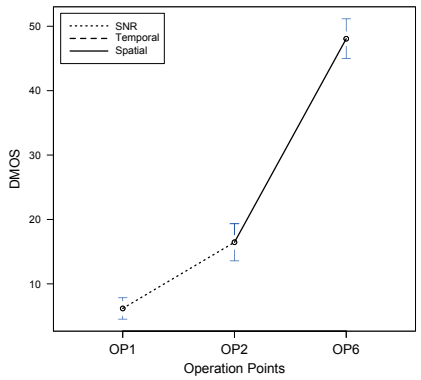
After carrying out the experiment, we calculated the differential mean opinion scores (DMOS) per operation point after quantizing the raw scores obtained from each assessor. The DMOS scores represent the subjective evaluation of the visual distortions in test sequences. Higher value of DMOS indicates lower perceived quality. To improve the accuracy of the study, we screened the scores for outliers and inconsistencies as defined by International Telecommunications (2002) and checked the reliability with Cronbach's alpha coefficient (Cronbach, 1951). As normality assumptions for DMOS scores were violated, we used conservative non-parametric statistics for further processing.

The objective of this analysis is to test the effect of downscaling (bit-rate reduction in three dimensions) on the perceived video quality, and whether the QoE degradations depends also on content or scaling dimension. For this purpose, we selectively compare two operation points and check if their DMOS scores differ significantly. A paired Wilcoxon test is used to test the significance. When testing the effect of downscaling, the alternative hypothesis is directional as we expect higher values of DMOS for lower-layer operation points. For the effect of scaling dimension, a two-tailed hypothesis is assumed.

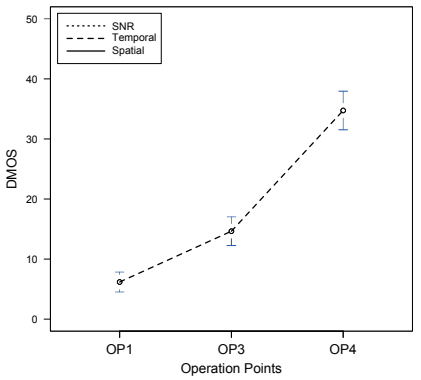
To provide further confidence in our observations, we also specify Cohen's statistical effect size and power (Cohen, 1988). Effect size helps to diagnose validity and discern consistent from unreliable results, e.g., a small effect size reflects a weak effect caused by small difference between scores. Power is the probability of not making a type-II error, that is, with low power, we might find a real existing effect as not significant.

4.5.2 Effect of bit-rate reduction

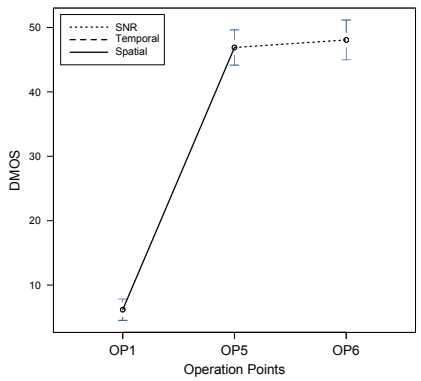
Bit-rate reduction attendant upon video scalability in any dimension will inevitably impair the physical quality of a video. To examine the severity of impairment, we connect any two OPs between which scaling possibility exists, and then we compare the DMOS



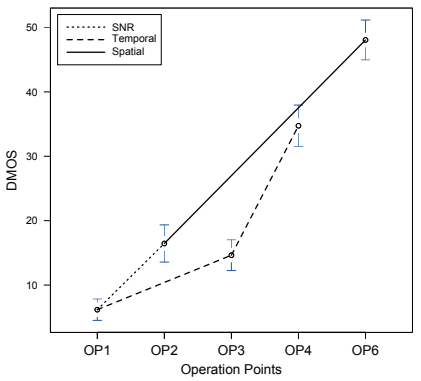
(a) Path: SNR ↦ Spatial



(b) Path: Temporal only



(c) Path: Spatial ↦ SNR



(d) Branch: SNR versus Temporal

Figure 4.1: QoE degradation along different adaptation paths, the DMOS values were averaged over six different content genres.

Dimension	**Temporal**	**Temporal**	**Temporal**	**Spatial**	**Spatial**	**SNR**	**SNR**
from	25 fps	12 fps	25 fps	320H	320L	320H	160H
to	12 fps	6 fps	6 fps	160H	160L	320L	160L
Scaling Option	$OP1 \mapsto OP3$	$OP3 \mapsto OP4$	$OP1 \mapsto OP4$	$OP1 \mapsto OP5$	$OP2 \mapsto OP6$	$OP1 \mapsto OP2$	$OP5 \mapsto OP6$
Animation	+++	+++	+++	+++	+++	+++	+
Cartoon	∘	∘	∘	+++	+++	++	∘
Documentary	++	+++	+++	+++	+++	∘	∘
Short Movie	+++	+++	+++	+++	+++	+++	∘
News	+++	+++	+++	+++	+++	∘	∘
Sports	+++	+++	+++	+++	+++	+++	∘
Total	++	+++	+++	+++	+++	++	∘
Saving(kbps)	207.3	187.5	394.7	951.1	976	99.6	124.5

Legend: ∘ not significant, + small effect, ++ medium effect, +++ large effect

Table 4.2: Effect of bit-rate reductions of available scaling options.

scores between the two OPs. The paired differences of DMOS scores serve as measures of noticeable QoE degradations.

Table 4.2 summarizes the analytical results. Figure 4.1 further illustrates the QoE evaluation along all possible adaptation paths enabled by the six operation points. The results show QoE degradations were noticed with a large effect size and sufficient power in almost all dimensions for almost all sequences. Although SNR downscaling gives the lowest bit-rate reduction ($\approx$ 99.6 kbps), the marginal means of DMOS scores at OP1 and OP2 still differ largely enough to show the significant main effect of bit-rate reduction. However, we do not find significant main effect on QoE degradation for the scaling option $OP5 \mapsto OP6$. Similarly, as the option $OP1 \mapsto OP2$, $OP5 \mapsto OP6$ reduces video bit-rate by the use of SNR scalability. But, in spite of the higher bit-rate reduction of $OP5$ and $OP6$, the DMOS scores of $OP6$ are not significantly higher than $OP5$, as demonstrated by the overlapping confidence intervals of $OP5$ and $OP6$ in figure 4.1c. This observation indicates downscaling may have no effect on the perceived quality if the bit-rate of video is already below a certain threshold, because people tends to be less sensitive to further QoE reductions when the quality was already poor.

4.5.3 Effect of scaling dimension

With multi-dimensional scalability, a video adaptation path may have one to several branches as shown in figure 4.1d. When more than one scaling options are available at the same time, we would have to answer the question: which scaling dimension gives the best trade-off between distortion and bit-rates? For this purpose, we compared operation points of scaling in different dimensions. To block the influence of bit-rate reduction, the comparison need to be performed among operation points at similar bit-rates. In this experiment, operation points OP2 and OP3 are two adaptation branches whose average bit-rates across six content types are at 1099 kbps and 991.5 kbps, respectively. By comparing their DMOS scores, we want to find out whether there exists a general preferred scaling order between SNR and temporal scalability.

However, we find no significant evidence for a general preference of SNR scalability over temporal scalability and vice verse (Wilcoxon T = 1.04, p=0.3). The DMOS scores of OP2 and OP3 do not differ much, as demonstrated by the overlapping confidence intervals in figure 4.1d. Counter-intuitively, OP3 received a better QoE score than OP2 although

it has a lower average bit-rate. This observation tells us that QoE degradations may be non-monotonic to bit-rate reduction when multi-dimensional scalability is assumed.

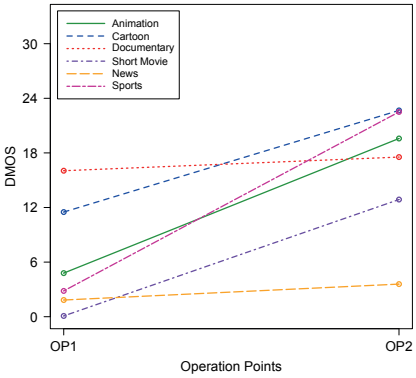
4.5.4 Interaction of content type

The third independent variable under investigation is the content related factor. We want to check the interaction effect of downscaling and content related factor, in other words, whether the effect of downscaling would differ depending on the video content. In table 4.2, we have already found the severity of QoE degradations are insignificant for some video sequences although there is a general main effect of downscaling. These exceptions of significant downscaling effect are the results of interaction of downscaling and content related factor. To have an better understanding of the interaction effect, we illustrate the variations of DMOS scores per content in figure 4.2.

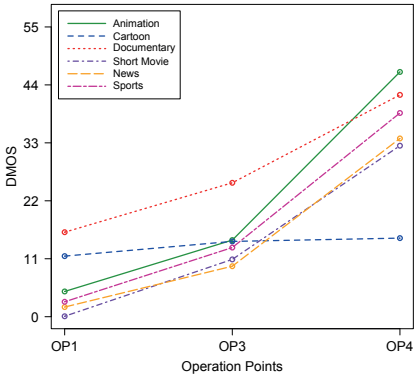
For the SNR scalability shown in figure 4.2a, the QoE degradations are only insignificant for both the Documentary and News sequence. The Documentary sequence has high complexity in texture details and medium high motion speed. For the same target bit-rate, the SVC encoder used coarser quantization levels to encode this video material, which results in larger impairments already in its layer OP1. Meanwhile, the Documentary sequence shows natural scenic textures of monkeys and jumbles of reedy grass. We think this is why its layer OP2 received relative lower DMOS scores since visual artefacts caused by quantization process such as blockiness, noise or edge busyness are less visible in nature and stochastic textures than smooth and homogeneous textures (International Telecommunications Union, 1996). The high DMOS score of OP1 and low DMOS score of OP2 demonstrated jointly the insignificant visibility of QoE degradation caused by SNR downscaling from OP1 to OP2. The same association between texture features and SNR related artefacts was also confirmed by the observation of the Cartoon sequence which is characterized by high contrast edges and large areas of monotonic textures. It is very likely due to these texture features that the Cartoon sequence received the highest DMOS score for its layer OP2 although relative smaller QPs were used during its encoding process.

As another exception of insignificant visibility of QoE degradation, the News sequence has medium complex texture detail and low motion speed. Low motion reduces the occurrence of the artefact of mosquito noise, but we suspect that other content related factors may also have their effects on the visibility of QoE degradation. The News sequence is a series of medium shots on an anchorman standing against the background of universe. The camera was tracked and zoomed in slowly so that the anchorman stays a constant size and motionless while the background shows a growing movement. This dynamic change of background attracts often viewer's attention and people tend to be less critical to fine details of objects from long distance or in the background. Therefore, we believe that the weak impact of SNR downscaling on this sequence may be attributed to people's expectations and attentions.

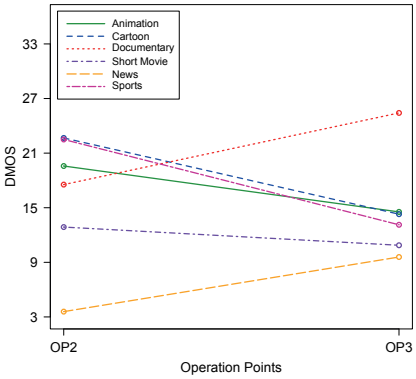
For the temporal downscaling, our observation is that video with high speed and unidirectional motion is more negatively affected by temporal downscaling. If we classify the six sequences into two groups of content with high and low amount of motion according to the calculated motion activities in table 3.2, it shows in figure 4.2b that the former group has significantly higher DMOS scores than the latter. Among the low motion



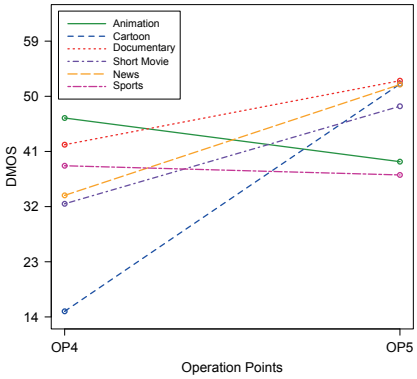
(a) SNR scaling



(b) Temporal scaling



(c) SNR versus Temporal scaling



(d) Temporal versus Spatial scaling

Figure 4.2: Interaction effect of content and scaling

sequences, Cartoon is particularly unaffected by frame-rate reduction. Even at a very low frame-rate (6fps at OP4), our assessors seemed to regard the QoE as sufficient. The reason is that the content already is non-naturally jerky. People do not expect object's movement to be as smooth as reality in this content genre. Again, the content related factors that can interact with quality scaling involve also people's expectations and daily practices.

The interaction effects between content related factors and scaling dimensions are illustrated in figure 4.2c, 4.2d. Obviously, the preferred scaling order of scaling dimensions varies for different content types. No main effect but interaction effect can be found when SNR scalability is compared with temporal scalability, as shown in figure 4.2c. Interestingly, figure 4.2d shows that the DMOS scores for spatial downscaling are higher than temporal downscaling for the Animation and Sports sequences despite the significant larger bit-rate difference between OP4 ($\approx 800kbps$) and OP5 ($\approx 256kbps$). Again, this indicates that monotony assumptions about the relation between bit-rate and QoE should be reconsidered for multi-dimensional scaling.

4.6 Limitations

Field studies generally suffer from less controlled presentation conditions. We therefore designed our study carefully by selecting more participants than required by ITU-R BT.500-11 and strictly removed outliers (6 in total among 30). To alleviate effects of an audio track which can influence video quality perception (Jumisko-Pyykkö and Häkkinen, 2006), we used undistorted, perfectly synchronised and normalised signals for all sequences. Although we are likely to miss effects that might have been observed in a laboratory, we still found significant results at significance level $p < 0.01$ of high statistical power and effect size in all tests. According to the power the number of participants was also sufficient for obtaining all results presented here.

DSCQS is sensitive to small differences in quality and used as quasi-standard in many subjective studies. For scalable content, however, it has two drawbacks. First, DSCQS is impractical to assess large numbers of operation points at several scaling dimension due to the limited amount of time before assessors become exhausted. Hence, we selected representative operation points only. Second, the scale used by DSCQS is ambiguous because QoE perception is not necessarily linear for people and individual participants may interpret scores differently (Watson and Kreslake, 2001). Hence, assuming DMOS scores obtained by DSCQS are interval-scaled is statistically incorrect. We address this by lowering our assumptions to ordinal data and non-parametric statistics. Despite these facts, we still found significant results and regard unnoticed effects as insignificant for mobile system design.

To reduce fatigue effects, only six video layers were included in this experiment. We therefore tested only the combination of SNR and spatial downscaling for the quality impact of combined scalability, while the temporal downscaling was only examined alone. In addition, the bit-rate granularity of the spatial layers was fairly coarse due to the dyadic rescaling method that was used to encode successive spatial layers, which makes it impossible to compare the quality influence of spatial downscaling with other downscaling options under the same bit-rate constraint.

Metric	CC	SROCC
Y-PSNR (copy)	-0.532	-0.562
Y-PSNR (skip)	-0.534	-0.555
SSIM (copy)	-0.271	-0.390
SSIM (skip)	-0.443	-0.451
NTIA GM	0.288	0.365
SVQM	-0.661	-0.684

Table 4.3: Correlation Results for Objective Quality Models. The higher the absolute value of a coefficient, the stronger correlation, namely the better prediction of the subjective scores. CC - Pearson Product-Moment Correlation Coefficient, SROCC - Spearman Rank-Order Correlation Coefficient.

4.7 Objective Model Performance

In this section, we analyse the performance of some existing objective video quality assessment models. Among many existing models, we selected three popular ones: Y-PSNR, SSIM (Wang et al., 2004) and the NTIA General Model (Pinson and Wolf, 2004). In addition, we implemented a recently proposed model which is specifically designed for video streams with multi-dimensional scalability (Kim et al., 2008). For simplicity, we call this model SVQM.

All the four objective models are full-reference quality metrics. For each test sequence, we compared the quality of all the extracted and decoded OPs with the original video sequence using these objective models. We omitted temporal and spatial registration because all decoded OPs are perfectly aligned with the reference video. For those OPs with lower frame-rate, the missing video frames were either skipped or the available frames were duplicated to replace the dropped frames. We performed skipping only for PSNR and SSIM to understand the influence of frame repetition and temporal scalability on those models. Finally, the video quality of each OP was quantified into a single value by averaging the quality values of each single or pair of frames. Then, we measured the objective model performance using Pearson's and Spearman's correlation coefficients between the objective scores and the subjective scores. Correlation was found to be significant with $p < 0.01$ at high power.

As table 4.3 reveals, SSIM and NTIA GM perform bad for scalable content on mobile screens. Although other studies reported good performance at television resolutions, both models are not tailored to multi-dimensional scalability and small screen sizes. PSNR performs only slightly better. SVQM achieved the best results of all examined models, but it is still far from being ideal. Although our version of SVQM is trained for the sequences used by Kim et al. (2008), it still creates reasonable results for our content. This indicates that the general idea of considering motion, frame-rate and spatial resolution in an objective model can yield some benefits. In contrast, a simple extension to traditional metrics like PSNR or SSIM which skips missing frames at low temporal resolutions does not create considerably better results.

4.8 Conclusion

We performed a subjective field study on mobile devices to investigate the effects of multi-dimensional scalability supported by H.264-SVC on human quality perception. In this experiment, three types of factor, bit-rate reduction, scaling dimension and content related factors have been investigated simultaneously. Not surprisingly, the magnitude of bit-rate reduction resulted from downscaling has been found having significant effect on the perceived video quality, but the effects differ depending on the content related factor and the actually scaling dimension. Due to the interaction of scaling and content related factor, we found no significant evidence for a general preference for one scaling dimension over the other. Videos with different content characteristics are influenced to various extent by the multi-dimensional bit-rate adaptation techniques. We observe that QoE degradation followed by bit-rate downscaling are more evident on videos with smooth textures and high motion activities. Moreover, a user's expectation and attention to the concrete video content play also an important role on human perception of video quality. Based on our experimental results, we concluded also that adaptation decisions for SVC bitstreams should not only be based on bit-rate and layer dependency information alone, as the perceived quality degradation may be non-monotonic to bit-rate reduction.

Despite the MGS encoding mode, the scaling granularity of H.264-SVC video streams may still be insufficient for efficient bit-rate adaptation. In the experiment introduced in this chapter, the bit-rate intervals between adjacent video layers were even found to be uneven. To mind large gaps between the average bit-rates of two video layers, streaming schedules need to change layers alternatively from high to low quality and vice verse. For scalable video streams that have multi-dimensional scalability, it is unclear how the perceived quality would be influenced by active layer variations and whether layer variations in different scaling dimensions would be perceived differently. We investigate these issues in chapters 5 and 6.

Chapter 5

Frequent layer switching

Frequent switching between the layers of a scalable video can enable finer grained bandwidth adaptation, but also generates additional visual artefacts. In this chapter, we will present another three subjective quality studies that evaluated the feasibility of switching techniques.

5.1 Introduction

In the context of scalable video streaming, the term "granularity" describes bit-rate difference between quality layers in video encoded in layered structure. The finer grained scalability a video has, the more precisely it can adapt to the available bandwidth, because that fine-grained scalability allows rate to be added or reduced in smaller amount. Therefore, it is often preferable to stream video with fine grained scalability. However, fine-grained scalable video suffers from a considerable coding penalty. For instance, the H.264 Scalable Video Coding (SVC) standard included the medium grain scalability (MGS) feature in addition to the coarse grain scalability (CGS) method for higher adaptation granularity, but this comes at the cost of increased signaling overhead, see the comparison of bit-rates of video with difference scaling granularities in table 5.1. For better bit-rate efficiency, SVC has to limit the number of supported enhancement layers and the number of MGS partitions. Furthermore, the bit-rate increments between adjacent quality layers in SVC encoded videos are not equal. In chapter 4, we reported a user study on material that has been encoded by the H.264-SVC encoder. From that study, we found that at low bit-rates less than 200 Kbps, scalable streams with a fixed set of predetermined quality layers have sufficient granularity; in the higher bit-rate range, the granularity becomes significantly coarser though. This also results in insufficient adaption granularity, either wasting resources or decreasing the quality of experience (QoE) more than necessary. The question arises as to whether more or less frequent switching between the layers of a coarse-grained scalable video could yield similar bandwidth adaptation while providing better perceived quality.

This chapter summarizes and extends our previous work (Ni et al., 2009, 2010) with more analyses and conclusions about the technique of **frequent layer switching** (FLS), a method for fine-grained bit-rate adaptation of scalable bitstreams with few scaling options. This technique is an alternative to static **downscaling**, which is the selection of a single, lower-quality operation point for an entire video sequence. In contrast to downscaling,

frequent layer switching alternates between two or multiple operation points in order to meet a given bit-rate constraint over a short time-window. FLS brings about repeated quality fluctuation due to frequent layer variation. Different FLS schedules can create different layer variation patterns, which may cause different impacts on perceived visual quality. In addition, H.264-SVC increases perceptual uncertainty dramatically because of its multi-dimensional scaling possibility. Layer variations that occur in different scaling dimension may not be perceived in the same way, either. To examine these issues, we conducted a **FLS study** that investigated the perceptual effects and usefulness of FLS. In this thesis, we consolidate the findings of this study by more analyses than our original paper included in Appendix D. Our aim was to provide recommendations on how to best incorporate FLS into practical streaming systems. In general, we were interested in two central questions:

- Is FLS a useful alternative to downscaling in streaming scenarios with limited and fluctuating bandwidth?
- How do switching schedules and streaming environments influence the subjective quality perception of human observers?

Layer switching can achieve a bandwidth consumption different from the long-term average of any operation point of a coarse-grained scalable video without the extra costs of MGS. This ability makes FLS suitable in several streaming scenarios:

- FLS can be used to achieve a long-term average target bit-rate that differs from average bit-rates of available operation points in coarse-grained scalable videos. This works even for variable-bit-rate SVC streams. Every average target bit-rate above the base layer's bandwidth demand can be achieved by switching enhancement layers on and off repeatedly, if necessary at different on and off durations.
- FLS can be used as an alternative means to exploit the temporary availability of bandwidth that exceeds the demands of the base layer, but does not suffice for the bandwidth demands of an enhancement layer. Through variations of the retrieval speed (implicitly in pull mode, explicitly in push mode), receivers can use the excess bandwidth during a period of base-layer playout to prefetch data for a period of enhanced-quality playout. The period duration depends on the available space for a prefetching buffer, but it also depends on the perceived playout quality which forbids an arbitrary choice.
- FLS can be used for bandwidth sharing in fixed-rate channels, in particular, for multiplexing multiple scalable bitstreams over Digital Video Broadcasting channels. With FLS, a channel scheduler gains more selection options to satisfy quality and bit-rate constraints. In addition to coarse operation point bit-rates, FLS can offer intermediate bit-rates at a similar quality of experience.

In all the above scenarios, the choice of switching pattern and switching frequency are of central importance because they may considerably impact the perceived quality. To identify the feasibility of switching techniques and advice design constraints, our FLS study includes several subjective quality assessment experiments collecting human observers' preferences when watching video clip pairs impaired with different switching and scaling patterns.

The experiments were performed in three different scenarios, i.e., mobile displays in private spaces, mobile displays in public spaces and HTDV displays in private spaces. Due to the differences of the scenario environments, we used multiple assessment methods to carry out these experiments in the FLS study.

The visual effects of quality impairment in temporal and SNR dimensions are significant different. McCarthy et al. (2004) performed several subjective tests to compare the effects of image fidelity and frame-rate downscaling. It was shown that high frame-rate is not always more preferable than high image fidelity, even for high motion video. Probably closest to our FLS study, Zink et al. (2003) investigated quality degradation caused by layer variations. In contrast to our work, they did not treat the layer switching and its related impairment in different dimensions separately, although different types of layer variation deserve an in-depth study. In our FLS study, we identified the noise flicker, motion flicker and blur flicker as three specific effects caused by FLS in separate dimensions. Our work compared the two-dimensional video impairment systematically and investigated how the visual effects are related to content, device and adaptation strategy.

Our results indicate that the perceived quality of different switching patterns may differ largely, depending on scaling dimensions, content and display device. In some cases, there are clear preferences for one technique while in other cases both, switching and downscaling, are liked or disliked equally. In several cases, FLS is a practical alternative for achieving fine-grained scalable streaming from coarse-grained videos, specifically, if the switching period is long enough to avoid flicker effect, then layer switching is even preferred over downscaling to a lower SVC quality layer.

5.2 Frequent layer switching study

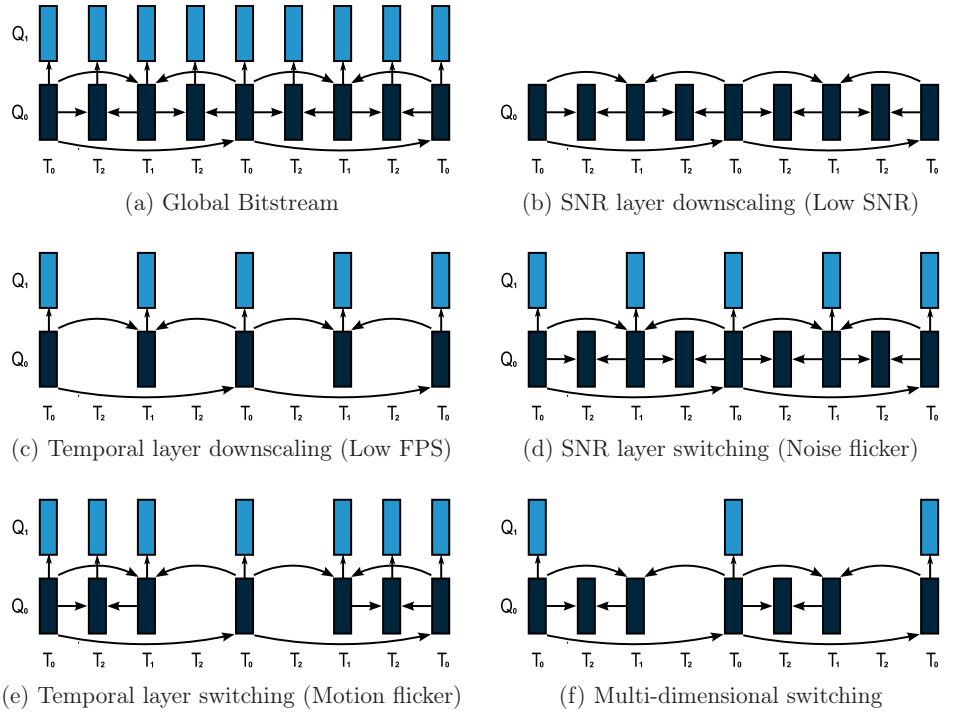
One of the main goals of this study is to see if the FLS technique can be used to achieve a more efficient fine-grained streaming solution considering the high overheads of MGS coding schemes. In this section, we identify the factors under investigation and describe the choice of factor levels and ranges.

5.2.1 FLS

In contrast to adaptation approaches that downscale a SVC bitstream to a particular operation point among many MGS or CGS quality levels that was predetermined at encoding time, FLS alternates between a few fixed operation points in order to meet a given bit-rate constraint over a short time-window without the extra overhead of defining additional operation points. For video with multi-dimensional scalability, layer switching is not limited to one single dimension. For instance, figures 5.1b –5.1c show two different approaches for downscaling; while figures 5.1d–5.1f illustrate three different switching patterns, two that perform switching in a single dimension (temporal or SNR) and one pattern that combines layer switching in the two multi-dimensions.

Thus, FLS introduces intermediate scaling options, but it also causes three perceptible effects on the users QoE:

- **Noise flicker** is a result of varying the signal-to-noise-ratio (SNR) in the pictures. It is evident as a recurring transient change in noise, ringing, blockiness or other



(a) Global Bitstream (b) SNR layer downscaling (Low SNR)

(c) Temporal layer downscaling (Low FPS) (d) SNR layer switching (Noise flicker)

(e) Temporal layer switching (Motion flicker) (f) Multi-dimensional switching

Figure 5.1: Bitstream layout for downscaling and layer switching options used in the experiments. Typical associated visual effects are given in parenthesis.

still-image artefacts in a video sequence.

- **Blur flicker** is caused by repeated changes of spatial resolution. It appears as a recurring transient blur that sharpens and unsharpens the overall details of some frames in a video sequence.

- **Motion flicker** comes from repeated changes in the video frame-rate (FPS). The effect is a recurring transient judder or jerkiness of naturally moving objects in a video sequence.

The choice of switching pattern and switching frequency are therefore of central importance due to the possible high impact on the perceived quality. Questions such as "under which conditions (e.g., viewing context, display size and switching frequency) these effects become noticeable" and "how they influence the perceived quality impression" are therefore important research issues. Furthermore, to identify the feasibility of switching techniques and advice design constraints, we were interested in answering the following questions:

- Do people perceive a difference in quality between scaling and switching techniques?

- Is there a general preference of one technique over the other?

Genre	Content	Detail	Motion	Audio	CGS Bitrate		MGS Bitrate	
					Max	Min	Max	Min
Animation	BigBuckBunny	3.65	1.83	sound	530.8	136.1	823.6	175.5
Cartoon	South Park	2.75	0.90	speech	533.8	158.8	767.5	199.7
Docu	Monkeys & River	3.64	1.61	sound	1,156.1	192.1	1,244.3	208.7
Movie	Dunkler See	1.85	0.58	sound	255.2	67.9	419.9	92.4
News	BBC News	2.92	0.69	speech	268.6	74.0	453.1	101.0
Sports	Free Ride	3.32	1.90	music	734.8	121.1	745.9	129.1
HD-Animation	BigBuckBunny	2.88	4.13	sound	10,457.0	10,32.4	14,210.0	1,021.7
HD-Docu	Canyon	3.09	3.33	sound	25,480.0	24,07.0	28,940.0	2,394.0

Table 5.1: Expanded test sequences for the experiments on both mobile device and HDTV monitor. The video source material were the same as used in a previous study, illustrated in table 3.2 and figure 3.2. Detail is the average of MPEG-7 edge histogram values over all frames (Park et al., 2000) and Motion is the MPEG-7 Motion Activity (Jeannin and Divakaran, 2001), i.e., the standard deviation of all motion vector magnitudes. Bit-rates are given in Kilobits per second (Kbps) for the SVC bitstream at the highest enhancement layer (max) and the base layer (min).

- Does a preference depend on genre, switching frequency or the scaling dimension?
- Are there frequencies and dimensions that are perceived as less disturbing?
- How general are our observations, i.e., do location, device type, display size and viewing distance influence the results?

5.2.2 Subjective quality evaluation

From the above questions, we defined several factors that may affect the perceived quality of video. Frequent switching and static downscaling are two different adaptation strategies. Combined with different scaling dimensions, it creates multiple streaming options. We use categorical variables to represent these options. In addition to streaming strategy and scaling dimension, content and switching frequency are also factors of interest. The switching frequency is determined by the playback duration of the scalable video layers. Since we examine periodic and symmetric layer switching pattern shown in figures 5.1d and 5.1e in this study, the term period, the playback duration of each layer, is used as the equivalent measure of the switching frequency. Finally, we use a single variable defined as streaming scenario to cover all the device and environment related factors. Subjective video quality can only be evaluated separately in different streaming scenarios. Thus this study was performed as a set of independent experiments. We deliberately controlled the variations of the other four factors in each experiment.

In this study, as more factors were involved in the evaluation, the test workload was increased considerably. To reduce the complexity of evaluation tasks, the Pairwise Comparison method was used as the main form of quality assessment instead of the DSCQS method. As mentioned earlier, with the PC method, our assessors need simply make a choice based on preference when watching video clip pairs.

To test different kinds of content with varying detail and motion, we selected eight sequences from different genres (see table 5.1), i.e., 6 for the small mobile devices and 2 for the HDTV. We obtained the content from previous user studies (see chapters 3, 4)

which allowed for a comparison with earlier results. From each sequence, we extracted an 8 second clip without scene cuts. After extraction, the texture complexity and motion activity are measured according to MPEG-7 specification.

We encoded the SVC bitstreams with version 9.16 of the JSVM reference software. The encoder was configured to generate streams in the scalable high profile with one base layer and one CGS enhancement layer, a GOP-size of 4 frames with hierarchical B-frames, an intra period of 12 frames, inter-layer prediction and CABAC encoding. Note, SVC defines the set of pictures anchored by two successive key pictures together with the first key picture as a group of picture, where key pictures are usually encoded as P-frames within an intra period (Schwarz et al., 2007). Due to the lack of rate-control for quality enhancement layers in JSVM, we used fixed quantization parameters (QP). Based on our previous experience and in order to obtain a perceivable quality difference, we selected QP36 and Q28 for encoding the base-layer and the enhancement layer respectively.

From the encoded SVC bitstreams, we extracted three scalable operation points with high variability in the bit-rates (see figures 5.1a–5.1b). The full bitstream (figure 5.1a) is the operation point containing the base-layer (Q_0) and the SNR enhancement layer (Q_1) at the original frame-rate, while the two other operation points are each downscaled in a single dimension to the lower SNR layer at full temporal resolution or a lower temporal resolution T_1 (12fps), but with SNR enhancement. These operation points were then used to generate streams with different switching patterns and to compare the switched streams' quality. Note that we only focused on SNR scalability and temporal scalability in this study. We did not consider spatial scalability, because it is undesirable for FLS due to the large decrease in perceived quality as shown in chapter 4.

Next, we have performed experiments in three different scenarios: mobile displays in both private and public spaces and HDTV displays in private spaces trying to find suitable switching patterns from the downscaling operation points (figures 5.1c–5.1b) resulting in patters like the ones shown in figures 5.1d–5.1f, i.e., better and more efficiently matching the available bit-rates between the downscaling operation points giving better video quality than the lower base layer only.

5.3 Mobile scenario - Field study 1

In our first experiment, we were interested in how user perception over FLS compared to static layer downscaling. The experiment is performed in a private, in-door environment (lab), and each participant evaluated all the video content.

5.3.1 Experimental design

The first experiment was designed based on the standardized Full factorial PC method (F/PC). Since the comparison between the two streaming strategies, frequent switching and static downscaling is of primary interest, we only paired up the switching pattern with one static operation point. The dimension varied between two levels, the temporal and SNR dimension. Then, we decided to use 3 and 6 levels of the factor switching frequency and content genre. In total, the F/PC design included $2 * 2 * 3 * 6 = 72$ treatment combinations.

In order to assess the reliability of votes from each participant and detect inconsistent ratings, each pairwise comparison was repeated twice during a test session, once in each possible presentation order of a pair of video clips. The overall presentation order of all clip pairs was a random permutation. Between subsequent pairs, there was a 6-second break, displaying a mid-grey image with black text that called for voting and announced the following clip pair.

For each clip pair, we obtained a single measure about which clip a participant preferred to watch. A three categorical scale was used for reporting the preferences. The participants were asked to judge whether they preferred the first or the second video clip in a clip pair or whether they had no preference if undecided. All ratings from both clip-pair orders (AB, BA) were used in our analysis. For statistical analysis, we first ran Binomial tests to see if a significant majority of ratings for one of the three rating categories existed. A significant majority was defined as the number of ratings that have occurred more frequently than the summed frequency of the other two kinds of ratings. The hypothetical probability of obtaining such a majority is $1/3$ assuming there is no effect of the independent variables. In the case that a group of data did not pass the significance test, we further examined whether the independent variables still have some weaker effects on people's preference. The ratings for no preference were then divided into halves and regarded as ratings for either A or B clip. The second Binomial test was called to examine if the number of ratings for A clip differs significantly than the ratings for B clip. If none of the two significance tests can identify a clear preference between the pair of clips, it indicates a lack of consensus of people's perception, which is in practice equivalent to an undecided preference for the given test conditions.

5.3.1.1 Material

In this experiment, we again (as used in chapters 4, 3) tested video from all the six different genres listed in the top of table 5.1. The selected six sequences were downscaled and eventually cropped from their original resolution to QVGA (320x240) in order to fit the screen size of our display devices. We simulated layer switching in the SNR dimension and the temporal dimension according to the patterns illustrated in figures 5.1d and 5.1e. The switching periods that were chosen for this experiment were 2, 24 and 48 frames, which correspond to about 0.08, 1 and 2 seconds for a 25 fps playback rate.

5.3.1.2 Participants

Twenty-eight payed participants (25% female) at mean age of 28 participated in the test. Among the participants, 90% are at an age between 18-34 while 10% are at an age between 35-39. All of the participants are college students with different education but no one has majored in multimedia technologies. All of the them are familiar with concepts like digital TV and Internet video streaming while 75% of them claimed that media consumption is part of their daily life. We obtained a total of 2016 preference ratings of which 44% indicated a clear preference (consistent ratings on both clip orders), 31% a tendency (one undecided rating) and 10% no difference (two undecided ratings). We observed 15% conflicting ratings, where participants gave opposite answers to a test pattern and its hidden check pattern. Participants with more than 1.5 times the interquartile range of conflicting ratings above the average were regarded as outliers. In total,

we removed two outliers from our data set. Regardless of remaining conflicts we found statistically significant results.

5.3.1.3 Procedure

As mobile display devices, we again used the iPod classic and the iPod touch from 2008. The two iPod models contain respectively a 2.5-inch and 3.5-inch display and have pixel resolutions of 320x240 and 480x320 at 163 pixel per inch. The selected display size is sufficient for depicting content at QVGA resolution according to Knoche and Sasse (2008). All videos had an undistorted audio track to decrease the exhaustion of test participants.

Although quite a few of assessors have previous experience in watching video on handheld devices like iPod, a brief introduction about how to operate the iPods during the experiments was given to the assessors prior to a test session. A whole test session lasted for about one hour, including two short breaks. Each participant completed the entire test session. During the session the assessors were free to choose a comfortable watching position and to adjust the watching distance. For example, they could choose to sit on sofas or in front of a desk. They were also free to decide when they wanted to continue the test after a break.

5.3.2 Results

Results are reported as preference for layer switching or layer scaling with 0.01 confidence intervals. If a preference was verified by only one Binomial test, we reported it as weak preference. Table 5.2 displays preferences between switching and scaling across different content genres and period lengths. The 'all' line in the table contains general results averaged over all periods and all genres.

	Motion flicker		**Noise flicker**	
Group	**Low FPS**	**Low SNR**	**Low FPS**	**Low SNR**
Animation	∘	+	−	∘
Cartoon	∘	+	−	(+)
Documentary	∘	+	−	(−)
Short Movie	∘	+	−	∘
News	∘	+	−	∘
Sports	∘	+	−	∘
2f			−	−
24f	∘	+	−	(+)
48f	∘	+	−	+
All	∘	+	−	(+)

Legend: + switching preferred, − downscaling preferred
∘ no preference, (*) weak tendency

Table 5.2: Private space mobile - quality preference for layer switching vs. downscaling, Empty cells are not covered by this study.

5.3.2.1 Temporal layer switching

Participant ratings indicate no clear preference when layer switching is compared to layer downscaling in temporal dimension. One possible reason for this observation is that temporal resolution changes between 25 fps and 12 fps have a minor impact on quality perception. This confirms results of previous studies reported in section 4. Using more bandwidth for a temporally switched stream (92%) compared to a temporal downscaled stream (85%) is thus not justified by a significant increase in quality perception. It may be interesting to further investigate whether this observation also applies to switching to lower temporal resolutions (below 10 fps).

When layer switching in the temporal or SNR dimension is compared to downscaling in the other dimension (Motion flicker vs. Low SNR and Noise flicker vs. Low FPS, respectively), the results indicate a clear preference towards decreasing the temporal resolution rather than the fine details of a video. With high significance, our results are consistent across all genres and independent of the switching period. The result again confirms previous findings reported by McCarthy et al. (2004). People seem to be more sensitive to reductions in picture quality than to changes in frame-rates when watching video on mobile devices, i.e., lowing the temporal resolution will, however, be much more visible on larger screens like HDTV monitor (see section 5.5). This clearly indicates that switching is a viable option for frequent temporal resolution changes on mobile devices. Although temporal base-layers consume the main bit-rate and potential savings are small, switching can still yield fine-grained adaptation in the upper bit-rate range of a stream. For a fair comparison, it is noteworthy that the temporal layer switching scheme in the FLS study made the bit-rate considerably higher than the static SNR downscaling option (92% vs. 28%). However, the comparison between the SNR layer switching pattern and the Low FPS option (89% vs. 85%) shows, that a lower bit-rate stream can yield a higher subjective quality regardless of the content.

5.3.2.2 SNR layer switching

When the visual artefact associated with SNR layer switching (Noise flicker) is compared to Low SNR effect associated with SNR downscaling, the combined results over all period sizes shows only a weak tendency of preference for Noise flicker. There is also no general preference towards a single adaptation technique that can be attributed to content characteristics alone. However, we observed a significant preference for SNR layer switching at long periods (48f) while for shorter periods (2f) a preference for static SNR downscaling exists.

Noise flicker is caused by fast switching between high- and low-quality encodings which leads to rapid iteration of high and low frequency textures. It was perceived as disturbing by almost all participants. At longer switching periods, this effect becomes less annoying and disappears eventually. We call the limit at which flicker effect disappears the *flicker threshold*. Only above the flicker threshold, people can pay enough attention to the fine details of a video. Interestingly, it seems that the flicker threshold is within the range of the selected period levels in this study, and long switching period above the flicker threshold makes the fluctuant visual quality more preferable than a constant low quality.

We just conducted tests with equally long intervals of high and low quality. Hence, the bit-rate demand of a SNR layer switching scheme is still much higher than that of

the Low-SNR downscaling option (89% vs. 28%). Asymmetric patterns with longer low-quality intervals will have a much lower bit-rate consumption and offer a wider range of bit-rate adaptation. We will investigate whether such patterns can also yield a better visual quality. We assume, however, that the flicker threshold plays an important role for asymmetric patterns as well.

5.4 Mobile scenario - Field study 2

The first FLS study was conducted in an indoor environment, but it is also interesting to see if the environment also influence the results. Therefore, the second FLS study was conducted in a more bustling public space environment. The purpose of this experiment was to verify whether the conclusion drawn in section 5.3 applies to people with a more varied background in a different streaming scenario.

5.4.1 Experimental design

The primary concern that had arisen from the first FLS study (section 5.3) was the long duration of each assessor's viewing time; about one hour. Although assessors had been allowed to take breaks, they were generally annoyed with the test itself, and we were concerned that this can have had unpredictable effects on the quality of their evaluation. Furthermore, the video quality tests in our first FLS study were mostly performed at Simula and on the university classrooms. In public spaces, it is not realistic to recruit volunteers to a test that lasts for an hour.

With this consideration, we designed our second field study based on the Randomized Pairwise Comparison method (R/PC). As described in chapter 3, R/PC is a flexible and economic extension to traditional pairwise comparison (PC) designs. Conventionally, it presents stimuli as pairs of clips. In contrast to traditional full factorial PC design that collects a full data sample for all pairs from every assessor, R/PC uses random sampling to select small subsets of pairs and thus creates a shorter but unique experiment session for each assessor. With this method, our assessors were allowed to stop at anytime, viewing and evaluation was better integrated, so that this study contained the same number of treatment combinations as the first FLS study.

5.4.1.1 Material

In this field study, we used the same video material to generate our test sequences as in section 5.3. We used iPod touch devices from 2008 to perform the tests and used encoding settings that were similar to those of the first field study, except that the resolution was changed. Instead of scaling the video on the devices itself, all sequences were downscaled and cropped from their original resolution to 480x272 pixels in order to fit the 3.2-inch screen size of iPod touch and keep the 16:9 format.

From the first field study, we already found out that quality changes at a very high frequency can make FLS significantly worse than static downscaling in terms of visual quality. In this study, the selected period levels were prolonged correspondingly. This level selection enabled the exploration of more values of switching periods, but it also made the comparison between the two studies not straightforward. The switching periods that

were chosen for this experiment were 12, 36 and 72 frames, which correspond to about 0.5, 1.5 and 3 seconds for a 25 fps playback rate.

5.4.1.2 Participants

The field study was performed under conditions that differ from the first one in several ways. Participants were approached by students in public locations in Oslo in the summer and autumn. They were approached in situations that we considered realistic public areas for the use of a mobile video system. We had 84 participants who had mostly been approached when they were idle, e.g., waiting for or sitting on a bus. They were asked for 15 minutes of their time.

Among the participants, 74% are between the age of 18-34, 20% are between the age of 35-59 and 6% are at an age under 18. 96% of the participants have normal visual acuity with or without glasses while 4% have limited visual acuity in spite of glasses. The field study was mostly conducted indoors (95%) in different locations (restaurant, bus station, cafeteria), while 3 participants were en-route and 1 person was outdoors. We gathered in total 2405 ratings of which 30% indicated a clear preference (consistent ratings on both clip orders), 36.3% a tendency (one undecided rating), 24.4% no preference (two undecided ratings) and 8% conflicting ratings. Using the same criterion introduced in section 5.3, we filtered out 3 unreliable participants.

5.4.1.3 Procedure

Consistently with an experiment that was as close to the real world, we did not control lighting or sitting conditions. Assessors were not protected from disturbances that are consistent with those that a user of a mobile video service would experience. They experienced distractions by passerbys, or the urge to check departure times or the station for the next stop. In case of such a short-term disturbances, they were allowed to continue watching clips pairs in the same test sessions.

Assessors were not shown training sequences, but they received a brief introduction by the student, explaining that clips might look identical. The suggested number of clips to be watched by an assessor was 30, but considering the experience of fatigue and annoyance with the first experimental design and the situation of the assessors, they could terminate the experiment at any time. The downside of this possibility was that the consistency of an individual assessor's answers could not be checked, and that every vote for a clip pair needed to be considered an independent sample. Lacking the control mechanism, we required 20 or more votes for each clip pair. Following this method, the test participants were asked to assess the quality of two sequentially presented clips. A subset of clip pairs was randomly chosen for each assessor from a base of 216 clip pairs (including reference pairs, equal reference and matched contrast pairs).

The evaluation procedure was changed from the paper questionnaire approach taken in section 5.3. This field study integrated both testing and evaluation into the iPod. Thus, users were given the opportunity to first decide whether they had seen a difference between the clips after each pair of clips that they had watched. If the answer was yes, they were asked to indicate the clip with higher quality.

5.4.2 Results

The results of the second field study are processed in the same way as those for the first FLS study. Confidence intervals are reported as 0.01. Table 5.3 presents preferences between switching and scaling across different content genres and period lengths. The 'all' line contains general results for all periods and all genres.

	Motion flicker		**Noise flicker**	
Group	**Low FPS**	**Low SNR**	**Low FPS**	**Low SNR**
Animation	∘	+	−	+
Cartoon	∘	+	(−)	(+)
Documentary	∘	+	∘	(+)
Short Movie	∘	+	∘	∘
News	∘	+	−	(+)
Sports	∘	+	∘	∘
12f	∘	+	(−)	(+)
36f	∘	+	(−)	(+)
72f	∘	+	(−)	∘
All	∘	+	(−)	(+)

Legend: + switching preferred, − downscaling preferred
∘ no preference, (*) weak tendency

Table 5.3: Public space mobile - quality preference for layer switching vs. downscaling.

5.4.2.1 Temporal layer switching

Two series of ratings provided by the assessors yielded results that were identical independent of genre. In the comparison of temporal switching (Motion flicker) and temporal downscaling (Low FPS) in table 5.3, our random, untrained assessors did not favor either option for any type of content independent of motion speed in the clip. This makes it very clear that a frame-rate difference of 25 fps versus 12 fps on a mobile device has minimal impact to the casual viewer. Additionally, temporal layer switching is given a clear preference over SNR downscaling (Low SNR) for all types of content. This repeats the equally clear findings of the first field study. Both of these comparisons stay the same when different switching periods are considered.

5.4.2.2 SNR layer switching

The general observation about SNR layer switching (Noise flicker) shows people still tend to prefer Low FPS rather than Noise flicker in public space. However, the preference is detected much less clearly in the second field study than in the first. Similar to the first study, a tendency of preference for Noise flicker over Low SNR was also found in table 5.3 for the entire data sample. As the result of extended period levels, no clear preference for Low SNR or Noise flicker was found in table 5.3 though. A remarkable new finding is that assessors did not prefer Noise flicker or Low SNR when the switching period reached 72 frames. One possible reason for this indecisiveness is that people are able to separate the Noise flicker and Low SNR as two different kinds of artefacts at this longer time-scale,

they prefer neither the constant low picture quality nor the instability associated with quality changes.

Content related factor has also demonstrated some influences on user's preferences in this study. Significant preference for Noise flicker over Low SNR has been detected for some content genre and people's preference for Low FPS over Noise flicker was not completely independent on video content.

5.4.2.3 Comparison between studies

The effect of content related factor was hidden in the first FLS study because of the different period level selection. To reduce the confounding and suppression effects from the period factor, we further analyzed the experimental data from the first study by excluding the exceptional period level at 2f. Correspondingly, the period level at 72f was filtered out from the second FLS study. This helps us to make a side-by-side comparison between the two studies, as the differences between the remaining period levels become relatively small and all the period levels have been verified to have similar effects on people's preference. We report the analytical results in table 5.4.

	Noise flicker vs. Low FPS		**Noise flicker vs. Low SNR**	
Environment	Private	Public	Private	Public
Animation	–	–	+	+
Cartoon	–	(–)	+	+
Documentary	○	○	(+)	○
Short Movie	–	○	○	○
News	–	–	(+)	(+)
Sports	(–)	○	(+)	○
All	–	(–)	(+)	(+)

Legend: + switching preferred, – downscaling preferred
○ no preference, (*) weak tendency

Table 5.4: The perceived mobile video quality in private and public spaces - quality preference for layer switching vs. downscaling.

Table 5.4 shows that the experimental findings of FLS study 1 and FLS study 2 are in substantial agreement. But, people's preference seems to be less clear in public spaces. For instance, the general observation shows only a weaker tendency of preference for Low FPS in the second FLS study, while Low FPS was clearly preferred in the first study. We attribute this uncertainty to the environmental disturbance. The effect of content related factor was revealed in the private spaces by this analysis. Based on the observations of two field studies, we found that video with non-natural texture are less influenced by flicker artefacts. The Animation and Cartoon clips are typical examples for artificial texture that appears sharper than natural texture due to higher edge contrast. In our experiments, clear preference for Noise flicker over Low SNR are found repeatedly only for the two clips. Similarly, people may not expect the movements of computer generated figures to be as smooth as natural movements. Therefore, despite of high motion activities, the frame-rate loss seems to be less severe in the Animation clip than the other high motion clips such as Document and Sports.

Finally, it can be mentioned that the SNR layer downscaling strategy never got more ratings than SNR layer switching in the experiments, which indicates that our assessors, untrained and randomly chosen, can differentiate the amount of normal compression artefacts in a video (8 QP differences) even in a noisy viewing environment.

5.5 HDTV scenario - Field study 3

With respect to both environment and device, there are large differences between small mobile devices like iPods and large, high-resolution devices like a 42-inch HDTV. The goal of our third experiment was to validate whether the results obtained in the mobile scenarios are general observations or whether the results depend on the screen size and viewing distance.

5.5.1 Experiment design

As we did in the first experiment described in section 5.3, the standardized Full factorial PC method was applied to test whether either the downscaling or the switching video adaptation options did significantly affect whether a user perceived the one or the other as better. The assessors could select if they preferred layer switching or layer downscaling, or if they had no preference. After gathering enough votes, we ran binomial tests to see if a significant majority of the ratings exist among the three rating categories.

5.5.1.1 Material

We prepared the test sequences in a similar way to our previous experiments. We encoded one base layer and one CGS enhancement layer using fixed quantization parameters of 36 and 28, respectively. The original spatial resolution of 1920x1080 was preserved.

Two HD video sequences (see table 5.1) were selected to represent video with natural and non-natural textures. The HD-Animation test sequence had the same content as the animation movie in the mobile tests. The HD-Docu sequence was extracted from the same Documentary movie accordingly, but another part to fit the visual characteristics and potential to HDTV presentation.

5.5.1.2 Participants

The study was conducted with 30 non-expert participants in a test room at Oslo university. All of them were colleagues or students between the age of 18 and 34. Three of them claimed to have limited visual acuity even with glasses. In total, we gathered 720 preference ratings of which 49% indicated clear preference, 29% a tendency and 12% no preference. In the results, there were 10% conflicting ratings. We removed three outliers from our data set using the same criterion as that introduced in section 5.3.1.2.

5.5.1.3 Procedure

The visual setup was a 32-inch, 1080p HDTV monitor. Our assessors were seated directly in line with the center of the monitor with a distance of about three monitor screen heights (3H distance). Since we conducted the test as a field study, we did not measure

	Motion flicker		Noise flicker	
	Low FPS	**Low SNR**	**Low FPS**	**Low SNR**
Animation	(+)	+	(+)	+
Canyon	∘	−	+	∘
12f	∘	(−)	+	∘
36f	(+)	(−)	+	(+)
72f	(+)	(−)	+	(+)
All	(+)	(−)	+	(+)

Legend: + switching preferred, − downscaling preferred
∘ no preference, (*) weak tendency

Table 5.5: HDTV scenario - quality preference for layer switching vs. downscaling.

the environmental lighting in the test room, but the lighting condition was adjusted to avoid incident light being reflected from the screen. We displayed the video clip pairs in two different randomized orders. The duration of a whole continuous test session was 20 minutes and none of the accessors requested break during the test.

5.5.2 Results

In the same way as in the two previous sections, the results of this study are reported with 0.01 confidence intervals. We demonstrate the correlations between the preferences, content genres and switching period lengths in table 5.5.

5.5.2.1 Temporal layer switching

The results found in HDTV test scenario differs significantly from what we found out in mobile test scenarios. The general observation shows a weak tendency towards SNR layer downscaling (Low SNR) than temporal layer switching (Motion flicker), in spite of the significant larger bit-rate reduction given by SNR layer downscaling. There is also a weak tendency of preference for layer switching than downscaling in temporal dimension. A possible conclusion is that people detect frame-rate loss easier and disapprove clearly of low frame-rate when watching HD video.

The effect of the period is weak as there is no significant variations in the data samples between different period levels. However, we noticed that shorter switching period seems to makes the impact of Motion flicker more pronounced than how longer period does. Temporal layer switching got less ratings at shorter period levels no matter whether it was compared to layer downscaling in SNR or temporal dimension. At 12f, it seems that our assessor did not differentiate the motion jerkiness caused by the Motion flicker and Low FPS.

Significant effect was found for the content related factor. When Motion flicker is compared with Low SNR, preferences differ between genres. The majority of our assessors preferred temporal switching over SNR downscaling when watching the Animation video. Watching the Canyon clip, on the other hand, they indicated the opposite preference, which contradicts also all the results from the two mobile FLS studies.

5.5.2.2 SNR layer switching

The observation of layer switching in SNR dimension repeats what we found in temporal dimension. In the HDTV scenario, people seem to be more sensitive to frame-rate changes than quality loss at the picture level. When SNR layer switching (Noise flicker) is compared to temporal downscaling (Low FPS), participant ratings indicate a clear preference towards Noise flicker instead of Low FPS, which is again different than the test results obtained from mobile scenarios. The results are consistent across genres and the preference for Noise flicker applies for different switching periods.

When layer switching is compared with downscaling in the single SNR dimension (Noise flicker against Low SNR), we do not find any significant results except for the animation clip. This confirms previous finding that non-natural texture with high contrast edges are less influenced by flicker effect.

According to the actual number of ratings, we believe the length of switching period affects the perceived video quality in a similar way both in HDTV and mobile scenarios. Namely, if the switching period is below a certain level (flicker threshold), the flicker effect would appear distinctly and impair severely user's visual experience. In this experiment, although the statistic does not indicate any preference at 12f, the Low SNR streaming solution received actually the most number of ratings. Compared to mobile scenarios, the flicker threshold seems to be higher in HDTV scenario.

5.6 Discussion

In this section, we provided an analysis of the perceived quality of FLS and its usefulness to adapt to a given average bandwidth. We also took a critical look at the assessment methods itself.

5.6.1 Range of experiments

We performed three field studies in order to understand whether people who watch video consider it beneficial to adaptively change video quality frequently, and whether the answer to this question changes with the switching frequency. That it is beneficial to exploit available bandwidth to its fullest and adapt video quality quickly to use it, is an assumption that has frequently been made in the past. Through prefetching or buffering on the client side, even coarse- and medium-grained scalable video codecs would be able to come close to exploiting all available bandwidth in the long-term average.

Our investigations considered only options that are available in the toolset of SVC as implemented by the reference encoder. We considered bandwidth changes through temporal scalability and through SNR scalability separately. We investigated only switching patterns where half of the frames belong to an upper and half to a lower operation point. A finer adaptation granularity can be achieved by adaptively tuning this ratio, but the 8-second clip length used in our tests in accordance with the PC approach prevent an exploration of other ratios. When analyzing the results from all three studies, we found that preference indicators depend highly on the scenario.

5.6.2 Mobile devices

In our two field studies that examined mobile devices, we found that temporal switching and also temporal downscaling down to 12 fps result in better subjective quality than any type of SNR layer adaptation. When directly comparing switching versus downscaling in the temporal domain alone, no preference became apparent. Hence, temporal adaptation could be employed at any desired ratio in the observed range between 25 and 12 fps. The reason for this is that human observers regard all frame-rates above a margin of 10 fps as sufficiently smooth, when they watch videos on small devices at typical viewing distances. This observations has been reported by McCarthy et al. (2004) and our earlier study in chapter 4, and was confirmed by the FLS study. Our suggestion from this observation is to use a lower frame rate than 12 fps for the base-layer when encoding videos for mobile devices.

For SNR layer switching, the period length is a crucial design criteria. Very short periods (less than 12f or 0.5 sec) should be avoided, because they introduce annoying flicker effect especially for natural textures. Reluctant quality adaptation such as switching layer every 2 seconds, on the other hand, decreases the bandwidth utilization and is either unacceptable for users.

5.6.3 Applicability of findings

The layer switching pattern must be supported by the SVC encoding structure and synchronized to the decoder operation to avoid prediction errors. The switching patterns used in our study assumed short GOP sizes and frequent intra-updates to allow for short switching periods. Due to inter-frame prediction, switching may not be possible at every frame boundary. Frequent layer switching points are usually in conflict with practical encoder setups that use multiple reference pictures, long GOPs and rare intra-updates for increased coding efficiency. This requires a trade-off at encoding time.

The results of our studies are not limited to layer switching in the coarse-grain encoded versions of H.264-SVC streams alone. Any adaptation strategy in streaming servers, relaying proxies and playout software that can alternate between different quality versions of a video may benefit from our findings.

5.7 Conclusion

We have investigated whether we can achieve fine-grained video scalability using coarse-grained H.264 SVC without introducing the high overhead of MGS in different streaming scenarios including mobile TV and HDTV. This was tested by switching enhancement layers on and off to achieve the target bit-rate between CGS operation points. According to the scaling dimensions, we have identified three types of visual artefacts caused by frequent layer variations, i.e., the noise, blur and motion flicker effects. We tested different switching patterns against different downscaling patterns. The flicker artefacts associated with these switching patterns were compared with regular compression artefacts in downscaled video. Our subjective tests indicate:

- Switching patterns with sufficient perceptual quality exist.

- Human perception of quality impairment in FLS is content and context specific.

At first, the screen size of display devices has significant influence on perceived video quality in FLS. The mobile test scenarios reveal a clear preference of temporal switching over SNR scaling regardless of content and switching period. In our investigation of HD screens, we found nearly the opposite picture. There, people prefer a regular SNR reduction over motion flicker which becomes apparent on large screens even when the frame rate is reduced from 25 fps to 12 fps. The explanation for this can be found in the human visual system. Visual acuity in human's foveal field-of-vision decreases from the center towards the outside while sensitivity to motion activities increases (Rix et al., 1999; Beeharee et al., 2003; Nadenau et al., 2000). The part of vision outside the center of gaze, referred to as peripheral vision, is better at detecting motion than the central vision. Mobile devices are best viewed from 7-9.8 screen heights distance (Knoche and Sasse, 2008), which keeps the entire screen inside the visual focus area. HDTV screens, on the other hand, are best viewed from 3 screen heights distance, where the display still covers most of the human field of vision. This difference influences the minimal required angular resolution of the human eye and foveal field-of-vision (Knoche and Sasse, 2008; McCarthy et al., 2004). Due to the small size of screen, mobile video is usually not perceived by peripheral vision. The complete screen on mobile devices is in the central high acuity region and therefore details are resolved throughout the displayed image at almost the same fidelity. On HDTV screens, the image covers a larger region of the field-of-vision. Hence, humans focus on particular details within the image, which are seen with high acuity, while outer regions of the image cover the temporally sensitive area perceived by peripheral vision. Thus, temporal abnormalities (jerkiness, jumping objects, flicker) are detected much easier and may even be annoying for the viewer.

Given the different quality impacts of screen size, video adaptation strategies for different end devices should be chosen accordingly. For mobile devices, temporal layer switching was shown to perform better than SNR layer downscaling, but not better than temporal downscaling. Hence, when bandwidth adaptation is required, the streamed video can select to first downscale its temporal resolution but no less than 10 $\sim$ 12 fps without introducing perceptual quality degradation. After that, SNR layer switching and downscaling alone can be compared to determine whether FLS should be applied for additional bandwidth saving. In the SNR dimension, the comparison of layer switching and downscaling on mobile devices shows that SNR layer switching with an 2frames (80ms) period leads to a visually disturbing flicker effect, while switching above a 72frames (3seconds) period is not clearly preferable to downscaling. Between these points, however, SNR layer switching, and thus FLS, has a beneficial effect that grows until a period length of 48-frames (2seconds). For large screens, adaptation in the temporal dimension was deemed to be generally undesirable comparing to SNR layer adaptation. Therefore, SNR layer adaptation should be given higher priority for bandwidth adaptation. When switching layer in the SNR dimension, short switching periods below 24frames (1second) should also be avoided to prevent annoying flicker effect.

In terms of resource consumption, both the temporal switching pattern (figure 5.1e) and SNR switching pattern (figure 5.1d) can achieve bit-rates between the encoded SVC base layer and the enhancement layer. Both switching patterns were preferred over the SNR downscaled operation point (figure 5.1b). Thus, we claim that such fine grained adaptation is possible in different scenarios. FLS is mostly suitable for video content

characterized by high contrast edges, artificial texture and movements.

However, based on our preliminary tests, we cannot say which switching pattern will give the *best* possible result. We need to investigate further how the detectability and acceptability of flicker is related to various kinds of factors such as flicker threshold and bit-rate intervals between high and low switching points. To investigate these factors, an additional subjective study has been performed, and we introduce the study in chapter 6. In practice, popular HD videos are not only streamed to large displays, but also can be watched on displays with smaller size. Additional studies can be done to investigate if the same temporal downscaling strategy also applies to HD video on smaller screens. At this point, we tested also only clips without scene changes. To further limit the perceived quality degradation of switching techniques, scene changes can for example be used as switching points.

Chapter 6

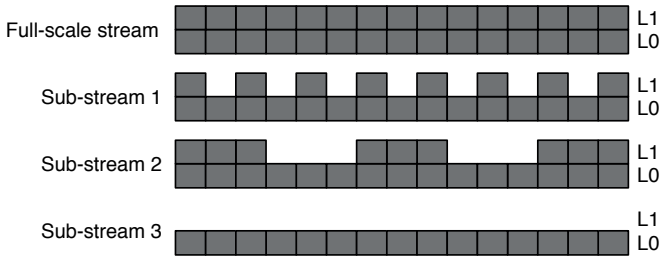
Flicker effects

This chapter reports on an in-depth investigation of the three types of flicker artefacts that are specific for frequent bandwidth adaptation scenarios.

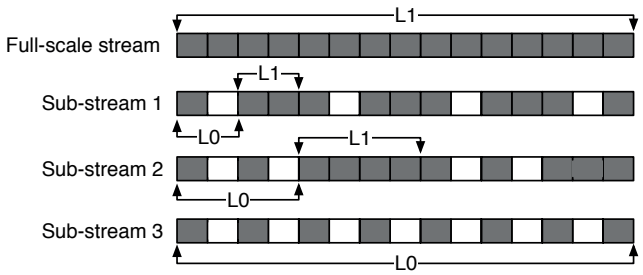
6.1 Introduction

In the previous chapter, we presented the FLS study that investigated the quality impacts of frequent layer variation on scalable video. It was shown that active layer switching may create *flicker artefacts* that specifically come from repeated quality fluctuations in the streamed video. The *flicker artefact* degrades usually the experienced subjective quality. However, the subjective tests included in the FLS study indicate that noise and motion flicker can not generally be considered deficiencies. In practice, active video adaptation to changes in available bandwidth is generally preferable to random packet loss or stalling streams, and not every quality change is perceived as a flicker effect. Essentially, the perceptual effect of flicker is closely related to the adaptation pattern, which is usually characterized by *amplitude* and *frequency* of the quality changes. So, the question remains how to make adaptation patterns to control these flicker artefacts in order to ensure sufficient video quality. With this question in mind, this chapter, which extends the paper of Ni et al. (2011a) with more analyses and conclusions, explores the acceptability of flicker for a handheld scenario. The original paper is included in Appendix E.

In figure 6.1, we show sketches of simple streaming patterns for both spatial and temporal scaling. Figure 6.1a depicts a video stream encoded in two layers; it consists of several subsequent segments, where each segment has a duration of t frames. The full-scale stream contains two layers (L0 and L1), and the low quality stream (sub-stream 3) contains only the lower layer (L0), it is missing the complete L1 layer. For these, the number of layers remains the same for the entire depicted duration, meaning that neither of the two streams flickers. The other two examples show video streams with flicker. The *amplitude* is a change in the spatial dimension, in this example the size of the L1 layer (in other scenarios, this may be the number of layers). The *frequency* determines the quality change period, i.e., how often the flicker effect repeats itself. In this example, sub-stream 1 changes its picture quality every t frames (2 blocks in the figure), whereas sub-stream 2 changes every $3t$ frames (6 blocks in the figure). Figure 6.1b shows a similar example of how the amplitude and frequency affect the streaming patterns in the temporal dimension. Here, the *amplitude* is a change in the temporal dimension (frame rate). In this example,



(a) Scaling in spatial dimension



(b) Scaling in temporal dimension

Figure 6.1: Illustration of streaming patterns for scalable video.

we index video segments by their temporal resolutions since only temporal scalability is our concern. The full-scale stream can be displayed at a normal frame rate. Sub-stream 3 drops frames regularly and can be displayed at a constant low frame-rate. Neither of the two streams flickers in the temporal dimension. Hence, we say that the full-scale stream contains layer L1, whereas sub-stream 3 contains only layer L0. Sub-stream 1 and 2 halve the normal frame-rate at a regular interval of $2t$ and $4t$ time units, respectively. Therefore, the layer variations in sub-streams 1 and 2 have the same amplitude, but the changes appear at different frequencies.

To provide the best possible video quality for a given available bandwidth, the applications need to select the most suitable options from several streaming patterns. Considering the alternatives in figures 6.1a and 6.1b, three sub-stream alternatives can be used if the full quality stream cannot be provided. Therefore, to get a better understanding of human quality perception of *flicker*, we have performed a subjective field study with a special focus on handheld devices. We have considered state-of-the-market encoding techniques represented by the H.264 series of standards. Our goals are to evaluate the influence of the main influential factors on acceptability, and to find the range of these factors' levels. With these answers we hope to minimize the flicker effect in layer variation.

6.2 Experiment Design

6.2.1 Randomized Block Design

We conduct subjective experiments to explore the impact of noise, blur and motion flicker on the perception of video quality. In addition to the three different adaptation domains (SNR for noise flicker, spatial resolution for blur flicker and temporal resolution for motion flicker), the overall video quality is influenced by other factors including amplitude, frequency and content characteristics (see section 6.2.2). All of these are design factors studied in our experiment. We do not limit ourselves to a single genre of video content, but we do not aspire to cover all semantic categories. We explore four content types, which are selected as representatives for extreme values of low and high spatial and temporal information content. In our experiments, the subjects are asked to rate their acceptance of the overall video quality. Due to the fluctuating state of videos that flicker, we predict flicker to be perceived differently than other artefacts. We add a Boolean score on perceived stability, which we expect to provide us with more insight into the nature of the flicker effects (see section 6.2.4). Finally, we measure participants' response time, which is the time between the end of a video presentation and the time when they provide their response.

The repeated measures design (Coolican, 2004) of these experiments ensures that each subject is presented with all stimuli. The repeated measures design offers two major advantages: First, it provides more data from fewer people than, e.g., Randomized Pairwise Comparison (R/PC) studies. Second, it makes it possible to identify the variation in scores due to individual differences as error terms. Thus, it provides more reliable data for further analysis. This study employs an alternative to the traditional full factorial repeated-measures design that is called Randomized Block Design (Mason et al., 2003). It blocks stimuli according to flicker type and amplitude level. A stimuli block consists of a subset of test stimuli that share some common factor levels and can be examined and analyzed alone. Stimuli are randomized within each block and blocks are randomized to an extent that relies solely on the participant, as they are free to choose which block to proceed with.

The randomization of stimuli levels ensures that potential learning effects are distributed across the entire selection of video contents and frequency levels, and, to a degree, also amplitudes and flicker type. Moreover, we hope to minimize the effect of fatigue and loss of focus by dividing stimuli into smaller blocks and allowing participants to complete as many blocks as they wish, with optional pauses between blocks.

6.2.2 Content Selection and Preparation

As the rate distortion performance of compressed video depends largely on the spatial and temporal complexity of the content, the flicker effect is explored across four content types at different extremes. Video content is classified as being high or low in spatial and temporal complexity, as recommended by International Telecommunications Union (1999) and measured by spatial information (SI) and temporal information (TI) metrics, respectively. Four content types with different levels of motion and detail are selected based on the metrics (figure 6.2). To keep the region of interest more global and less

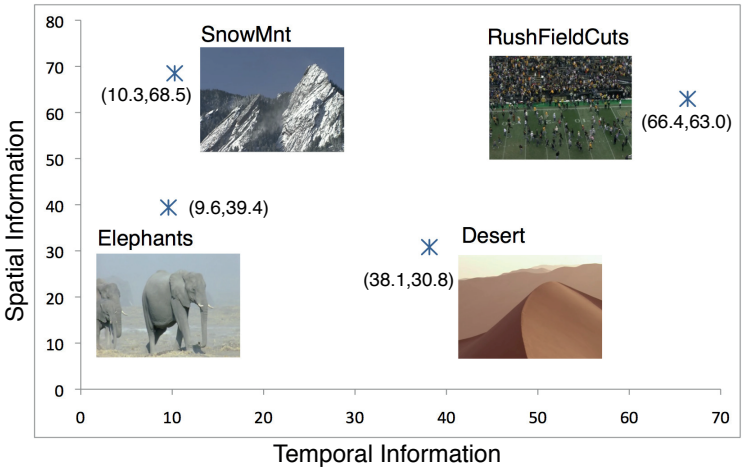


Figure 6.2: Test sequences.

focused on specific objects, we avoid videos with dictated points of interest, such as a person speaking. We avoid also video with artificial textures and focus video with natural textures instead. It is beyond the scope of the current investigation to generalize the results to all video content.

Raw video material is encoded using the H.264-SVC reference software, JSVM 9.19, with two-layer streams generated for each type of flicker, as portrayed in figure 6.1. The amplitude levels of the layer variations are thus decided by the amount of impairment that separates the two layers. Table 6.1 summarizes the factor levels of amplitude, frequency, and content, according to the different flicker stimuli, noise, blur, and motion. For noise flicker stimuli, constant quantization parameters (QP) are used to encode a base layer L0 and an enhancement layer L1. Since the latter is encoded with QP24 for all test sequences, the amplitude levels and variations in video quality are represented by the QPs applied to L0 for noise flicker stimuli. Similarly, with blur flicker stimuli, amplitude is represented by spatial downscaling in L0, and temporal downscaling in L0 defines the amplitude for motion flicker stimuli.

To simulate the different flicker effects that can arise in streamed video, video segments from the two layers are alternately concatenated. Different frequencies of layer variation are obtained by adjusting the duration of the segments. For simplicity, we assume periodic layer switching scheme, and set the duration constant for the segments in a video sequence. Therefore we use period to refer to the segment duration. Corresponding to six frequency levels, six periods in terms of the L1 frame-rate are selected, which include 6, 10, 30, 60, 90 and 180 frames for both noise and blur flicker stimuli. Since short durations for changes in frame-rate are known to lead to low acceptance scores (Ni et al., 2010), the periods for motion flicker stimuli are limited to 30, 60, 90 and 180 frames.

a) Noise flicker		
Amplitude	L1	QP24
	L0	QP28, QP32, QP36, QP40
Period		6f, 10f, 30f, 60f, 90f, 180f
Content		RushFieldCuts, SnowMnt, Desert, Elephants

b) Blur flicker		
Amplitude	L1	480x320
	L0	240x160, 120x80
Period		6f, 10f, 30f, 60f, 90f, 180f
Content		RushFieldCuts, SnowMnt, Desert, Elephants

c) Motion flicker		
Amplitude	L1	30fps
	L0	15fps, 10fps, 5fps, 3fps
Period		30f, 60f, 90f, 180f
Content		RushFieldCuts, SnowMnt, Desert, Elephants

Table 6.1: Selection of factor levels

6.2.3 Participants

In total, 28 participants (9 female, 19 male) were recruited at the University of Oslo, with ages ranging from 19 to 41 years (mean 24). They volunteered by responding to posters on campus with monetary compensation rewarded to all. Every participant reported normal or corrected to normal vision.

6.2.4 Procedure

This field study was conducted in one of the University of Oslo's library with videos presented on 3.5-inch iPhone of 480x320 resolution and brightness levels at 50%. Participants were free to choose a seat among the available lounge chairs but were asked to avoid any sunlight. They were told to hold the device at a comfortable viewing distance and to select one of the video blocks to commence the experiment. The 12-second long video segments were presented as single-stimulus events, in accordance with the ITU-T Absolute Category Rating method (International Telecommunications Union, 1999). Each video stimulus was displayed only once. Video segments were followed by two response tasks, with responses made by tapping the appropriate option on-screen. For the first, participants had to evaluate the perceived stability of the video quality by answering "yes" or "no" to the statement "I think the video quality was at a stable level". The second involved an evaluation of their acceptance of the video quality, where they had to indicate their agreement to the statement "I accept the overall quality of the video" on a balanced 5-point Likert scale. The Likert scale includes a neutral element in the center and two

opposite extreme values at both ends. A positive value can be interpreted as an acceptable quality level, a neutral score means undecidedness, while a negative score indicates an unacceptable quality level. Upon completion of a block, participants could end their participation, have a short break, or proceed immediately to the next block. Participants spent between 1.5 and 2 hours to complete the experiment.

6.3 Data analysis

6.3.1 Method of Analysis

The current study explores the influence of amplitude and frequency of video quality shifts for three types of flicker stimuli, noise, blur and motion, as well as video content characteristics, on the perception of stability, the acceptance of video quality and response time. Control stimuli with constant high or low quality are included as references to establish baselines for the scores provided by participants. Stability scores and rating scores are processed separately, grouped according to flicker type. Thus, responses are analyzed in six different groups, with control stimuli included in all of them. Since the perception of stability relies on detection, scores are binary and are assigned the value "1" for perceived stability of quality, and the value "0" for the opposite. Rating scores are assigned values ranging from -2 to 2, where "2" represents the highest acceptance, "0" the neutral element, and "-2" the lowest acceptance.

Consistency of acceptance scores is evaluated by comparing scores for control stimuli of constant high or low quality. Whenever a low quality stimulus scores better than the corresponding high quality stimulus, this is counted as a conflict. Conflicts are added up for each participant. If the acceptable number of conflicting responses is exceeded, the participant is excluded as an outlier. An acceptable number of conflicts stays within 1.5 times the interquartile range around the mean as suggested by Frigge et al. (1989); Coolican (2004). For the blur stimuli group, this excluded two participants (12.5%), two for the motion stimuli group (10.5%), and none for the noise stimuli group.

The consistency of response times is also evaluated in order to eliminate results that reflect instances in which participants may have been distracted or taken a short break. Thus, any response time above three standard deviations of a participant's mean is not included in the following analyses.

Stability scores are analyzed as ratios and binomial tests are applied to establish statistical significance. As for acceptance scores, these are ordinal in nature and are not assumed to be continuous and normally distributed. They are therefore analyzed with the non-parametric Friedman's chi-square test (Sheldon et al., 1996). The Friedman test is the best alternative to the parametric repeated-measures ANOVA (Howell, 2002), which relies on the assumption of normal distribution; it uses ranks to assess the differences between means for multiple factors across individuals. Main effects are explored with multiple Friedman's chi-square tests, applied to data sets that are collapsed across factors. Confidence intervals are calculated in order to further investigate the revealed main effects, assessing the relations between factor levels. Multiple comparisons typically require adjustments to significance levels, such as the Bonferroni correction. Yet, such adjustments can increase the occurrence of Type II errors, thus increasing the chances of rejecting a valid difference (Perneger, 1998). In light of this, we avoid the use of adjust-

a) Period

Options	Stable	Unstable	P-value	Signif.
HQ	95.3%	04.7%	2.04e-71	+
6f	30.6%	69.4%	3.32e-12	–
10f	30.0%	70.0%	6.18e-13	–
30f	30.3%	69.7%	1.44e-12	–
60f	31.6%	68.4%	3.71e-11	–
90f	32.5%	67.5%	3.65e-10	–
180f	41.2%	58.8%	0.002	–
LQ	71.3%	28.7%	1.80e-14	+

b) Amplitude

Options	Stable	Unstable	P-value	Signif.
QP28	65.8%	34.2%	3.66e-12	+
QP32	27.7%	72.3%	4.49e-23	–
QP36	21.7%	78.3%	3.51e-37	–
QP40	15.6%	84.4%	8.74e-56	–

Table 6.2: Perceived quality stability for Noise flicker (+ Stable, - Unstable, (*) not significant), HQ = constant high quality, LQ = constant low quality.

ments and instead report significant results without corrections. This procedure requires caution; we avoid drawing definite conclusions and leave our results open to interpretation. Repeated-measures ANOVA tests are finally introduced when analyzing response times.

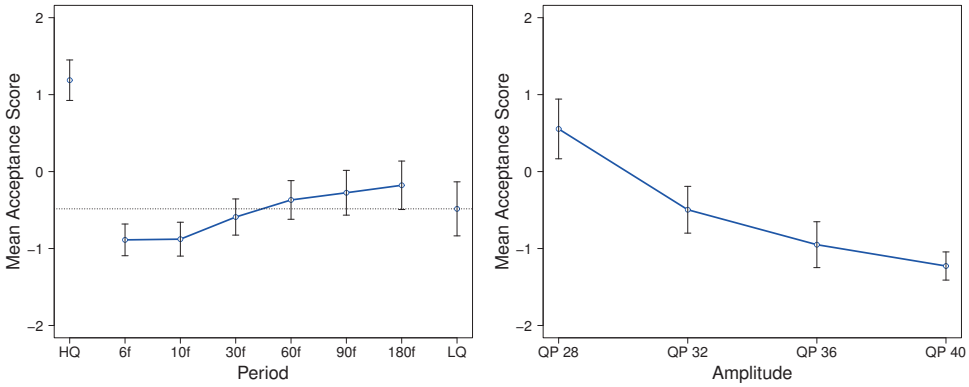
6.3.2 Response Times

None of the repeated-measures ANOVA tests reveals any effect of amplitude, frequency or content on response time, for any type of flicker. In fact, response times seem to vary randomly across most stimuli levels. Possibly, this may be related to individual effort in detecting stability. If so, the video quality variation did not increase the decision-making effort. We may even surmise that participants evaluated the stability of video quality with a fair degree of confidence.

6.3.3 Noise Flicker Effects

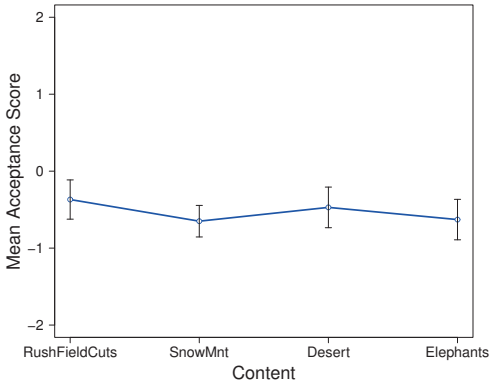
The perceived stability of noise flicker stimuli is generally low and varies little over the different periods, as seen in table 6.2(a). However, the response percentage reflecting stable video quality is slightly higher for video segments of 180 frames. A significantly larger share of responses for the control stimuli reports video quality to be stable, as opposed to unstable, refer to the top and bottom lines in table 6.2(a). Due to the small difference between layers for QP28, it is plausible that the vast majority of participants do not perceive the flicker effect, which would explain why two thirds report stable quality, see the top line in table 6.2(b). Meanwhile, the higher rate of reported stability for non-flicker stimuli fits well with predictions. It indicates that participants detect and identify flicker as instability, whereas constant quality is experienced as stable, even when it is poor.

Main effects are found with Friedman's chi-square tests for period ($\chi^2(5) = 69.25, p<.001$), amplitude ($\chi^2(3) = 47.98, p<.001$) and content ($\chi^2(3) = 27.75, p<.001$). The means and



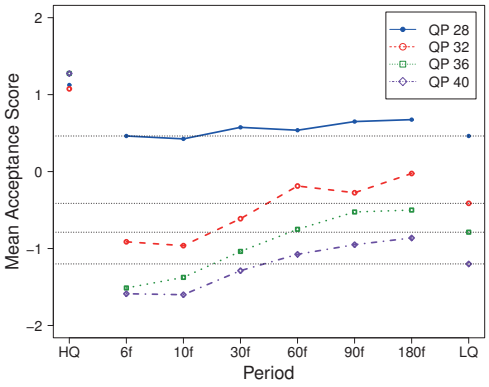
(a)

(b)

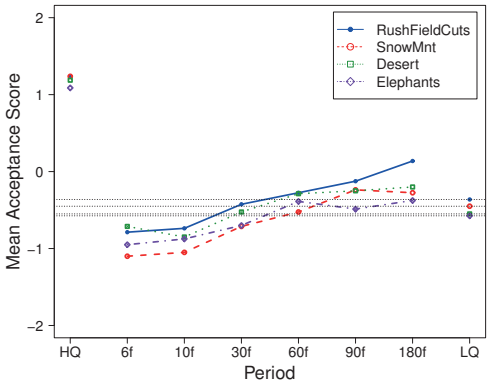


(c)

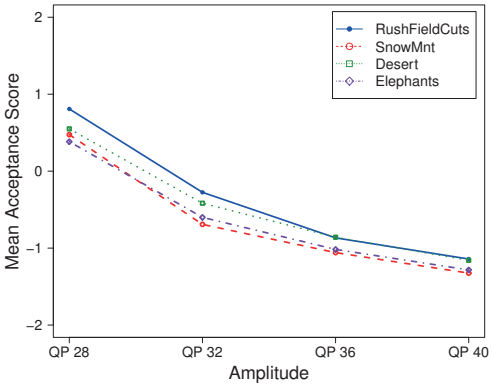
Figure 6.3: Effects of period, amplitude and content on Noise flicker stimuli. Error bars represent 95% confidence intervals.



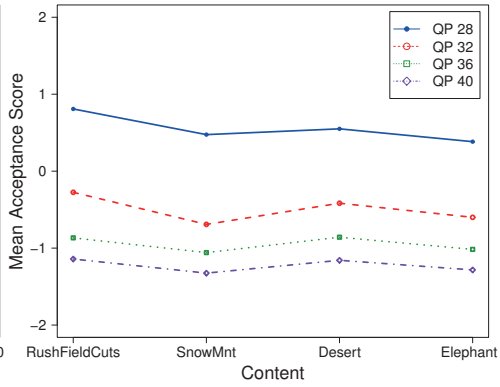
(a) Period effects per amplitude level



(b) Period effects per content type



(c) Amplitude effects per content type



(d) Content effects per amplitude level

Figure 6.4: Explored interactions between influential factors of Noise flicker. (HQ = constant high quality, LQ = constant low quality)

a) Period

Options	Stable	Unstable	P-value	Signif.
HQ	100%	00.0%	3.85e-34	+
6f	11.6%	88.4%	1.50e-17	–
10f	11.6%	88.4%	1.50e-17	–
30f	11.6%	88.4%	1.50e-17	–
60f	13.4%	86.6%	7.12e-16	–
90f	12.5%	87.5%	1.08e-16	–
180f	17.0%	83.0%	6.75e-13	–
LQ	81.2%	18.8%	1.42e-11	+

b) Amplitude

Options	Stable	Unstable	P-value	Signif.
240x160	19.3%	80.7%	4.89e-31	–
120x80	06.6%	93.5%	2.57e-67	–

Table 6.3: Perceived quality stability for Blur flicker (+ Stable, - Unstable, (*) not significant).

confidence intervals presented in figure 6.3a show that acceptance scores become increasingly higher than the constant low quality controls for periods of 60 frames and above. Figure 6.3b displays the decrease in acceptance with larger amplitudes, while figure 6.3c shows only small variations in acceptance scores depending on content type. The effect of content factor is shown only in the case of low amplitudes such as QP28 and QP32, as illustrated in figure 6.4d. As for potential interactions, figure 6.4 illustrates how mean acceptance scores vary across levels of multiple factors. It shows in Figure 6.4c,6.4b that the mean acceptance scores tend to increase as amplitude decreases or period increases, irrespective of the content factor. However, the scores in figure 6.4a point to noteworthy interactions between period and amplitude.

6.3.4 Blur Flicker Effects

For blur flicker stimuli, perceived video quality stability is again low across the different periods, accompanied by high perceived stability ratios for control stimuli, summarized in table 6.3(a). Furthermore, participants tend to judge the video quality as unstable at both amplitude 240x160 and amplitude 120x80, see table 6.3(b). This is also consistent with expectations, suggesting again that flicker is detectable and perceived to be unstable.

Friedman's chi-square tests reveal main effects for period ($\chi^2(6) = 41.79, p<.001$), amplitude ($\chi^2(1) = 14.00, p<.001$) and content ($\chi^2(3) = 33.80, p<.001$). Similar to noise flicker, mean acceptance scores tend to increase as as amplitude decreases or period increases as seen in figures 6.5a and 6.5b. But, the mean acceptance scores are generally low across period and amplitude levels. Frequent changes of spatial resolution seems to be unacceptable. Only at 60 frames and above the mean acceptance scores of fluctuated quality approach the acceptance of constant low quality. Compared to noise flicker, larger effect of content factor was manifest on blur flicker. As shown in figure 6.5c, acceptance scores for the Desert and Elephants clips appear to be higher than the RushFieldCuts and SnowMnt clips.

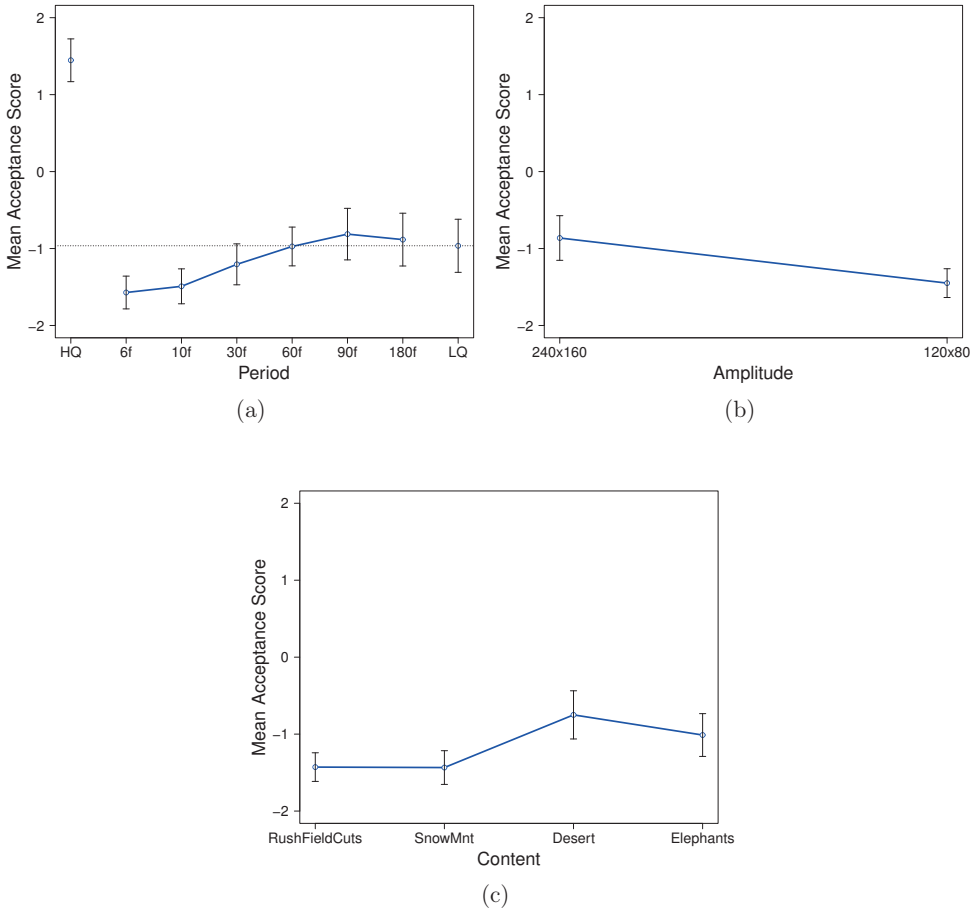
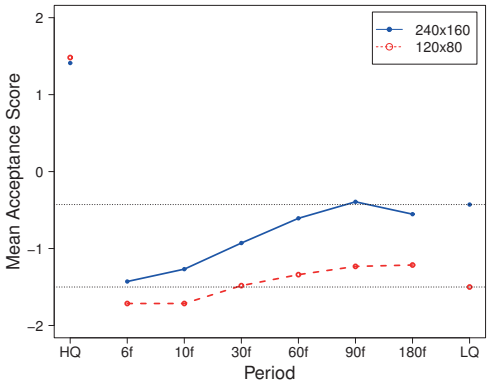
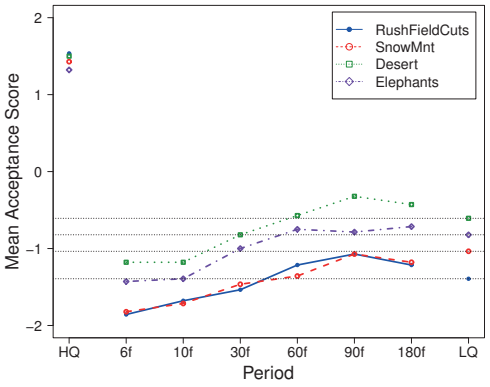


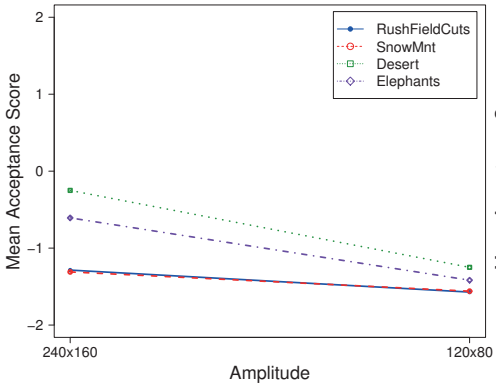
Figure 6.5: Effects of period, amplitude and content on Blur flicker. Error bars represent 95% confidence intervals.



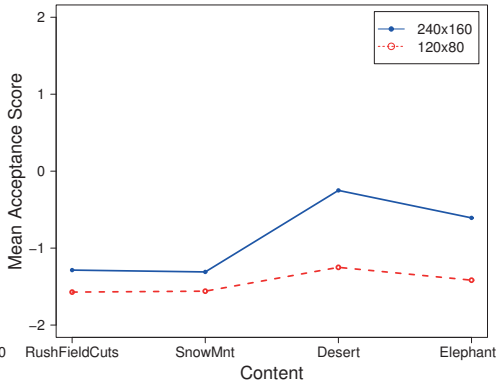
(a) Period effects according to amplitude



(b) Period effects according to content



(c) Amplitude effects according to content



(d) Content effects according to amplitude

Figure 6.6: Explored interactions for Blur flicker. (HQ = constant high quality, LQ = constant low quality)

a) Period

Options	Stable	Unstable	P-value	Signif.
HQ	90.8%	09.2%	4.43e-47	+
30f	14.3%	85.7%	7.85e-35	–
60f	16.2%	83.8%	4.08e-31	–
90f	18.0%	82.0%	1.08e-27	–
180f	20.6%	79.4%	2.44e-23	–
LQ	40.8%	59.2%	0.0029	–

b) Amplitude

Options	Stable	Unstable	P-value	Signif.
15fps	43.8%	56.2%	0.045	(*)
10fps	15.1%	84.9%	2.62e-33	–
5fps	07.4%	92.6%	2.82e-52	–
3fps	02.9%	97.1%	1.82e-67	–

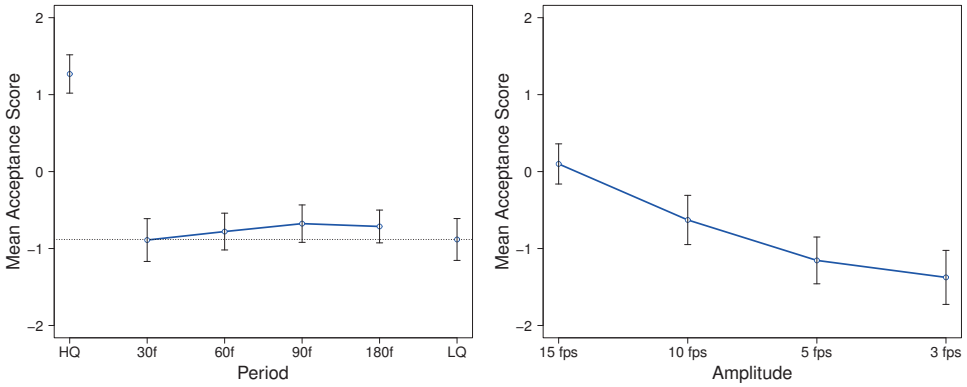
Table 6.4: Perceived quality stability for Motion flicker (+ Stable, - Unstable, (*) not significant).

Figure 6.6 further illustrates the effect of each influential factor when considering interactions. Period seems to have larger effect when the amplitude is smaller as shown in figure 6.6a. Similarly, the interaction of amplitude and content shows more markedly by the larger variation between the different content groups at resolution 240x160, as seen in figures 6.6c and 6.6d. However, we note that changing resolution to 240x160 is large enough to produce detectable flicker artefact, as reported in table 6.3(b).

6.3.5 Motion Flicker Effects

Low perceived stability ratios are evident across all periods for motion flicker stimuli, presented in table 6.4(a). As expected, the vast majority of participants think that the video quality is stable for constant high quality control stimuli but not for constant low quality; there are more responses that correspond to perceived instability for low quality control stimuli. This is potentially explained by the lack of fluency of movement that occurs at lower frame-rates. The stability scores for amplitude may also reflect a bias towards reporting jerkiness as instability, as listed in table 6.4. However, stability is reported more frequently for larger periods and better frame-rates; this indicates influences from both period and amplitude on perceived quality stability.

Friedman's chi-square tests uncover main effects for all factors, including period ($\chi^2(3) = 7.82$, $p < .05$), amplitude ($\chi^2(3) = 41.62$, $p < .001$), and content ($\chi^2(3) = 27.51$, $p < .001$). However, the main effect for period is very close to the significance threshold (p=0.0499), which is likely the reason for the relatively flat distribution of acceptance scores observed in figure 6.7a. Amplitude and content type, on the other hand, have larger effects on quality acceptance, as seen in figures 6.7b, 6.7c and 6.8. The effect of content appears most significant at 10 fps.



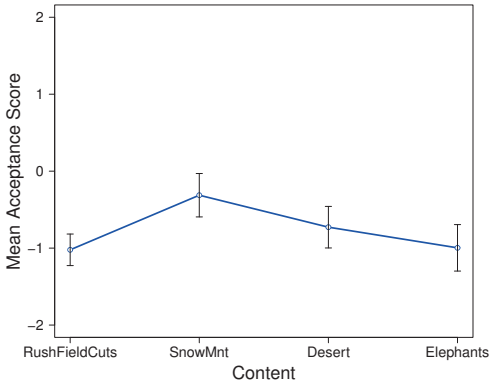
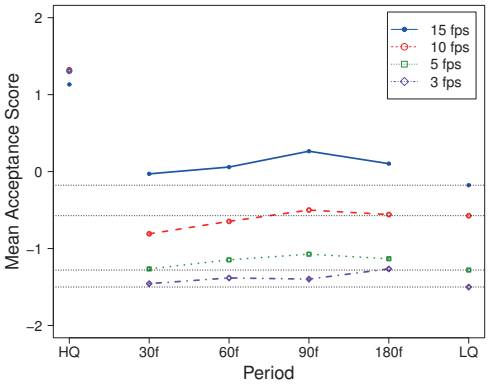
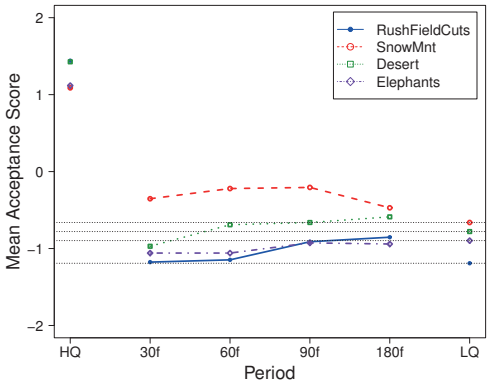


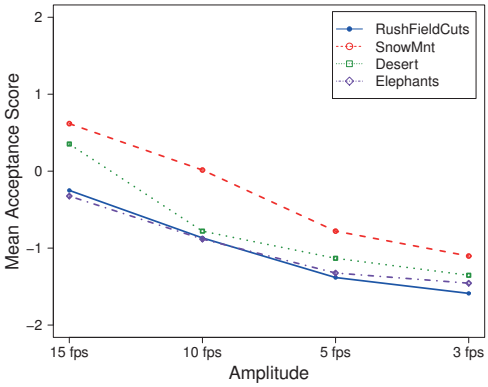
Figure 6.7: Effects of period, amplitude and content on Motion flicker stimuli. Error bars represent 95% confidence intervals.



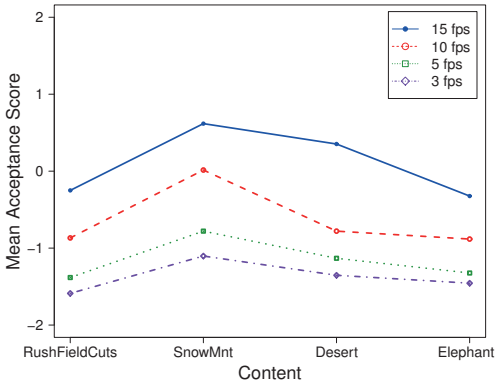
(a) Period effects according to amplitude



(b) Period effects according to content



(c) Amplitude effects according to content



(d) Content effects according to amplitude

Figure 6.8: Explored interactions for Motion flicker. (HQ = constant high quality, LQ = constant low quality)

6.4 Discussion

6.4.1 Period Effect

The period of flicker is a major influential factor for flicker in the spatial dimension. Significant differences between acceptance scores given to different periods in noise flicker can be found in figure 6.3a, and for blur flicker in figure 6.5a. In figures 6.6a and 6.6b, we can highlight three period ranges that influence the overall quality acceptance: low acceptance for short periods, acceptance higher than the low-quality control stimuli for moderate periods, and stagnating for long periods. Stagnation is less pronounced in figures 6.4a and 6.4b.

In figure 6.4b, the average across all amplitudes is shown for individual contents, reinforcing that the effect is independent of the content. At high frequencies ($< 30f$ corresponding to $< 1sec.$), the flicker is perceived as more annoying than constant low quality for all different content types. Starting at moderate frequencies ($30 \sim 60f$ or $1 \sim 2s$), the quality is considered as better than a constant low quality for some content types. At low frequencies ($> 60f$ or $> 2s$), it is more or less established that the quality of a flicker video is in most cases regarded as better than a constant low quality. For both flicker types in the spatial dimension, this is significant across amplitudes (figures 6.4a and 6.6a), content (figures 6.4b and 6.6b), but counter-examples exist (see the top line in figure 6.6a).

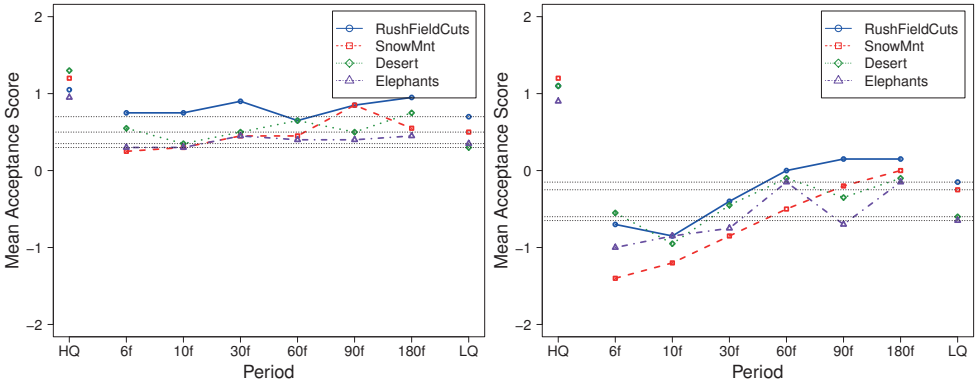
In the temporal dimension, the period does not seem to have a significant influence on the motion flicker. There are only small differences between acceptance scores for different periods, ranging from $30f$ to $180f$ (see figures 6.7a, 6.8a and 6.8b). When the amplitude of temporal downscaling is small, scores are higher than for the low-quality control stimuli (figures 6.8a, 6.10a). No period ranges can be highlighted.

A general observation for all three flicker types is that adaptive video streaming can outperform constant low quality streams, but the switching period must be considered in relation to the flicker amplitudes.

6.4.2 Amplitude Effect

The amplitude is the most dominant factor for the perception of flicker. This seems reasonable since the quality differences become more apparent with increasing amplitude when alternating between two quality versions. Our statistical results, presented in section 6.3, show this and evaluate the strength of the influence. At low amplitude where visual artefacts are less obvious, the flicker effect can be unnoticeable for the majority of our participants, but the detectability of quality fluctuation grows with the increase of flicker amplitudes for all three types of flicker. The period effect becomes significant only if the flicker effects are detectable from the increase of flicker amplitude. It is possible to obtain a benefit by choosing a suitable period for SNR and resolution variation, but it seems that only amplitude is critical for frame-rate variation.

We asked the assessors in our experiments to report whether they perceived the quality of the presented videos to be stable or not. Stability scores serve as measures of amplitude thresholds above which flicker effects become perceptible. With 65.8% of 480 responses (see Q28 in table 6.2(b)) agreeing on stable quality when QPs differ by 4 or less, the



(a) Amplitude = QP28 (b) Amplitude = QP32

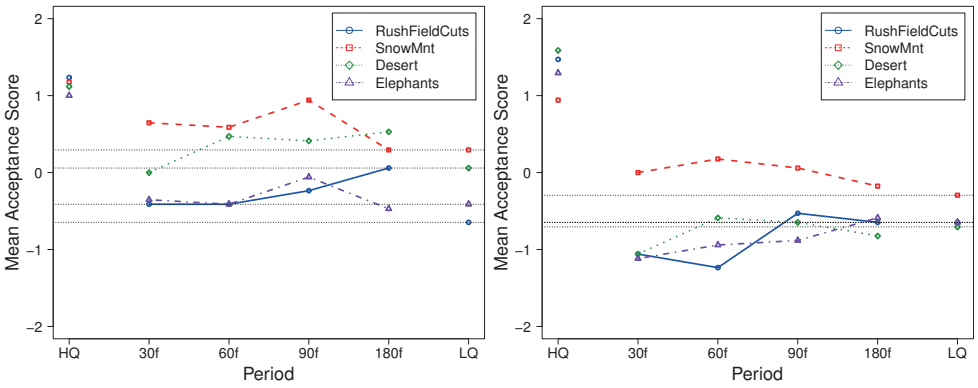
Figure 6.9: Mean acceptance scores for two top amplitude levels in Noise flicker. (HQ = constant high quality, LQ = constant low quality)

amplitude threshold is established at 4 QP levels for noise flicker. As for motion flicker, the amplitude threshold is set to a 50% reduction of frame-rate, with consensus on stability for 56% of the 272 responses for adaptations varying between 15 and 30 fps. For the blur flicker, however, higher resolution than 240x160 has not been evaluated in this study. Nevertheless, the amplitude threshold should be no more than 50% reduction.

In this study, we kept the higher quality layer at a constant level while changing the lower quality layer to different levels, therefore the amplitude level represents also the total amount of visual artefacts in a video. Large amplitude levels may result in severe visual distortion and make the video quality unacceptable. As shown in figures 6.9 and 6.10, frequent adaptation with amplitude of more than 4 QP differences or 50% frame-rate reduction may be generally rated as unacceptable for all content types. For blur flicker, user experience of watching up-scaled video that was originally half or a quarter of the native display resolution of a handheld device turned out to yield low acceptance. Given the fact that our content is chosen from a wide range of spatial and temporal complexities (figure 6.2), this indicates that the change of spatial resolution should not exceed half the original size in order to deliver a generally acceptable quality. Further investigations are necessary to find amplitude thresholds for blur flicker.

6.4.3 Content Effect

Content seems to play a minor role for flicker, but its effect varies across different flicker types. For noise flicker, the effect of content is not significant (figure 6.3c). We observe weak interaction effects between period and content (figure 6.4b), but no interaction between amplitude and content. In figure 6.4d, we see that the acceptance scores vary only slightly between content for the noise flicker although the chosen amplitudes cover a large part of the scale. However, a significant effect of content can be found in both blur and motion flicker (figures 6.5c and 6.7c). Content interacts slightly with amplitude as well. For blur flicker, the Desert and Elephant sequences get significantly different scores than



(a) Amplitude = 15 fps (b) Amplitude = 10 fps

Figure 6.10: Mean acceptance scores for two top amplitude levels in Motion flicker. (HQ = constant high quality, LQ = constant low quality)

RushFieldCuts and SnowMnt, see figure 6.6d. For motion flicker, the SnowMnt sequence is least influenced by the loss of frame-rate and always has significantly higher scores, see figures 6.8b, 6.8d and 6.10. The observation means different content characteristics can influence the perception of flicker.

The SnowMnt and RushFieldCuts sequences have more complex texture details then the other two content types and are therefore more strongly affected by the loss of spatial resolution. Additionally, SnowMnt contains significantly less motion; half of the sequence moves slowly around the snow mountain at fairly constant distance. The lack of relative movement between objects in the scene may limit the visible effect of frame dropping. However, video classification based only on two simple metrics of spatial and temporal information does not cover enough content features that are related to human perception. Region of interest, the scope and direction of motion etc. may also have influences on visual experience. In our experiments, 15 fps has the effect that the scores for two test sequences are on the negative part of the scale (see figure 6.10a), while the two sequences have quite different temporal complexity according to the TI metric, introduced in section 6.2. More advanced feature analysis is needed for further explanation of these phenomena.

6.4.4 Applicability of the Results

The discussion in section 6.4.1 highlights the detrimental effects of high-frequency quality fluctuations on perceived quality of video streams. Short switching periods are therefore not recommended for practical bit-rate adaptation schemes. To explore the differential effects of the remaining streaming options, we examine again the experimental results after filtering out period levels lower than 2 seconds.

In our study, a bit-rate adaptation scheme is represented by a type of quality down-scaling operation and the magnitude of quality change. Beside the adaptation schemes that actively switch between multiple video layers, streaming only a single layer without switching can be also regarded as a special case of adaptation scheme with zero magnitude

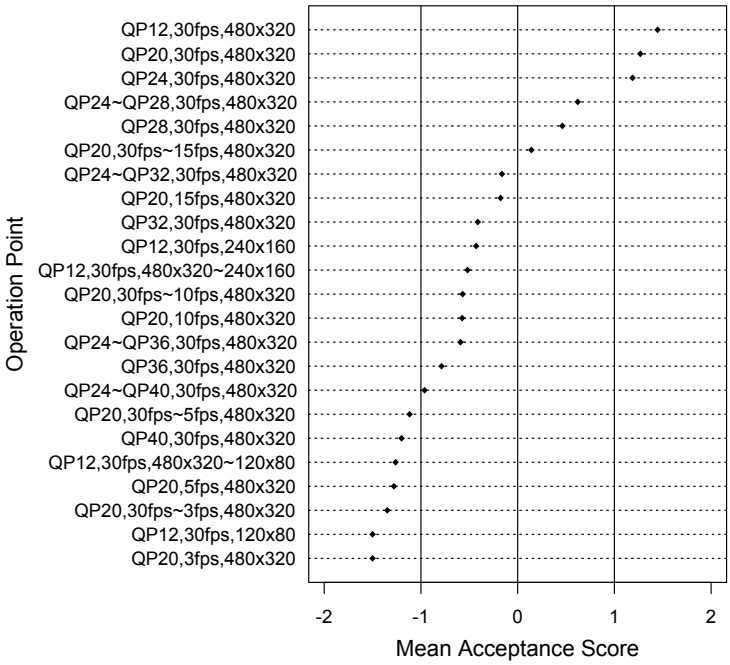


Figure 6.11: Bit-rate adaptation schemes ranked according to the mean acceptance scores. Scores are averaged across quality-shift periods of 2, 3 and 6 seconds, excluding the shorter periods perceived as flicker.

of quality change. Comparing to active layer switching, streaming only a single sub-layer provides constant quality at lower bit-rates. To make an overall comparison between these schemes, we sort these schemes according to their mean acceptance scores, as illustrated in figure 6.11.

The overall picture shows that quality switching options have almost always higher mean acceptance scores than their respective single low layer options, regardless of the amplitude of changes and the adaptation dimension. But, because that large amplitudes were used in our experiments, the acceptance scores for many switching options in figure 6.11 are not sufficiently high. Higher acceptance scores may be obtained if we introduce an intermedia quality layer in between the two layers of any switching options.

The flicker amplitude and frequency determine the severity of flicker effect. When flicker effect is eliminated, the acceptance of overall quality is mainly related to the magnitudes of quality degradation, which reflect the amount of quality artefacts created during encoding process. In this respect, figure 6.12 depicts scores and statistics for these single layer streaming options, arranged from highest to lowest mean acceptance score. As seen from the figure, median and mean acceptance scores are below neutral for all adaptations

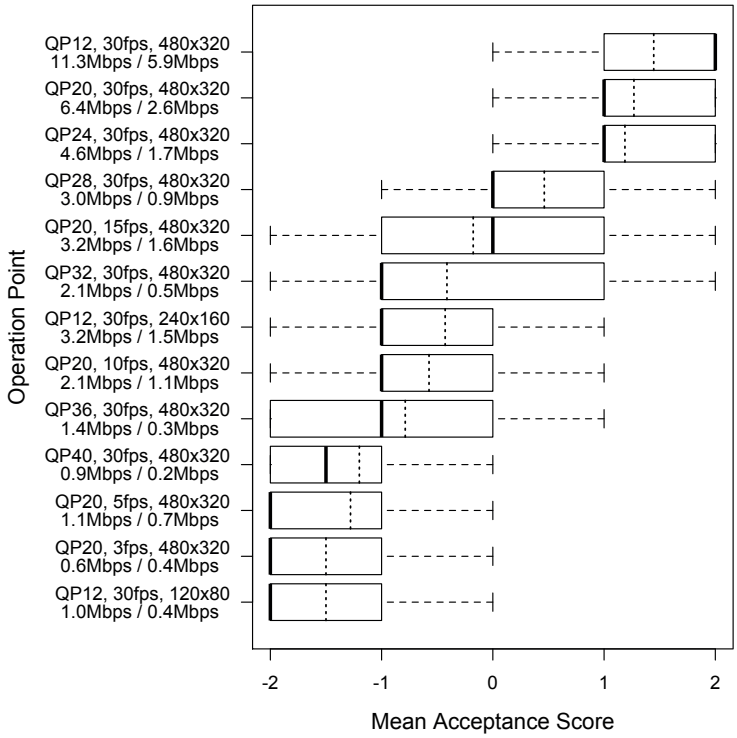


Figure 6.12: Box plot of acceptance scores for compression, resolution, and frame-rate adaptations. The central box spans the interquartile range, with minimum and maximum scores illustrated by "whiskers" to the left and right. Within the box, the bold line corresponds to the median, whereas the dotted line represents the mean acceptance score. The resulting bit-rates are also included for each step. The first bit-rate is when using I-frames only, which is used in the subjective assessments in order to maintain focus on the given quality parameters and avoid irrelevant artefacts. A real-world scenario would include inter-frame coding (like IBB* used in second bit-rate) giving a lower rate (we did not observe any visual difference between the I* and IBB*-videos)

with compression ratio at QP32 or above, frame-rate at 10 fps or below, and resolution at 240x160 pixels or below. These findings imply that video quality degradation at these levels are generally perceived as unacceptable for mobile devices with 480x320 pixel screens.

When it comes to frame-rate, previous research has suggested that 6 fps is sufficient for acceptable video quality (McCarthy et al., 2004), yet our data set does not provide support for this threshold. We found mean acceptance scores below neutral even at 15 fps. This decrease in acceptance scores could be related to the larger screens of today's mobile devices, and possibly to an increase in the use and familiarity of watching mobile video. Judging from the implemented levels of compression and spatial resolution, we surmise that the acceptance thresholds for SNR and spatial scaling techniques are located around QP32 and 240x160 pixels. These thresholds for each scaling technique serve as guidelines to the extent a physical quality trait can be reduced without risking an adverse impact on user satisfaction.

It shows in figure 6.12 that the only four levels with mean acceptance scores better than neutral are all different levels of SNR downscaling, ranging from QP12 to QP28. Going by these results, we can conclude that SNR scalability is the most efficient adaptation option. When switching SNR layer, quality differences should be limited to less than 4 QP levels to avoid making flicker artefact visible. However, if a larger quality shift is necessary, a quality level should be kept stable for at least 2 seconds in order to relieve the annoyance of the flicker effect. Combination of different scaling techniques above their respective acceptance threshold is recommended.

The results of our study can help to improve video adaptation strategies in streaming systems or bit-rate controllers for processing scalable video and non-scalable video. The knowledge is applicable for both SVC-type and AVC-type systems. We have used SVC, but the results should be equally important/relevant for AVC-type systems like those used in modern HTTP streaming systems. For SVC, this knowledge helps to schedule the different enhancement layers and decide which to drop in case of congestion. For AVC, it helps determining how to code video segments of different quality in order to optimise the overall viewing experience if congestion forces the application to choose another quality segment.

6.5 Conclusion

To understand the human perception of video quality adaptation in fluctuating bandwidth scenarios, like streaming to handheld devices over wireless networks, we have performed a series of subjective assessment experiments using iPhones and iPods. We evaluated the effect of noise, blur and motion flicker on several different types of video content. For each video type, we investigated how users experience quality changes at different amplitudes and frequencies. In total, we performed 5088 individual assessments.

From our results, we observe that the perception of quality variation is jointly influenced by multiple factors. Amplitude and frequency have significant impact on subjective impression. Most notably, when decreasing the quality switching frequency for flicker in the spatial domain, including noise and blur flicker, users' acceptance scores of the video quality tend to be higher. However, although low frequency can relieve the annoyance of flicker effect in the spatial dimension, decreasing frequency further below a threshold

(on the scale of a few seconds) does not have significant effect. On the other hand, the amplitude has a dominant effect across spatial and temporal dimensions and should be kept as low as possible for satisfactory visual quality. Finally, flicker type and content type are found to influence perceived video quality in their own ways. For instance, both blur and motion flicker demonstrated different effects for different video content types in our experiments. The experimental results indicate that videos with complex spatial details can be easily affected by blur flicker and videos with global motion demand often high frame-rate to be perceived as smooth.

Chapter 7

Conclusion

In this chapter, we will summarize our work on evaluating user experience of scalable video streaming, provide suggestions to video adaptation strategies and outline possible directions of future work.

7.1 Perception and adaptation

In this thesis, we have presented a sequence of user studies that evaluate the effects of multidimensional video adaptation techniques on human quality perception. The results of our subjective tests show that the visual artefacts followed by adaptive operations in different scaling dimension are not perceived in the same manner, so that we can no longer assume a monotonic rate-distortion function for scalable video. With multidimensional scalability, more dynamic bit-rate variations may be introduced in a video stream, which negatively impacts the perceived quality stability. The subjective experience of scalable video becomes even harder to estimate. As objective metrics were generally found unreliable for video quality estimation, we sum up our observations of human quality perception based on our experimental results and provide some initial suggestions for perceptually preferable quality adaptation accordingly.

Our user studies are mainly on the subject of multimedia experience on mobile devices. The conclusions from our subjective quality experiments tell that active video adaptation is a viable way of achieving fine-grained scalability and sufficient perceptual quality, but care should be taken to prevent the detrimental flicker effects. Our experiments reveal that the flicker effect can be very pronounced in the spatial domain if quality changes are more frequent than once per second, but the effect diminishes with the decrease of frequency. To ensure stability of the perceived video quality, we suggest thus not to change the spatial components (pixel values and frame size) for at least 2 seconds when streaming videos online. Not surprisingly, the magnitudes, or the amplitudes of the quality changes, are also bound to affect the perceived quality. In the spatial domain, the flicker effect resulting from great amplitude may significantly impair the perceived quality, while for small amplitudes, the effect could go unnoticed even on high-frequency layer switching. In our subjective studies, we found that the fluctuations of SNR are generally not noticeable for most of people if the QP differences in consecutive video frames are limited to 4 levels. On the other hand, dyadic downscaling of frame-size or QP increment of more than 8 levels bring always noticeable quality degradation. We suggest to avoid such a

large quality downscaling as much as possible. When it comes to quality changes in the temporal domain, the rules of spatial components do not apply. We found that the play-out smoothness is mainly influenced by the frame-rate of the lower temporal layer while cyclic or irregular variations of frame-rate do not have obvious impacts on the multimedia experience on mobile devices.

Besides the flicker thresholds, the acceptance thresholds for quality downscaling in separate dimensions have been explored by our studies. The experimental results imply that compression ratio given by QP 32 or above, frame-rate at 10 fps or below, and resolution at 240x160 or below will generally result in unacceptable quality of video-watching experience on mobile devices with 480x320 screen size. These findings serve as guidelines to the extent a physical quality trait can be reduced without risking an adverse impact on user satisfaction.

Among the three dimensional adaptation techniques, increasing compression ratio by coarse quantization makes the most efficient tradeoff between bandwidth saving and perceived video quality, so that it should be used as the primary adaptation method. However, with very limited bandwidth, these compression ratios below their acceptance threshold may not yield sufficiently low bit-rates, in which case it would be advisable to reduce the frame-rate. Resolution adaptation appears to be the last resort, only to be applied under severely poor conditions. For the best subjective experience, combination of all the three adaptation techniques is recommended in order to avoid exceeding their respective acceptance thresholds. To avoid flicker effects, we also recommend stepwise quality layer switching so that the amplitude and frequency of changes at each step do not exceed their flicker thresholds either.

Quality adaptations do not operate uniformly across video contents according to our experimental results. We found that both the spatial and temporal characteristics of different video contents interacting with the applied adaptation technique. In the spatial domain, the quality acceptance for video contents with complex textural details was more negatively affected by resolution adaptations compared to contents with low spatial complexity, but compression artefacts are more visible in video with smooth or simple texture than in video with complex texture. On the other hand, we observed that high edge contrast counteracts to some extent the influence of flicker effect. As to frame-rate adaptation, video with fast or unidirectional motion received lower evaluation scores than content with slow or non-orientable motion. In addition, people usually do not expect artificial movements to be as smooth as true-life movements. These differences make small but noticeable discrepancies in the actual acceptance and flicker thresholds for concrete video materials. Hence, it would be prudent for service providers to consider video content characteristics before applying an adaptation technique.

Besides the investigations of user perception on mobile devices, we also conducted a preliminary user study on large screens. The study shows that human perceptions change in relation to the viewing environment. Especially, a higher degree of humans sensitivity to motion jerkiness was found on the large screen. The acceptance threshold of frame-rate seems to be significantly higher on large screens than on small screens. The effect of content characteristics seem to be also stronger, which makes larger discrepancies in the acceptance and flicker thresholds for quality fluctuations. But similar to the effect found on mobile devices, artificial movements and high contrast edges make frequent quality changes easier to accept.

7.2 Contributions

The work presented in this dissertation addressed several issues in the field of audiovisual quality assessment, which include visual artefact analysis, experimental design, and subjective evaluation methodology etc. Here, we summarise the main contributions derived from this work.

Development of method for audiovisual quality assessment: A field study is the fundamental means for the exploration of a realistic multimedia experience. However, the practicality of subjective studies is often threatened by prohibitive requirements, in particular by the participant's time and the budget for recompensation. We introduced Randomized Paired Comparison (R/PC), an easy-to-use, flexible, economic and robust tool for conducting field studies. With the use of R/PC, an experimenter can easily obtain stable results with an accuracy close to traditional experiment designs at a much lower cost. We demonstrate the efficiency and practicality of R/PC by simulations. For the first time, we quantify, in a heuristic study, the performance difference between R/PC and classical evaluation method. We prototyped also a software program on iOS to automate the experimental design.

Gathering of subjective evaluations of perceived video quality: We spent a considerable amount of time conducting experiments of subjective evaluation. A large number of reliable subjective evaluation scores were recorded and can be used as reference when comparing or validating different objective quality metrics. We do not limit ourselves to a single genre of video content, and we therefore collected a rich data set that has wide applicability in video streaming systems.

Subjective evaluation of Scalable Video Coding (SVC): The Scalable Video Coding extension of the H.264-AVC standard provides three different types of scalability for efficient and flexible video adaptation. However, the increased number of scaling options increases also the difficulty of visual quality assessment. We conducted the first study that evaluated the subjective performance of multi-dimensional scalability features in SVC. The study reveals that adaptation decisions for SVC bitstreams should not only be based on bit-rate and layer dependency information alone, as the perceived quality degradation may be non-monotonic to bit-rate reduction and the preferred adaptation paths depend on content and user expectations. The experimental results can help improving the design of objective quality models towards multi-dimensional video scalability and the evaluation scores from this study can be used to validate the performance of existing and future objective models.

Subjective evaluation of frequent bit-rate adaptation: Optimal bandwidth adaptation is usually achieved via frequent switching between different bit-rate versions of video segments. To investigate the visual effects and usefulness of frequent bit-rate adaptation, we performed several subjective quality assessment experiments in different scenarios. Our results show that frequent quality variations may create additional visual artefacts denoted flicker effects, and it is not worthwhile making quality changes unless the negative impact of flicker on visual quality is eliminated. We associated the clear definition of flicker effect with different types of quality variations. In addition, we found that people can detect slow or irregular frame-rates

much easier on large HDTV screens than small screens of mobile devices. Therefore, our suggestions of how to make video adaptation strategies were given with the consideration of screen size of the end devices.

In-depth study on flicker effect: The perception of flicker effects is jointly influenced by multiple factors. To get a better understanding of human quality perception of the flicker effects, we performed a comprehensive set of subjective tests on handheld devices. From the study, we were able to identify the main influential factors on the visibility of flicker effects and determine the threshold quantities of these factors for acceptable visual quality of video. These findings can help improving video adaptation strategy or bit-rate controllers deployed in video streaming services. Since our observations were made about the visual artefacts in general terms, the experimental findings are applicable for both scalable or non-scalable video. This is especially useful for modern HTTP streaming systems, which use segmentation to achieve dynamic bandwidth adaptation for non-scalable video. Finally, the flicker effects were explored across different types of video content. We provided some preliminary analyses of content effects on human quality perception.

7.3 Limitations

Audiovisual quality assessment is a relatively new research field for computer scientists. The work presented in this dissertation is still at the exploration stage. Our subjective quality studies provided only some initial suggestions for perceptually preferable quality adaptation. More quality adaptation levels and quality adaptations in more than one dimension should be included for a more thorough survey. Furthermore, we focused mainly on perceived video quality on handheld devices. Only a small-scale experiment that tested fewer types of video content was performed on a HDTV monitor. Therefore, many experimental results apply only to small screens. For a better comparison between the viewing experiences on small and large screen, experimental repetition on large screens is necessary.

In addition, our experiments were performed on mobile devices produced before year 2008. With the rapid development of hardware technology, displays with higher resolution are nowadays used on even small mobile devices. The average screen size of mobile phones increases as well. Our experimental findings may not apply directly on these state-of-the-art devices.

In mobile-viewing environments, both the user and the environment influence the user experience in a much more complex way than in home-viewing environments with large screen. There are more variations in the video streamed over wireless networks, while user's intention of watching mobile video can vary largely from video conversation to aimless kill-time. All of these uncontrolled factors make it impossible to find a close form to express their combined effects on human quality perception. Therefore, we have not developed an objective metric for perceived video quality on mobile devices in this thesis.

7.4 Future work

Our work covers the two research topics: audiovisual quality assessment and adaptive video streaming. We proposed a new method for conducting quality assessment studies in the field. By quality assessment studies of scalable video, we investigated factors that were most relevant for adaptive streaming services. It lies beyond the scope of this thesis to develop an objective quality metric, but we obtained knowledge of the perceived effects of these factors and their interactions with the quality of users' viewing experience. We provided advice on how to deliver satisfactory service quality while avoiding unnecessary bandwidth expense. However, there are still unaddressed issues that need further exploration. In this section, we outline some of the topics that may extend the work presented in this thesis.

In chapter 3, we examined the accuracy of the randomized pairwise comparison method using simulations, but the simulation results are based on the data set from only one study. Data sets from one or more additional new quality assessment studies could be used for further verifying the robustness of our method. Moreover, an analytical model could also be future work to offer a more solid theoretical foundation. In addition, there still exists a demand for new methodologies that are applicable to the evaluation of long viewing experiences in the field.

In all the reported studies (chapter 4, 5, 6), our analyses of video content effects were based on simple classification of visual differences in texture and motion features. However, according to our experience, it could be useful to find alternative ways to classify different video genres. For example, more accurate and complete descriptions of video content can be obtained if we extract and include other content features such as region of interest, focus point, global motion and motion trajectory etc. Then, we could investigate further the influences of these high-level semantic features on user perceived quality. The quality of viewing experience is also influenced by user's expectation and attention. Experiments could also be performed to find out how human's usual practices affects the subjective evaluation of visual quality.

Finally, as described in chapter 5, human perception of visual information on large screens is different than it is on small screens. Additional user studies should be performed to find out the flicker and acceptance thresholds for HD displays. To verify if the suggested streaming strategy also applies to HD video on small screens, we should repeat our user studies on state-of-the-market mobile devices.

Terminology

Flicker effect: A visual artefact that appears due to a repeated fluctuation in video properties such as brightness, sharpness, noise or motion smoothness. Flicker effects can appear as quick flashes, or constant fluctuations at different frequencies, from brief flicker (high frequency), to medium flutter and even slow shifts. In this thesis, we mainly use the term to describe the unstable visual quality caused by adaptation operations in video streaming systems.

Flicker threshold: The flicker threshold defines the frequency at and above which a quality fluctuation appears to be statistically stable for human observers.

Frequent layer switching: A general term used to describe active bandwidth adaptation strategies for streaming scalable video.

Operation point: A subset of a scalable video stream that is identified by a set of particular values of scaling parameters. A bitstream corresponding to an operation point can be decoded to offer a representation of the original video at a certain fidelity in terms of spatial resolution, frame-rate and reconstruction accuracy.

Spatial scalability: The ability to support multiple display resolutions of the same video sequence.

Temporal scalability: The ability to reduce the frame-rate of a video sequence by dropping packets.

SNR/Quality scalability: The ability to have multiple quality versions of the same video sequence.

Scaling/adaptation granularity: The extent to which a scalable video stream can be broken into decodable video streams with smaller sizes.

Visual attribute: The sensory characteristics of video material that are perceived by our senses of sight, i.e blurriness, fogginess etc.

Acceptance threshold: The level or amount of a given quality degradation that fulfills a user's minimal expectations and needs as a part of user experience.

Switching amplitude: The amount of changes in quality level or bit-rate when a scalable video is switched from one video layer to the other.

Switching frequency: The number of times a scalable video repeats switching between its layers within a specified interval.

Independent variable: variable that is deliberately changed by the experimenter to examine its effect on the dependent variable.

Dependent variable: variable that is assumed to be influenced by the changes in the independent variables in an experiment.

Factor: A controlled independent variable in a multi-factorial design.

Treatment: A treatment is a test condition that an experimenter administered to experimental units. It represents by a level of a factor.

Experimental unit: An unit is the basic object upon which the experiment is carried out. For example, a video material is the experimental unit in an audiovisual experiment.

Observation unit: An observation unit is the entity on which information is received and statistical data are collected. In an audiovisual experiment, the observation unit is one individual person.

Test stimulus: A detectable change in the experimental unit after applying a specific treatment. Usually a test stimulus corresponds to a treatment / factor level.

Field study: Study carried out outside the laboratory.

Factorial design: Experiment in which more than one independent variable is involved.

Repeated measure / Within-subjects design: An experimental design in which all subjects experience every treatment condition.

Independent-measures / Between-subjects design: An experimental design in which different groups of subjects are randomly assigned to only one treatment condition.

Blocking: Blocking is a technique that manages relatively homogeneous experimental units into groups (blocks). Comparisons among the factors of primary interest are made within each block.

Replication: Repetition of the basic experiment.

Randomization: Randomization is the design principle used to guard against unknown or uncontrollable factors. By randomization, both the allocation and presentation order of stimuli to observation units in each individual run of the experiment are randomly determined.

Randomized block design: In this type of experimental design, experimental units are divided into several blocks, and the complete randomization is restricted within each block.

Correlation analysis: The study that analyzes the extent to which one variable is related to the another.

Significance level: A fixed probability that is used as criterion for rejecting the null hypothesis.

Main effect: The effect of a factor on a dependent variable averaged across the levels of any other factors.

Interaction effect: The effect of one factor differs significantly depending on the level of another factor.

Type I error: Mistake made in rejecting the null hypothesis when it is true.

Type II error: Mistake made in retaining the null hypothesis when it is false.

Effect size: The size of effect being investigated (difference or correlation) as it exists in the population.

Power: The probability of not making a Type II error if a real effect exists

Confounding: A confounding variable is an extraneous variable that is statistically related to the independent variable, which leads to an error in the interpretation of what may be an accurate measurement.

Bibliography

Adobe (2010). HTTP dynamic streaming on the Adobe Flash platform. http://www.adobe.com/products/httpdynamicstreaming/-pdfs/httpdynamicstreaming_wp_ue.pdf.

Bech, S. and N. Zacharov (2006, July). *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley.

Beeharee, A. K., A. J. West, and R. Hubbold (2003). Visual attention based information culling for distributed virtual environments. In *VRST '03: Proceedings of the ACM symposium on Virtual reality software and technology*, New York, NY, USA, pp. 213–222. ACM.

Callet, P. L., C. Viard-Gaudin, S. Péchard, and É. Caillault (2006). No reference and reduced reference video quality metrics for end to end qos monitoring. *IEICE Transactions 89-B*(2), 289–296.

Chi, C. and F. Lin (1998). A Comparison of Seven Visual Fatigue Assessment Techniques in Three Data-Acquisition VDT Tasks. *Human Factors 40*(4), 577–590.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Comer, D. E., D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, and P. R. Young (1989, January). Computing as a discipline. *Commun. ACM 32*(1), 9–23.

Coolican, H. (2004). *Research Methods and Statistics in Psychology* (4 ed.). Hodder Arnold.

Cranley, N., P. Perry, and L. Murphy (2006). User Perception of adapting Video Quality. *International Journal of Human-Computer Studies 64*(8), 637–647.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 3*(16), 297–334.

Dosselmann, R. and X. Yang (2011). A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing 5*(1), 81–91.

Eichhorn, A. and P. Ni (2009, June). Pick your Layers wisely - A Quality Assessment of H.264 Scalable Video Coding for Mobile Devices. pp. 1019–1025. IEEE.

Eichhorn, A., P. Ni, and R. Eg (2010). Randomised pair comparison: An economic and robust method for audiovisual quality assessment. In *International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*. ACM: ACM.

Evensen, K., T. Kupka, H. Riiser, P. Ni, R. Eg, C. Griwodz, and P. Halvorsen (2014, January). Adaptive media streaming to mobile devices: Challenges, enhancements, and recommendations. *Adv. MultiMedia 2014*, 10:10–10:10.

Frigge, M., D. C. Hoaglin, and B. Iglewicz (1989, February). Some implementations of the boxplot. *The American Statistician 43*(1), 50–54.

Gottron, C., A. König, M. Hollick, S. Bergsträßer, T. Hildebrandt, and R. Steinmetz (2009, December). Quality of experience of voice communication in large-scale mobile ad hoc networks. In *Proceedings of the second IFIP Wireless Days 2009*. IEEE Computer Society.

Goyal, V. K. (2001, September). Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine 18*(5), 74–93.

Gunawan, I. P. and M. Ghanbari (2008). Efficient Reduced-Reference Video Quality Meter. *IEEE Transactions on Broadcasting 54*, 669–679.

Ho, Y.-S. and S. H. Kim (2006). Video coding techniques for ubiquitous multimedia services. In F. Stajano, H. J. Kim, J.-S. Chae, and S.-D. Kim (Eds.), *ICUCT*, Volume 4412 of *Lecture Notes in Computer Science*, pp. 1–10. Springer.

Howell, D. C. (2002). *Statistical Methods for Psychology* (5 ed.). Duxberry.

Huang, J., C. Krasic, J. Walpole, and W. Feng (2003). Adaptive live video streaming by priority drop. In *Proc. of IEEE AVSS*, pp. 342–347.

Huang, Z. and K. Nahrstedt (2012). Perception-based playout scheduling for high-quality real-time interactive multimedia. In A. G. Greenberg and K. Sohraby (Eds.), *INFOCOM*, pp. 2786–2790. IEEE.

Huynh-Thu, Q. and M. Ghanbari (2008, June). Scope of validity of psnr in image/video quality assessment. *Electronics Letters 44*(13), 800–801.

International Telecommunications (2002). *ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television picture*. International Telecommunications.

International Telecommunications Union (1990). *Report 1082-1. Studies toward the unification of picture assessment methodology*. International Telecommunications Union.

International Telecommunications Union (1996). *Principles of a reference impairment system for video*. International Telecommunications Union.

International Telecommunications Union (1999). *ITU-T P.910. Subjective video quality assessment methods for multimedia applications*. International Telecommunications Union.

International Telecommunications Union (2006). *SERIES P. P.800.1. Mean Opinion Score (MOS) terminology*. International Telecommunications Union.

International Telecommunications Union (2007). *Amendment 1. Definition of Quality of Experience (QoE)*. International Telecommunications Union.

International Telecommunications Union (2008a). *Objective perceptual multimedia video quality measurement in the presence of a full reference*. International Telecommunications Union.

International Telecommunications Union (2008b). *SERIES G. G.1080. Quality of experience requirements for IPTV services*. International Telecommunications Union.

International Telecommunications Union (2011). *Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*. International Telecommunications Union.

ITU-T and ISO/IEC JTC 1 (2003, April). *Advanced Video Coding for Generic Audiovisual services, ITU-T Recommendation H.264*. ITU-T and ISO/IEC JTC 1. ISO/IEC 14496-10(AVC).

Jeannin, S. and A. Divakaran (2001, Jun). MPEG-7 visual motion descriptors. *IEEE Trans. on Circuits and Systems for Video Technology 11*(6), 720–724.

J.Sullivan, G. and T. Wiegand (1998, November). Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine 15*(6), 74–90.

Jumisko-Pyykkö, S. and J. Häkkinen (2006). I would like to see the subtitles and the face or at least hear the voice: Effects of screen size and audio-video bitrate ratio on perception of quality in mobile television. In *4th European Interactive TV Conference*.

Kim, C. S., S. H. Jin, D. J. Seo, and Y. M. Ro (2008). Measuring video quality on full scalability of H.264/AVC scalable video coding. *IEICE Trans. on Communications E91-B*(5), 1269–1278.

Kirk, R. E. (1982). *Experimental Design: Procedures for Behavioral Sciences* (2 ed.). Wadsworth Publishing.

Knoche, H. O. and M. A. Sasse (2008). The sweet spot: how people trade off size and definition on mobile devices. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, New York, NY, USA, pp. 21–30. ACM.

Lewis, S. M. and M. Tuck (1985). Paired comparison designs for factorial experiments. *Appl. Statist. 34*(3), 227–234.

Loke, M. H., E. P. Ong, W. Lin, Z. Lu, and S. Yao (2006, Oct.). Comparison of Video Quality Metrics on Multimedia Videos. *IEEE Intl. Conf. on Image Processing*, 457–460.

Martinez-Rach, M., O. Lopez, P. Pinol, M. Malumbres, and J. Oliver (2006, Dec). A study of objective quality assessment metrics for video codec design and evaluation. In *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on*, pp. 517–524.

Mason, R. L., R. F. Gunst, and J. L. Hess (2003, April). *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*. Wiley.

McCarthy, J. D., M. A. Sasse, and D. Miras (2004). Sharp or smooth?: Comparing the effects of quantization vs. frame rate for streamed video. In *Proc. CHI*, pp. 535–542.

Montgomery, D. C. (2006). *Design and Analysis of Experiments*. John Wiley & Sons.

Move Networks (2008, November). Internet television: Challenges and opportunities. Technical report, Move Networks, Inc.

Nadenau, M. J., S. Winkler, D. Alleysson, and M. Kunt (2000). Human vision models for perceptually optimized image processing – a review. In *Proc. of the IEEE*.

Ni, P. (2009, November). Towards optimal quality of experience via scalable video coding. Licentiate thesis, Mälardalen University Press.

Ni, P., R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen (2011a). Flicker effects in adaptive video streaming to handheld devices. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, New York, NY, USA, pp. 463–472. ACM.

Ni, P., R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen (2011b, July). Spatial flicker effect in video scaling. pp. 55–60. IEEE.

Ni, P., A. Eichhorn, C. Griwodz, and P. Halvorsen (2009). Fine-grained scalable streaming from coarse-grained videos. In *Proc. NOSSDAV*, pp. 103–108.

Ni, P., A. Eichhorn, C. Griwodz, and P. Halvorsen (2010). Frequent layer switching for perceived quality improvements of coarse-grained scalable video. *Springer Multimedia Systems Journal 16*(3), 171–182.

Ni, P., F. Gaarder, C. Griwodz, and P. Halvorsen (2009, October). Video streaming into virtual worlds: the effects of virtual screen distance and angle on perceived quality. pp. 885–888. ACM. Short paper.

Ni, P. and D. Isovic (2008, May). Support for digital vcr functionality over network for h.264/avc. pp. 520–525. IEEE.

Ni, P., D. Isovic, and G. Fohler (2006). User friendly h.264/avc for remote browsing. New York, NY, USA, pp. 643–646. ACM Press.

Pantos, R., J. Batson, D. Biderman, B. May, and A. Tseng (2010). HTTP live streaming. http://tools.ietf.org/html/draft-pantos-http-live-streaming-04.

Park, D. K., Y. S. Jeon, and C. S. Won (2000). Efficient use of local edge histogram descriptor. In *Proc. of ACM workshops on Multimedia*, pp. 51–54.

Pereira, F. C. and T. Ebrahimi (2002). *The MPEG-4 Book*. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal 316*(7139), 1236–1238.

Pinson, M. and S.Wolf (2003). Comparing subjective video quality testing methodologies. In *SPIE'03*.

Pinson, M. and S. Wolf (2004, Sept.). A new standardized method for objectively measuring video quality. *IEEE Trans. on Broadcasting 50*(3), 312–322.

Rix, A. W., A. Bourret, and M. P. Hollier (1999). Models of human perception. *BT Technology Journal 7*(1), 24–34.

Sat, B. and B. W. Wah (2009). Statistical scheduling of offline comparative subjective evaluations for real-time multimedia. *IEEE Transactions on Multimedia 11*, 1114 – 1130.

Schwarz, H., D. Marpe, and T. Wiegand (2007, September). Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology 17*(9), 1103–1129.

Seshadrinathan, K., R. Soundararajan, A. C. Bovik, and L. K. Cormack (2010, June). Study of subjective and objective quality assessment of video. *Trans. Img. Proc. 19*(6), 1427–1441.

Shaw, R. G. and T. Mitchel-Olds (1993, Sep). Anova for unbalanced data: An overview. *Ecology 74*(6), 1638–1645.

Sheikh, H., Z. Wang, L. Cormack, and A. Bovik (2002, nov.). Blind quality assessment for jpeg2000 compressed images. In *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*, Volume 2, pp. 1735 –1739 vol.2.

Sheldon, M. R., M. J. Fillyaw, and W. D. Thompson (1996). The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International 1*(4), 221–228.

Sodagar, I. (2011). The mpeg-dash standard for multimedia streaming over the internet. *MultiMedia, IEEE 18*(4), 62–67.

Stockhammer, T. (2011). Dynamic adaptive streaming over http - design principles and standards. In *The Second W3C Web and TV Workshop*.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review 101*(2), 266–70.

Wang, Z., A. Bovik, H. Sheikh, and E. Simoncelli (2004, April). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Processing 13*(4), 600–612.

Wang, Z., H. R. Sheikh, and A. C. Bovik (2002). No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings of IEEE 2002 International Conferencing on Image Processing*, pp. 477–480.

Watson, A. B. and L. Kreslake (2001). Measurement of Visual Impairment Scales for Digital Video. In *Proc. SPIE*, Volume 4299, pp. 79–89.

Wiegand, T., G. Sullivan, G. Bjontegaard, and A. Luthra (2003). Overview of the h.264/avc video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on 13*(7), 560–576.

Wien, M., H. Schwarz, and T. Oelbaum (2007, Sept.). Performance analysis of SVC. *IEEE Trans. on Circuits and Systems for Video Technology 17*(9), 1194–1203.

Wikipedia (2015a). Just-noticeable difference — wikipedia, the free encyclopedia. [Online; accessed 11-December-2015].

Wikipedia (2015b). Quality of experience — wikipedia, the free encyclopedia. [Online; accessed 11-December-2015].

Winkler, S. and P. Mohandas (2008, sept.). The evolution of video quality measurement: From psnr to hybrid metrics. *Broadcasting, IEEE Transactions on 54*(3), 660 –668.

Wu, H., M. Claypool, and R. Kinicki (2006). On combining Temporal Scaling and Quality Scaling for Streaming MPEG. In *Proc. of NOSSDAV*, pp. 1–6.

Wu, W., A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang (2009). Quality of experience in distributed interactive multimedia environments: Toward a theoretical framework. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, New York, NY, USA, pp. 481–490. ACM.

Xu, J. and B. W. Wah (2013). Exploiting just-noticeable difference of delays for improving quality of experience in video conferencing. In *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, New York, NY, USA, pp. 238–248. ACM.

Zambelli, A. (2009). Smooth streaming technical overview. http://learn.iis.net/page.aspx/626/smooth-streaming-technical-overview/.

Zink, M., O. Künzel, J. Schmitt, and R. Steinmetz (2003). Subjective impression of variations in layer encoded videos. In *Proc. IWQoS*, pp. 137–154.

Appendix A

Publications

This appendix lists all the scientific publications of the candidate. Section A.1 contains those publications that are the basis for this thesis. Each publication is briefly described, and the individual contributions of the authors are explained.

A.1 Papers relied on in this thesis

A.1.1 Refereed Proceedings

ACM MM 2011 Ni, Pengpeng, Ragnhild Eg, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. "Flicker Effects in Adaptive Video Streaming to Handheld Devices", In ACM International Multimedia Conference (ACM MM), 2011 (Ni et al., 2011a).

This paper describes the flicker effects, the visual artifacts that are caused by adaptive layer switching, and investigates their perceptual impacts through subjective assessments. Pengpeng Ni was responsible for designing, conducting the quality assessment experiments, as well as statistical analyses of the subjective evaluation scores and writing this paper. Ragnhild Eg contributed by writing, as well as providing feedbacks on the text and the evaluation results. All authors took part in the technical discussions and contributed to the writing. This paper is an extension of the QoMEX 2011 work (Ni et al., 2011b), relied on parts of my work conducted for this thesis. Chapter 6 is based on this paper but extends it with more studies and conclusions. The paper is included in thesis as Appendix E.

NOSSDAV 2010 Eichhorn, Alexander, Pengpeng Ni, and Ragnhild Eg. "Randomised Pair Comparison - an Economic and Robust Method for Audiovisual Quality Assessment", In International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), 2010 (Eichhorn et al., 2010)

This paper introduces an easy-to-use, flexible and economic method for conducting subjective quality perception studies. Alexander Eichhorn contributed to the planning and writing this paper. Pengpeng Ni implemented the required software tools and was in charge of the evaluation experiment. The two were also both responsible for designing the method. Ragnhild Eg contributed by writing the section on fatigue and learning effects. All authors involved in technical discussions and

provided feedback on the paper's content. This paper relied on parts of my work conducted for this thesis. Chapter 3 is based on this paper but extends it with more studies and conclusions. The paper is included in thesis as Appendix B.

ICC 2009 Eichhorn, Alexander, and Pengpeng Ni. "Pick Your Layers Wisely - a Quality Assessment of H.264 Scalable Video Coding for Mobile Devices", In IEEE International Conference on Communications (ICC), 2009 (Eichhorn and Ni, 2009)

This paper relied on parts of my work conducted for this thesis. It presents the first subjective quality assessment we performed to investigate the effects of multi-dimensional scalability on human quality perception. Chapter 4 is based on this paper but extends it with additional analyses. The paper is included in thesis as Appendix C. Alexander Eichhorn designed the experiment and contributed to most of the writing. The experiment was carried out by Pengpeng Ni, who also contributed by performing objective quality evaluation and writing the section on related work.

A.1.2 Journal article

Multimedia Systems Journal Ni, Pengpeng, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. "Frequent Layer Switching for Perceived Quality Improvements of Coarse-Grained Scalable Video", In ACM Multimedia Systems Journal 16, 2010 (Ni et al., 2010)

This paper proposes and evaluates the idea of performing frequent layer switching in coarse-grained scalable video for finer-grained bitrate adaptation. It is an extension of the NOSSDAV 2009 work (Ni et al., 2009), which relied on parts of my work conducted for this thesis. Chapter 5 is based on this paper but extends it with additional analyses and conclusions. The paper is included in thesis as Appendix D. Pengpeng Ni designed and conducted the quality assessment experiments. All authors took part in the technical discussions, contributed to the writing, and provided feedback on the paper's content and structure.

A.2 Other papers co-authored by the candidate

A.2.1 Refereed Proceedings

QoMEX 2011 Ni, Pengpeng, Ragnhild Eg, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. "Spatial Flicker Effect in Video Scaling", In International Workshop on Quality of Multimedia Experience (QoMEX), 2011 (Ni et al., 2011b).

This paper investigates the perceptual impacts of some flicker effects through subjective assessments. Pengpeng Ni was in charge of the experiment and the paper writing. All authors took part in the technical discussions and contributed to the paper's content.

NOSSDAV 2009 Ni, Pengpeng, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen. "Fine-Grained Scalable Streaming From Coarse-Grained Videos", In International

Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), 2009 (Ni et al., 2009)

This paper proposes and evaluates the idea of performing frequent layer switching in coarse-grained scalable video for finer-grained bitrate adaptation. The idea came from Pengpeng Ni, who also designed and conducted the evaluation experiment. Alexander Eichhorn contributed to implementation and statistical analysis. All authors took part in the technical discussions, contributed to the writing, and provided feedback on the paper's content and structure.

ACM MM 2009 Ni, Pengpeng, Fredrik Gaarder, Carsten Griwodz, and Pål Halvorsen. "Video Streaming Into Virtual Worlds: the Effects of Virtual Screen Distance and Angle on Perceived Quality", In the ACM International Multimedia Conference (ACM MM), 2009 (Ni et al., 2009)

This paper presents a subjective study assessing how positioning of video in the 3D virtual environment influences the user perception of quality degradation. Fredrik Gaarder did the programming and carried out the assessment experiment. Pengpeng Ni contributed to statistical analysis and the writing. The other authors contributed with feedback on the textual content.

ICC 2008 Ni, Pengpeng, and Damir Isovic."Support for Digital VCR Functionality Over Network for H.264/AVC", In IEEE International Conference on Communications (ICC), 2008 (Ni and Isovic, 2008)

This paper is an extension of the ACM MM 2006 work (Ni et al., 2006), which proposes a self adaptive rate-control algorithm to optimize the bit allocation in multi-layered video sequences. The major contribution to both writing, implementation and evaluation was made by Pengpeng Ni. Damir Isovic contributed with input and feedback on the textual content.

ACM MM 2006 Ni, Pengpeng, Damir Isovic, and Gerhard Fohler. "User Friendly H.264/AVC for Remote Browsing", In ACM International Conference on Multimedia (ACM MM), 2006 (Ni et al., 2006)

This paper proposes a layered video coding structure and a transcoding scheme for H.264/AVC video to support trick-mode implementation. The idea came from Pengpeng Ni, who also were responsible for most of the writing. Damir Isovic contributed also to the writing, as well as creating figures used for illustration and providing feedback on the paper's content and structure. All authors involved in technical discussions.

A.2.2 Journal article

Advances in Multimedia Evensen, Kristian, Tomas Kupka, Haakon Riiser, Pengpeng Ni, Ragnhild Eg, Carsten Griwodz, and Pål Halvorsen. "Adaptive Media Streaming to Mobile Devices: Challenges, Enhancements, and Recommendations", In Advances in Multimedia, 2014 (Evensen et al., 2014)

This paper evaluates how different components of a streaming system can be optimized when serving content to mobile devices in particular and make recommen-

dations accordingly. All authors contributed with text input and feedback on the paper's content and structure.

A.2.3 Thesis

Licentiate Thesis Ni, Pengpeng. "Towards Optimal Quality of Experience via Scalable Video Coding", Licentiate Thesis, School of Innovation, Design and Engineering, Mälardalen University, 2009 (Ni, 2009)

This thesis collects our early works on the subject of QoE (Ni and Isovic, 2008; Ni et al., 2009; Eichhorn and Ni, 2009; Ni et al., 2009). It touches upon relatively a wider range of aspects of QoE, including trick-mode functionality, virtual screen distance and angle etc.

Appendix B

Paper I

Title: Randomised Pair Comparison - an Economic and Robust Method for Audiovisual Quality Assessment

Authors: Eichhorn, Alexander, Pengpeng Ni, and Ragnhild Eg

Published: In International Workshop on Network and Operating Systems Support for Digital Audio and Video, 2010

Abstract: Subjective quality perception studies with human observers are essential for multimedia system design. Such studies are known to be expensive and difficult to administer. They require time, a detailed knowledge of experimental designs and a level of control which can often only be achieved in a laboratory setting. Hence, only very few researchers consider running subjective studies at all. In this paper we present Randomised Pair Comparison (R/PC), an easy-to-use, flexible, economic and robust extension to conventional pair comparison methods. R/PC uses random sampling to select a unique and small subset of pairs for each assessor, thus separating session duration from the experimental design. With R/PC an experimenter can freely define the duration of sessions and balance between costs and accuracy of an experiment. On a realistic example study we show that R/PC is able to create stable results with an accuracy close to full factorial designs, yet much lower costs. We also provide initial evidence that R/PC can avoid unpleasant fatigue and learning effects which are common in long experiment sessions.

Randomised Pair Comparison - An Economic and Robust Method for Audiovisual Quality Assessment

Alexander Eichhorn[1], Pengpeng Ni[1,2], Ragnhild Eg[1]
[1]Simula Research Laboratory, Norway
[2]Department of Informatics, University of Oslo, Norway
{echa, pengpeng, rage}@simula.no

ABSTRACT

Subjective quality perception studies with human observers are essential for multimedia system design. Such studies are known to be expensive and difficult to administer. They require time, a detailed knowledge of experimental designs and a level of control which can often only be achieved in a laboratory setting. Hence, only very few researchers consider running subjective studies at all.

In this paper we present Randomised Pair Comparison (R/PC), an easy-to-use, flexible, economic and robust extension to conventional pair comparison methods. R/PC uses random sampling to select a unique and small subset of pairs for each assessor, thus separating session duration from the experimental design. With R/PC an experimenter can freely define the duration of sessions and balance between costs and accuracy of an experiment.

On a realistic example study we show that R/PC is able to create stable results with an accuracy close to full factorial designs, yet much lower costs. We also provide initial evidence that R/PC can avoid unpleasant fatigue and learning effects which are common in long experiment sessions.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Evaluation

General Terms

Human Factors, Measurement, Experimentation

Keywords

Quality assessment, Pair comparison, Experiment Design

1. INTRODUCTION

Audiovisual quality assessment fundamentally relies on subjective methods to capture the perceived quality experience of human observers. Subjective assessment in general is useful for measuring end-user acceptance, comparing alternative algorithms and finding optimal designs or configurations. Pair comparison is a particularly prominent assessment method because it involves a simple cognitive task, comparing two stimuli in a pair against each other. Results obtained with pair comparison tests are robust and known to closely reflect perceived sensations on a psychological scale [12].

However, the main drawback of pair comparison is that the number of pairs grows exponentially with the number of factors and factor levels under investigation. Audiovisual quality studies are known to contain a large number of factors and levels. Full factorial experiment designs that cover all possible combinations of influential factors at all levels are impractical. For example, the study of a scalable video encoder may require investigation of effects on multiple scaling dimensions at multiple scaling magnitudes on different content types and different display devices. Another example is a comparison of alternative error protection schemes under different loss patterns and loss rates, potentially generating a variety of decoding artifacts and distortions which may have to be considered separately.

Even fractional factorial designs and blocking strategies [7] which systematically reduce the number of pairs by excluding some factor combinations are of limited help. To stay within time and resource limits, an experimenter has to strictly limit the number of factors, also to avoid undesirable fatigue and learning effects. Screen-based tasks are especially susceptible to fatigue effects, even for durations as short as 15 minutes [2]. In video quality assessment, assessors can easily become tired, bored and uncooperative. Their responses will therefore be increasingly unreliable, leading to greater unexplained variance. Moreover, simple cognitive tasks are quickly mastered [8], and discrimination between two visual signals improves over time [13]. It follows that repeated exposure to the same content during an experiment session (although at different quality levels) may lead to undesired training. Assessors tend to focus on salient features of audio and video clips instead of reporting their overall quality impression. This may lead to stricter than necessary interpretations of salient artifacts.

We introduce *Randomised Pair Comparison* (R/PC) as an economic extension to traditional pair comparison designs which become increasingly counterproductive for audiovisual quality studies. The novelty is that, in contrast to full factorial designs, R/PC randomly selects small subsets of pairs and thus creates a unique experiment session for each assessor. An experimenter can control the session duration regardless of the number of factor-level combinations. This allows to make more realistic assumptions about the time

NOSSDAV'10, June 2–4, 2010, Amsterdam, The Netherlands.

assessors have to spend on a study and makes it easier to use the method on assessors with different background and age. Thus we believe R/PC is useful for laboratory experiments, observed field studies and self-controlled web-based studies. R/PC can offer a level of robustness close to traditional experiment designs while effectively avoiding fatigue and learning effects.

Randomisation in general is known to yield many benefits for statistical analysis, but the random pair selection in R/PC leads to unbalanced and missing data. Without balance, common statistical tools like ANOVA or GLMs become unstable. That's why the data analysis for R/PC has to either sacrifice some of the quality in the obtained data (e.g. ignore within-subject variability) or use computationally more expensive statistics. We will discuss some alternative ways for data analysis and we will show that it is still possible to find significant main effects.

In the remainder of this paper we first discuss related work in section 2 before we present our R/PC method in depth in section 3. Section 4 provides a first validation of R/PC which is based on an example study we conducted in order to compare a full factorial design against R/PC with the same video material, presentation settings and report scales. Finally, section 5 concludes the paper.

2. RELATED WORK

International recommendation such as ITU BT.500-11 [5], ITU-T P.910 [4], provide instructions on how to perform different types of subjective tests for the assessment of video quality in a controlled laboratory environment. The recommended test methods can be classified as Double Stimulus (DS), Single Stimulus (SS) or Pair Comparison (PC) method. The standard recommendations focus on common aspects of subjective assessment such as viewing conditions, measurement scales and basic statistics for data analysis, while stimuli selection and experiment organisation are left to the experimenter.

In DS methods, assessors are asked to rate the video quality in relation to an explicit reference. In contrast, SS and PC methods do not use explicit references. In SS methods, assessors only see and rate the quality of a single video with an arbitrary length. In the PC method, a pair of clips containing the same content in two different impairment versions is presented and the assessor provides a preference for one version in each pair. The rating procedure of PC method is simpler than that of DS and SS methods and the comparative judgement can be easily verified by examining the transitivity of the ratings.

A comparison of the DS and SS method in [10] shows that the SS method can generate quality estimates comparable to DS methods, but humans consider only the last 9 to 15 seconds of video when forming their quality estimate. While DS and SS methods are mainly used to test the overall quality of a video system, PC methods are well-suited for inspecting the agreement between different users [3].

Pair comparison is widely used in various domains. One example are online voting systems [1] for crowd-sourcing quality assessment tasks. The test design exerts loose control of stimuli presentation only. Assessors are allowed to skip between clips in a pair and they can also decide when to vote. This design is limited to test sequences with constant quality, which restricts its capability of evaluating quality fluctuation within sequences. Our R/PC method is also a variant of the PC test design as defined by ITU-T P.910 [4]. We partially follow standard recommendations and restrict the length of a test sequence to 8 to 10 seconds with the consideration of human memory effects. We also let the experimenter freely select content and quality patterns. One difference to standards is that we do not force assessors to vote in a given time. Instead we measure the timing of responses as well.

The experimental design of R/PC is closely related to designs commonly used in psychological, sociological and biological studies. In particular, completely randomised factorial designs and split-plot factorial designs [7] are closest to R/PC . Such designs are economic in a sense that they require the optimal number of factor combinations and responses to find desired main and interaction effects. They mainly assure that data is balanced so that common statistical assumptions are met. The main difference of R/PC is that our design creates unbalanced data due to random pair selection and early drop-outs and that R/PC allows to choose an arbitrarily small number of pairs per session which is independent of factorial combinations.

3. R/PC METHOD DESIGN

We designed the Randomised Pair Comparison Method with realistic expectations about the time assessors are willing to spend in a study and practical assumptions about the ability of experimenters to control environmental and content-related factors. R/PC is robust and easy to use in laboratory and field studies and is even suitable for web-based self-controlled studies.

Session duration is separated from factorial complexity of an experiment and an experimenter can balance between experiment costs and data accuracy. A session can have an arbitrary duration (down to a single pair) and assessors can quit their session anytime, e.g. when they get distracted by phone calls or have to leave a bus or train.

In contrast to traditional full factorial designs R/PC does not collect a full data sample for all pairs from every assessor. The randomisation procedure in R/PC guarantees that all pairs get eventually voted for, that an experiment session will be unique for every assessor and that all required reference pairs are contained in a session.

The overall costs of R/PC are typically lower than that for comparable full factorial designs. R/PC achieves this by shifting the resource consumption (time and number of assessors, resp. number of responses) to software-based randomisation and a computationally more expensive data analysis. Main effects will be visible with a minimum number of assessors, but with too few responses interaction effects may remain undiscovered. More assessors may increase reliability and the chance to find interaction effects. In total, R/PC may require more individual assessors to achieve significant results, but each assessor has to spend less time.

In the remainder of this section we present the general design of our method and recommendations for applying it to a particular problem. We will also discuss some implications on scales and statistical data analysis.

3.1 Presentation

Similar to conventional Pair Comparison methods [4], audiovisual stimuli are presented as pairs of clips. The duration of each clip should not be longer than 10s, but it can be adjusted to the displayed content and purpose of the

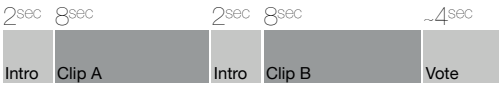


Figure 1: Presentation pattern for a single clip pair.

study. Each clip in a pair is introduced by a 2 second long announcement of the clip name and the letter A or B, displayed as a 50% grey image with black text. This results in the time pattern as shown in figure 1.

After each pair is presented, an assessor is expected to enter a response about the preference on one of the scales defined below. The time to enter the response is recorded and used for later analysis. The session continues with the next pair immediately after the response has been entered.

Because clips in each pair are presented sequentially an assessor's response may be influenced by order which may lead to a systematic bias. We compensate for that by displaying both possible clip orders within a pair (see section 3.3) and randomising the order of pair presentation.

3.2 Factorial Designs

Any two clips in a pair may differ in one or multiple factors, as defined by the experimenter. We call such pairs *contrast pairs*. They are used for actual exploration and hypothesis testing. An experimenter may, for example, base his research hypothesis on assumptions about the visibility and effect size of contrasting factors.

An experimenter should first identify factors which will be controlled in the study and the number of levels for each factor. Factors can be discrete or continuous and the number of levels may differ between factors. Uncontrolled factors such as, for example, an assessor's age and occupation or the time and location of an experiment session should at least be measured if they are regarded as relevant.

Pairs are created for all factor/level combinations. To reduce the overall number of pairs, it is worthwhile to identify factors for blocking. Blocking restricts combinations within each level of the selected factor. The blocking factor can be used in the later analysis to investigate differences between blocks. For example, to explore the effect of video clip content it is not necessary to assess all potential combinations of clips in separate pairs. Instead, only pairs made of the same clip can be assessed.

Blocking by content type also isolates a systematic effect that is introduced by the content itself. Because content is one of the major sources for unexplained variance in audiovisual quality assessment it is desirable to understand and limit its impact. Using a small set of standard sequences would be statistically reasonable, but practically it is undesirable due to the limited relevance findings would have.

3.3 Reference Conditions

Reference conditions are introduced to find unreliable assessors and outliers in the data, but also to understand the perceptual limits of individual assessors.

R/PC uses two types of reference conditions, (1) *equal reference pairs* that contain the same clip at the same quality level twice, and (2) *matched contrast pairs* that just differ in the presentation order of the contained clips, but are regular contrast pairs otherwise. For every contrast pair there should be a matched contrast pair of opposite order. Likewise, for every factor/level combination there should be an equal reference pair, but equal reference pairs don't need a matched counterpart. Reference conditions should be randomly distributed and hidden inbetween other pairs when presented to an assessor to avoid their detection.

Although reference pairs increase the duration of an experiment session, they are necessary to assure data quality and detect systematic problems during data analysis.

3.4 Random Pair Selection and Ordering

In contrast to full factorial designs, where each assessor has to respond to all combinations, the R/PC method creates a *unique random subset of pairs* for each assessor and then randomises the presentation order for pairs.

The procedure ensures that (1) the ratio of contrast to reference pairs is equal in each subset and is also equal to the ratio in a full design, (2) each selected contrast pair is contained in both possible presentation orders (both matched contrast pairs AB and BA are present), (3) equal reference pairs correspond to selected contrast pairs (there is no reference clip version which does not occur in a contrast pair as well), and (4) equal reference pairs are contained only once.

First, an experimenter must pre-determine the subset size. Assuming all pairs are of equal duration, the size s is calculated as $s = d_s/d_p$, where d_s is the total session duration as defined by the experimentor and d_p is the duration of a pair including an estimated time for voting. The subset size should be equal for all assessors in an experiment.

Then contrast pairs are randomly chosen and their matched contrast pair counterparts are added. Assuming there are in total p contrast pairs (in AB order), the same amount of matched contrast pairs (in BA order), and e equal reference pairs, then $s(p/(2p+e))$ (matched) contrast pairs and $s(e/(2p+e))$ equal reference pairs are selected. This ensures the same ratio between contrast pairs and equal reference pairs as in a full factorial design. Note that equal reference pairs have to match the selection of contrast pairs so that no reference pair exists which does not occur otherwise.

Randomisation in general has many benefits for statistical analysis. The randomised pair selection in R/PC , however, leads to an unbalanced data matrix, where (i) the total number of responses per item may be unbalanced, (ii) the number of responses per item can be zero, (iii) each assessor votes for a small percentage of pairs only and thus many empty within-subject cells exist, (iv) the number of responses per assessor may be unbalanced when the assessor quits a session before it ends. Statistical tools for data analysis have to be robust against these uncommon features or a pre-processing step will be required to create the desired features for statistical tests an experimenter would like to employ.

Because R/PC presents a small amount of pairs per session only, it is expected that the number of assessors may be higher than for full factorial designs to achieve stable statistical results. Overall, each assessor spends less time in a session and less responses are collected which may lead to confounding of estimated main effects with factor interactions. Due to the complete randomisation in R/PC the confounding effects are limited. This is because each pair that would be used in a full factorial design will eventually contribute in R/PC as well. For confounding effects to become negligible and results to become stable a sufficient number of assessors is required. A minimal number may dif-

fer between studies and we are investigating the influencing factors further.

3.5 Assessment Task and Reporting Scales

The main task for assessors is to *compare the overall quality* of the presented pairs and to report their preference using a self-report scale. At the beginning of a session a brief introduction about the purpose of the study and the scale which is used may be given. Assessors should be reminded to pay close attention, but an experimenter should avoid specific restrictions or hints which might guide an assessor's focus. A training session is not required.

For voting we suggest not to impose time limitation which would force a decision. Instead we propose to measure the time it takes an assessor to respond and use this in the data analysis. Assessors should report their preference on one of the following comparison scales:

- a *binary preference scale* which allows to express either a preference for clip A or a preference for clip B (forced preference selection results in random votes for equal reference pairs and close to just noticeable differences, JND)
- a *3-point Likert scale* which contains a neutral element in addition to a preference for A or B (promotes indecisiveness, but allows to detect JND thresholds)
- a *4-point Likert scale* which provides items for weak and strong preference for either clip, but lacks a neutral element (has a higher resolution than the binary scale and forces preference selection)
- a *5-point Likert scale* which contains a neutral element as well as items to express weak and strong preference (high resolution, may promote indecisiveness, but allows JND detection)

From a statistical perspective the data obtained with such scales is binomial or ordinal at most. Although some psychometrics researchers argue that data on a 5-point Likert scale can be considered as interval-type because this scale measures psychological units (perceptual differences in our case), we advise to apply non-parametric statistics.

3.6 Data Analysis

Proper data analysis for R/PC is currently a work in progress. In this paper we briefly discuss some implications of our method design and options on how to deal with unbalanced data. For an in-depth discussion on non-parametric procedures see [11].

The nature of the binomial and Likert scales suggests non-parametric statistics for data analysis. Whether the reason for a study is hypothesis testing or exploratory analysis, care should be exercised when responses from assessors are extremely skewed or inconsistent. Even though non-parametric statistics are robust against outlier values, because they rely on medians instead of means, unreliable assessors should be completely removed. Assessor validity can be verified based on reference pairs, in particular, by comparing matched contrast pairs for inconsistent responses.

Useful non-parametric statistical tools are Binomial tests or χ^2 tests for majority analysis on counts (to check whether a majority of preference ratings for one factor-level is significant). As non-parametric counterparts to t-tests and ANOVA, the Mann-Whitney U, Kruskal-Wallis and Friedman tests exist. Rank-order analysis for finding total transitive preference orders between pairs are provided by the zeta method [3] and Thurstone's law of comparative judgement [12]. For more thorough investigations on the underlying structure of the data and to find a linear combination of variables that explains how factors contribute to effects an exploratory factor analysis using generalised linear models or logit models should be considered.

Although we obtain repeated measures, for analysis we regard response data as independent (we drop subject-specific data). Our rationale is that although within-subject variability may be useful to explain effects, we have to sacrifice some of the data quality to compensate for the unbalanced design, in particular the large number of empty cells. Hence, the repeated measures design is just for convenience' sake of obtaining more data from each assessor. We could as well just collect a single response per assessor, which may be more adequate for web-based self-studies.

Ignoring subject-specific data for audiovisual experiments is reasonable because we are interested in general observations which are independent from individual assessors abilities, expectations or perceptual limits. Conclusions from audiovisual quality experiments are expected to apply to a broad spectrum of end-users. Hence we regard all assessors as relatively homogeneous. If a specific target group is of interest in a study, then assessors should be representative for that group.

4. METHOD VALIDATION

To validate the usability and reliability of our R/PC method we performed a simple quality assessment study. Purpose of the study was to obtain two data sets, one with a conventional pair comparison method based on a full factorial experiment design (F/PC) and a second data set with R/PC . We first explain the design of our example study and present an initial analysis of our findings afterwards.

4.1 Example Study

As a simple example of an audiovisual study, we examined the visibility of different video quality reductions in relation to already existing impairments. Our quality impairments originate in a loss of image fidelity between five different operation points, which have been created using different fixed quantisation parameters (QP) for encoding.

In this study, we focus on three factors that are assumed to have main effects, namely the original quality level, the amplitude of a quality change and the content type. These factors can mutually influence the user's perception. For example, the same amplitude in quality changes may be perceived differently depending on the original quality and the direction of the quality change. Interactions may also exist between the change amplitude and the content.

To test different kinds of content with varying detail and motion, we selected six 8 second long clips (200 frames) without scene cut from different genres (see table 1). All clips were downscaled and eventually cropped from their original resolution to 480x320 pixel in order to fit the screen size of our display devices. We used x264 to encode the original clip in constant quantiser mode so that the same amount of signal distortion was added over all frames in a test sequence. Since the visibility of quality impairments is not linearly related to the size of QP, we selected a set of five QPs with

logarithmically distributed values. In a pilot study, the corresponding QPs (10, 25, 34, 38, and 41) have been verified to yield perceptual differences. With five quality levels we can create $\binom{5}{2} = 10$ unique combinations of contrast pairs that have quality change amplitudes between 1 to 4 and five equal reference pairs per content type. In total, we created 120 contrast pairs in both orders and 30 (25%) equal reference pairs. In our example study, a F/PC test session lasted for 60 min while a R/PC test session lasted for only 12 min.

The clip pairs were displayed to assessors on an iPod touch which has a 3.5-inch wide-screen display and 480x320 pixel resolution at 163 pixels per inch. Display and voting were performed on the same device using a custom quality assessment application. The experiment was carried on in a test room at Oslo university. Overall, 49 participants (45% female) at an age between 19 and 39 performed the experiment. Among the participants, 34 people (50% female) were paid assessors who performed both the F/PC test and R/PC test while 15 participants (40% female) are volunteers who performed only the R/PC test. Half of the participants who did both tests, performed the R/PC method first, while the other half did the F/PC test first. During all test sessions the participants were free to choose a comfortable watching position and to adjust the watching distance. They were also free to decide when and for how long they needed a break.

Genre	Content	Detail	Motion
Animation	BigBuckBunny	3.65	1.83
Cartoon	South Park	2.75	0.90
Docu	Earth 2007	3.64	1.61
Movie	Dunkler See	1.85	0.58
News	BBC News	2.92	0.69
Sports	Free Ride	3.32	1.90

Table 1: Sequences used in the experiments. Detail is the average of MPEG-7 edge histogram values over all frames [9] and Motion is the MPEG-7 Motion Activity [6], i.e., the standard deviation of all motion vector magnitudes.

a) Full factorial Pair Comparison

Unique Subjects:	34
Unique Pairs:	150
Unique Responses:	5100
Resp/Subj (min/mean/max):	150 / 150 / 150
Resp/Pair (min/mean/max):	34 / 34 / 34

b) Randomised Pair Comparison

Unique Subjects:	49
Unique Pairs:	150
Unique Responses:	1470
Resp/Subj (min/mean/max):	30 / 30 / 30
Resp/Pair (min/mean/max):	4 / 9.8 / 19

Table 2: Grand totals and statistics for the two data sets in our example study.

4.2 Fatigue and Learning Effects

In order to assess learning and fatigue effects in the 60 minutes long F/PC test, we created a measure of accuracy by coding preference responses as correct, neutral or incorrect. For equal reference pairs, neutral and correct responses were equivalent. Learning and fatigue effects were explored separately, with both reference and contrast pairs.

We expected fatigue effects to become evident already after the first ten minutes, so we divided an experiment session into five equal duration groups, each consisting of 12 min (30 pairs). We also expected the impact of fatigue to be more prominent for video pairs with a fairly visible quality difference, hence video contrasts with one level quality difference were excluded from the analysis. A Pearson chi-square was run for all remaining contrast pairs, but no effects were uncovered ($\chi^2(8)$=11.18, ns). Due to the binary nature of the response categories for the equal reference pairs, a Cochran-Mantel-Haenszel chi-square was used for this analysis. It revealed that response accuracy was conditional of duration (χ^2=4.60(1), p>.05), thus indicating that neutral and incorrect responses varied across one or more duration groups. Binomial tests were applied to further explore this relationship. We found that neutral responses were more frequent in the final compared to the first duration group (S=43, p>.05), otherwise there were no differences in neutral or incorrect responses.

Learning effects were expected to be most relevant where quality differences were hard to spot; hence the analysis included only contrast pairs with one level quality difference. These were grouped according to content repetition, so that five content repetition groups were created based on how many times a video pair with the same content had previously been presented. However, Pearson chi-square revealed no variation according to the number of repetitions ($\chi^2(8)$=5.21, ns). Neither did Cochran-Mantel-Haenszel chi-square reveal any differences for the equal reference pairs (χ^2=3.76(1), ns).

The significant difference in neutral responses for equal reference pairs could indicate that assessors are suffering from fatigue. With more neutral responses towards the end of the experiment, a plausible proposal might be that they become more careless with responses when tired. However, such an effect should perhaps present itself earlier. Another plausible proposal is that the difference is not due to fatigue, but to learning. Although completely randomised, on average an assessor would have observed the presented video contents several times when embarking on the final 30 pairs. Thus the increase in neutral responses may represent an improved ability to detect the absence in difference between equal reference pairs. The current analyses do not provide sufficient data to form a conclusion, but they do suggest that responses change during the course of a F/PC test.

4.3 Reliability

Based on the two data sets we gathered using F/PC and R/PC we did some initial comparative analysis. We are interested whether an investigator using different statistical procedures on either data set would be able to find similar results. Hence, we first looked at the correlation between both data sets and second we tried to fit a linear model to the data in order to find factors which influence main effects.

For the correlation analysis we first calculated the arithmetic mean and the median of all responses per pair. Then we calculated Pearson, Spearman and Kendall correlation coefficients as displayed in table 3. All coefficients were significant below the p<0.01% level.

Metric	CC	SROCC	$\mathcal{T}$
mean	0.974	0.970	0.857
median	0.961	0.965	0.931

Table 3: Correlation between R/PC and F/PC data sets. CC - Pearson Product-Moment Correlation Coefficient, SROCC - Spearman Rank-Order Correlation Coefficient, $\mathcal{T}$ - Kendall's Rank Correlation Coefficient.

Despite the fact that responses in the R/PC data set are very unbalanced (min = 4, max = 19 responses for some pairs, see table 2) and that the total unique responses collected with our R/PC method are only <1/3 of the total F/PC responses, there is still a very strong correlation between both data sets. This supports the assumption that random pair selection may become a useful and robust alternative to full factorial designs for audiovisual quality assessment. However, further analysis is needed to find the minimal number of required assessors and responses.

In our second validation step we compared the results of fitting a generalised linear model (GLM) to both data sets. We used a binomial distribution with a logit link function and modelled the main effects original quality level (Q-max), amplitude of quality change (Q-diff) and content type (content), but no interaction effects. As table 4 shows, all main effects are significant, although the significance is lower in the R/PC case which was to be expected. Again, it is plausible to argue for a sufficient reliability of the R/PC method.

	Factor	Df	Dev	R.Df	R.Dev	P(>χ^2)
f/pc	Q-diff	4	718.43	5095	3505.5	< 2.2e-16
	Q-max	4	54.31	5091	3451.2	4.525e-11
	content	5	34.39	5086	3416.8	1.995e-06
r/pc	Q-diff	4	236.18	1465	1085.2	< 2.2e-16
	Q-max	4	20.48	1461	1064.8	0.0004007
	content	5	16.94	1456	1047.8	0.0046084

Table 4: Deviance analysis for a simple GLM considering main factor effects.

5. CONCLUSION

In multimedia system design the search for optimal solutions is often exploratory, necessitating large numbers of experimental factors which makes full-factorial studies excessively long and draining. In the current paper, we have presented Random Pair Comparison as a practical and economic method for exploratory quality assessment. R/PC provides the possibility of investigating numerous factors, while maintaining the freedom of both experimenters and assessors. We provided first evidence that R/PC is a robust assessment method suitable for finding main effects at reasonable costs.

However, R/PC comes at the expense of higher computational costs for randomisation and data analysis. Violations of normality, uneven response distributions and greater error variance complicate the statistical analysis. In future studies, we aim to further explore non-parametric tests and establish a robust statistical procedure for analysing data generated by R/PC .

One important question remains unanswered so far: what is the minimal number of assessors and responses required to achieve stable results and how much can R/PC really reduce the costs of a study. An answer is not simple since statistical results depend on many factors. In our example study we were able to find significant results with only 29% of responses and costs. A thorough statistical analysis and more data from studies using R/PC will provide further insights.

Acknowledgement

The authors would like to thank Beata Dopierala and Stian Friberg for their help in organising the example study as well as the numerous volunteer participants. This work was sponsored by the Norwegian research council under the Perceval project.

6. REFERENCES

[1] Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. A crowdsourceable QoE Evaluation Framework for Multimedia Content. In *MM '09: Proc. of the ACM Intl. Conference on Multimedia* (New York, NY, USA, 2009), ACM, pp. 491–500.

[2] Chi, C., and Lin, F. A Comparison of Seven Visual Fatigue Assessment Techniques in Three Data-Acquisition VDT Tasks. *Human Factors 40*, 4 (1998), 577–590.

[3] International Telecommunications Union. *Report 1082-1. Studies toward the unification of picture assessment methodology*, 1990.

[4] International Telecommunications Union. *ITU-T P.910. Subjective video quality assessment methods for multimedia applications*, 1999.

[5] International Telecommunications Union - Radiocommunication sector. *ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television picture*, 2002.

[6] Jeannin, S., and Divakaran, A. MPEG-7 Visual Motion Descriptors. *IEEE Trans. on Circuits and Systems for Video Technology 11*, 6 (Jun 2001), 720–724.

[7] Kirk, R. E. *Experimental Design: Procedures for Behavioral Sciences*, 2 ed. Wadsworth Publishing, 1982.

[8] Miller, J., Ruthig, J., Bradley, A., Wise, R., H.A., P., and J.M., E. Learning Effects in the Block Design Task: A Stimulus Parameter-Based Approach. *Psychological Assessment 21*, 4 (2009), 570–577.

[9] Park, D. K., Jeon, Y. S., and Won, C. S. Efficient use of Local Edge Histogram Descriptor. In *Proc. of ACM workshops on Multimedia* (2000), pp. 51–54.

[10] Pinson, M., and S.Wolf. Comparing subjective video quality testing methodologies. In *SPIE'03* (2003).

[11] Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures*, 2 ed. Chapman & Hall, 2000.

[12] Thurstone, L. L. A law of comparative judgment. *Psychological Review 101*, 2 (1994), 266–70.

[13] Yu, C., Klein, S. A., and Levi, D. M. Perceptual learning in contrast discrimination and the (minimal) role of context. *J. Vis. 4*, 3 (3 2004), 169–182.

Appendix C

Paper II

Title: Pick Your Layers Wisely - a Quality Assessment of H.264 Scalable Video Coding for Mobile Devices

Authors: Eichhorn, Alexander, and Pengpeng Ni

Published: In IEEE International Conference on Communications, 2009

Abstract: Multi-dimensional video scalability as defined in H.264-SVC is a promising concept to efficiently adapt encoded streams to individual device capabilities and network conditions. However, we still lack a thorough understanding of how to automate scaling procedure in order to achieve an optimal quality of experience (QoE) for end uses. In this paper we present and discuss the results of a subjective quality assessment we performed on mobile devices to investigate the effects of multi-dimensional scalability on human quality perception. Our study reveals that QoE degrades non-monotonically with bitrate and that scaling order preferences are content-dependent. We confirm previous studies which found common objective metrics to fail for scalable content, but we also show that even scalability-aware models perform poor. Our results are supposed to help improving the design of quality metrics and adaptive network services for scalable streaming applications.

Pick your Layers wisely - A Quality Assessment of H.264 Scalable Video Coding for Mobile Devices

Alexander Eichhorn
Simula Research Laboratory, Norway
Email: echa@simula.no

Pengpeng Ni
Simula Research Laboratory, Norway
IFI, University of Oslo, Norway
Email: pengpeng@simula.no

***Abstract*—Multi-dimensional video scalability as defined in H.264/SVC is a promising concept to efficiently adapt encoded streams to individual device capabilities and network conditions. However, we still lack a thorough understanding of how to automate scaling procedure in order to achieve an optimal quality of experience (QoE) for end uses.**

In this paper we present and discuss the results of a subjective quality assessment we performed on mobile devices to investigate the effects of multi-dimensional scalability on human quality perception. Our study reveals that QoE degrades non-monotonically with bitrate and that scaling order preferences are content-dependent. We confirm previous studies which found common objective metrics to fail for scalable content, but we also show that even scalability-aware models perform poor. Our results are supposed to help improving the design of quality metrics and adaptive network services for scalable streaming applications.

I. Introduction

H.264 Scalable Video Coding (SVC) is the first international video coding standard that defines multi-dimensional scalability [1]. SVC supports several enhancement layers to vary temporal resolution, spatial resolution and quality of a video sequence independently or in combination. This enables efficient adaptation of a compressed bitstream to individual device capabilities and allows to fine-tune the bitrate to meet dynamic network conditions without transcoding. Scaling even works at media aware network elements (MANE) in the delivery path. Hence, SVC is an ideal choice for large-scale video broadcasting like IPTV and content distribution to mobile devices.

SVC was designed for efficient and network-friendly operation [2], but the actual delivery over unreliable networks requires additional methods to protect data and avoid congestion. Such techniques inherently rely on objective video quality metrics (VQM) [3] for optimal performance. QoE, however, is a subjective measure, and current objective models fail to estimate human perception at low frame rates or in mobile environments [4], [5]. An objective metric that considers combined scalability in multiple dimensions and helps content producers or distributors to pick the right combination of layers when encoding, protecting or adapting a scalable video stream is missing so far.

In order to understand human quality perception of H.264/SVC scalability, we performed a subjective field study with a special focus on mobile devices. Our goals are to (1) identify when quality degradations become noticeable, (2) find optimal adaptation paths along multiple scaling dimensions and (3) examine whether objective VQMs can predict subjective observations with reasonable accuracy. To our knowledge, this is the first study that investigates the subjective performance of multi-dimensional scalability features in H.264/SVC.

In this study, we restrict ourselves to on-demand and broadcast delivery of pre-encoded content at bitrates offered by existing wireless networks. Because we are interested in QoE perception on real mobile devices in natural environments, we conduct a field study rather than a synthetic laboratory experiment. Due to lack of space, we focus on static relations between SVC scaling dimensions only. Dynamic aspects like SVC's loss resilience or the impact of layer switching and timing issues on quality perception are not investigated here.

Our results reveal that adaptation decisions for SVC bitstreams should not only be based on bitrate and layer dependency information alone. We found that quality degradation may be non-monotonic to bitrate reduction and that preferred adaptation paths depend on content and user expectations. Confirming previous studies, we also found that common objective VQM like Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index fail for scalable content and even scalability-aware models perform poor.

Our results are supposed to help improving the design of objective quality models towards multi-dimensional video scalability. Enhanced objective models will be useful for several applications and network-level mechanisms, such as bandwidth allocation for wireless broadcasting networks, streaming servers, packet scheduling, unequal error protection and packet classification schemes and quality monitoring.

The paper is organised as follows. Section II briefly summarises related work. Section III presents the design of our field study. Section IV analyses several bitstream properties and Section V reports and discusses our quality assessment results. Finally, Section VI concludes the paper.

II. Related work

The mean squared error based PSNR metric is widely used due to its simplicity, but it does not reflect well the video quality perceived by human observers [3]. To mimic the overall reaction of the human visual system (HVS), Wang et al. proposed the SSIM metric [6] that compares local patches of pixel intensities that have been normalised for luminance and contrast. In [7], the National Telecommunications and

Information Administration General Model (NTIA GM) was introduced for combining measures of the perceptual effects of different types of impairments such as blurring, blocking, jerk, etc. Despite of some reported superiority of the two latter objective models over PSNR, the evaluations performed in [5], [4] indicates that the SSIM and NTIA GM do not work well on multimedia video with low bitrates, various frame rates, and small frames size.

The scaling options of H.264/SVC increase the perceptual uncertainty dramatically. Due to the lack of encoders capable of full scalability, previous studies could not investigate the influence of three-dimensional scaling on quality perception. Additionally, many existing subjective tests like [8]–[10] were conducted on desktop monitors in a controlled laboratory environment. This differs from our testing scenario defined for mobile video applications.

In [8], a set of experiments were carried out to discover the Optimal Adaptation Trajectory (OAT) that maximizes the user perceived quality in the adaptation space defined by frame rate and spatial resolution. Meanwhile, an objective VQM multiplicatively combining the quantization distortion and frame loss was proposed in [11]. The effects of fidelity degradation and frame rate downscaling were also evaluated by subjective tests in [9]. Evaluations like [10] have been performed to investigate the relationship between quality impairment and layer switching at both temporal and quality dimensions. Further, other factors affecting video quality such as performance of codecs, picture ratio and synthetical audiovisual effects etc, were examined in [12]. Although codec performance is critical for decoded video quality, none of the above mentioned evaluations were performed for SVC encoded video, and SVC performance was only measured using PSNR metric in [13]. Recently, Kim et al. proposed a scalability-aware VQM [14] which incorporated the spatial resolution together with frame rate and quality distortion into a single quality metric. We examine this model's performance together with other VQMs in Section V-C.

III. Field Study Design

Our research method is based on ITU-R recommendations for subjective quality assessment BT.500-11 [15], we conducted a field study using iPods as mobile display device and television content that contains an audio track. This research method allows us to study natural user experience under familiar viewing conditions rather than quality perception in a single synthetic environment.

A. Content Selection and Encoding

We selected six sequences from popular genres which are potential candidates for mobile broadcasting (see table I). All sequences were downscaled and eventually cropped from their original resolution to QVGA (320x240). From each sequence, we extracted an 8 second clip (200 frames) without scene cuts. We encoded the SVC bitstreams with version 9.12.2 of the JSVM reference software[1]. The encoder was configured to generate streams in the scalable baseline profile with a GOP-size of 4 frames, one I-picture at the beginning of the sequence, one reference frame, inter-layer prediction and CABAC encoding. Due to the lack of rate-control for enhancement layers in JSVM, we determined optimal quantisation parameters (QP) for each layer with the JSVM Fixed-QP encoder.

Genre	Content	Detail	Motion	Audio
Animation	BigBuckBunny HD	3.65	1.83	sound
Cartoon	South Park HD	2.75	0.90	speech
Documentary	Earth HD	3.64	1.61	sound
Short Movie	Dunkler See	1.85	0.58	sound
News	BBC Newsnight	2.92	0.69	speech
Sports	Free Ride	3.32	1.90	music

Table I
Selected sequences and their properties. Detail is the average of MPEG-7 edge histogram values over all frames [16] and Motion is the MPEG-7 Motion Activity [17], i.e. the standard deviation of all motion vector magnitudes.

Since we are interested in quality perception along and between different scaling dimensions, we defined a full scalability cube with 2 spatial resolutions at QQVGA (160x120) and QVGA (320x240), 3 temporal layers of 25, 12.5 and 6.25 fps, and 4 quality layers with lowest/highest target rate points at 128/256 Kbit for QQVGA/25fps and 1024/1536 Kbit for QVGA/25fps. The target bitrates were chosen according to standard bitrates of radio access bearers in current wireless networking technologies such as HSDPA and DVB-H. For quality scalability, we used SVC's mid-grain scalability (MGS) due to its improved adaptation flexibility that supports discarding enhancement layer data almost at the packet level [1].

B. Scalable Operation Points

From the scalable bitstreams, we extracted six scalable operation points (OP) which cover almost the total bitrate operation range (see table II). Our selection lets us separately assess (a) the QoE drop for temporal scaling at the highest spatial layer (OP1, OP3, OP4), (b) the QoE drop of spatial scalability at two extreme quality points with highest frame rate (OP1 vs. OP5 and OP2 vs. OP6), and (c) the QoE drop of quality scalability at two resolutions with highest frame rate (OP1 vs. OP2 and OP5 vs. OP6).

C. Subjective Assessment Procedures

We performed subjective tests with the Double Stimulus Continuous Quality Scale (DSCQS) method as defined by the ITU [15]. Although this method was designed for television-grade systems, it is widely used as the standard method for several kinds of video quality assessment. DSCQS is a hidden reference method where the original and a distorted sequence (one of the operation points) are displayed twice in A-B-A-B order without disclosing the randomised position of the original. The assessors are asked to score the quality of both

[1] Available at http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm.

Operation Point	Spatial Resolution	Frame Rate	Quality	Layer ID	Target Bitrate
OP1	320x240	25.00	highest	23	1536 kbit
OP2	320x240	25.00	lowest	14	1024 kbit
OP3	320x240	12.50	highest	20	–
OP4	320x240	6.25	highest	17	–
OP5	160x120	25.00	highest	11	256 kbit
OP6	160x120	25.00	lowest	2	128 kbit

Table II
SELECTED OPERATION POINTS.

sequences on a continuous five-grade scale. We interspaced the A-B clips with 4 second breaks, displaying a mid-grey image with black text that announced the following clip or called for voting. We randomised the order of operation points as well as the order of sequences to avoid ordering effects.

Currently, there is no mobile device capable of decoding and displaying SVC bitstreams. Hence, we re-encoded the test sequences into H.264/AVC[2] and displayed them in fullscreen on an iPod classic (80GB model, generation 5.5) as a typical mobile video player. The average distortion introduced by re-encoding was 0.09 dB. Our iPod models contain a 2.5-inch display with 163 ppi and a QVGA resolution. The iPod further supports a low-complexity version of the H.264/AVC Baseline Profile at 1.5 Mbps bitrate. Low spatial resolutions were upscaled to QVGA using JSVM normative upsampling and low frame rates were upscaled by frame copy to the original 25 fps. The audio track was encoded into AAC-LC 48 KHz 120 KBit after the volume was normalised.

Thirty non-expert assessors (33% female) in age classes between 18 and 59 with different education participated in the test. At the beginning, an introduction was held and a training sequence covering the upper and lower quality anchors was shown. The test session lasted for half an hour. We calculated the differential mean opinion scores (DMOS) per operation point after quantising the raw scores obtained from each assessor. We then screened the scores for outliers and inconsistencies as defined in [15] and checked the reliability with Cronbach's alpha coefficient [18]. As normality assumptions for DMOS scores were violated, we used conservative non-parametric statistics for further processing. We also specify Cohen's statistical effect size and power [19] to provide further confidence in our observations. Effect size helps to diagnose validity and discern consistent from unreliable results, e.g. a small effect size reflects a weak effect caused by small difference between scores. Power is the probability of not making a type-II error, that is, with low power we might find a real existing effect as not significant.

D. Limitations

Field studies generally suffer from less controlled presentation conditions. We therefore designed our study carefully by selecting more participants than required by ITU-R BT.500-11 and strictly removed outliers (6 in total among 30). To alleviate effects of an audio track which can influence video quality perception [12], we used undistorted, perfectly synchronised and normalised signals for all sequences. Although we are likely to miss effects that might have been observed in a laboratory, we still found significant results at significance level $p < 0.01$ of high statistical power and effect size in all tests. According to the power the number of participants was also sufficient for obtaining all results presented here.

DSCQS is sensitive to small differences in quality and used as quasi-standard in many subjective studies. For scalable content, however, it has two drawbacks. First, DSCQS is impractical to assess large numbers of operation points at several scaling dimension due to the limited amount of time before assessors become exhausted. Hence, we selected representative operation points only. Second, the scale used by DSCQS is ambiguous because QoE perception is not necessarily linear for people and individual participants may interpret scores differently [20]. Hence, assuming DMOS scores obtained by DSCQS are interval-scaled is statistically incorrect. We address this by lowering our assumptions to ordinal data and non-parametric statistics. Despite these facts, we still found significant results and regard unnoticed effects as insignificant for mobile system design.

IV. BITSTREAM ANALYSIS

Compared to non-scalable video streams, a scalable video stream is more complex. In this section, we analyse a scalable bitstream to detect some of its structural properties.

A. Scaling Granularity and Diversity

Figure 1 displays the bitrate distribution in the Sports bitstream at different operation points. Each OP extracted from a SVC bitstream is identified by an unique combination of its spatial, temporal and quality layers tagged as $[S_m, T_n, Q_i]$. To further describe a scalable bitstream, we introduce two properties: *scaling granularity* and *scaling diversity*. Granularity is the difference between bitrates of two close-by scaling options.

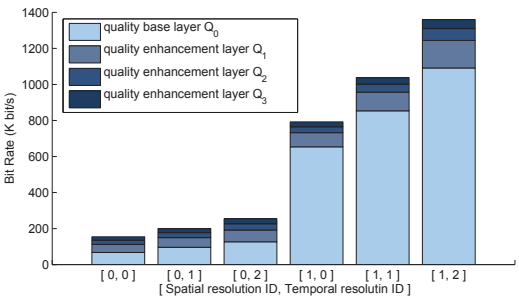


Figure 1. Bitrate allocation for scalable OPs $[S_m, T_n]$ in Sports sequence, where S_m represents m-th spatial resolution, T_n represents n-th temporal resolution. Each bar column can be additionally truncated into 4 quality layers identified by Q_i.

[2]AVC re-encoding was done with x264 version 2245 available at http://www.videolan.org/developers/x264.html.

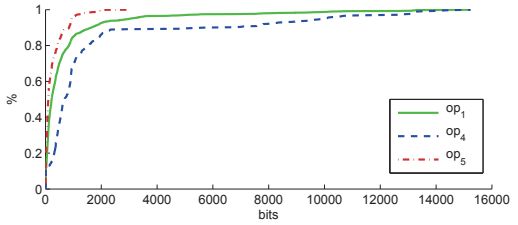


Figure 2. Cumulative distribution function (CDF) of NALU packet sizes for selected operation points of the Sports sequence.

Sequence	Dim from to	T 25 fps 12 fps	T 12 fps 6 fps	T 25 fps 6 fps	S 320H 160H	S 320L 160L	Q 320H 320L	Q 160H 160L
Animation		+++	+++	+++	+++	+++	+++	+
Cartoon		○	○	○	+++	+++	++	○
Documentary		++	+++	+++	+++	+++	○	○
Short Movie		+++	+++	+++	+++	+++	+++	○
News		+++	+++	+++	+++	+++	○	○
Sports		+++	+++	+++	+++	+++	+++	○
All		++	+++	+++	+++	+++	++	○

Table III
NOTICEABLE EFFECT OF QOE DROP WITHIN DIMENSIONS.
LEGEND: ○ NOT SIGNIFICANT, + SMALL EFFECT, ++ MEDIUM EFFECT, +++ LARGE EFFECT.

Smaller bitrate differences give higher granularity. Obviously, video streams with higher granularity can be more robust and adaptive to bandwidth variations. Scaling diversity, on the other hand, reflects the number of distinct scaling options for efficient utilisation of a given bandwidth. Higher diversity provides more adaptation paths to choose.

Scaling granularity and scaling diversity in figure 1 are higher in the range of lower bitrates and OPs with low spatial resolution. I.e., at the bitrate of approximately 192 Kbps, the scaling diversity becomes as high as 3 where $[S_0, T_0, Q_3]$, $[S_0, T_1, Q_2]$ and $[S_0, T_2, Q_1]$ overlap. On the other hand, in the range of high bitrates the granularity is coarser and diversity is reduced. I.e., at a bitrate of 600 Kbps no alternative scaling option exists besides dropping to $[S_0, T_2, Q_3]$ which wastes a considerable amount of bandwidth.

B. Packet Statistics

To further understand bitstream properties, we investigate size and distribution of Network Abstract Layer Units (NALU). This is of interest for protocol designer who need to fragment or aggregate NALUs into network packets.

In figure 2, OP1 is actually the global SVC bitstream which comprises all NALUs. OP4 has the same spatial and quality resolution as OP1, but the lowest temporal resolution. It contains a subset of the NALUs in OP1 only and according to figure 2 the maximum packet size in both OP1 and OP4 is 15235 bits. However, it appears that OP4 contains a larger percentage of NALUs compared to OP1. For example, about 6% of the NALUs in OP1 are larger than 2000 bits, while OP4 contains 14% of such NALUs. This reflects the fact that anchor/key frames in lower temporal layers require more bits than frames in higher layers. Meanwhile, OP5 at the lower spatial layer has a maximum packet size of 2935 bits. This reveals that low spatial layers usually contain small packets only, while the larger packets are contained in higher spatial layers.

V. SVC QUALITY ASSESSMENT

This section reports on our results of three statistical analysis we performed to gain initial insights into human perception of multi-dimensional scalability of SVC encoded video.

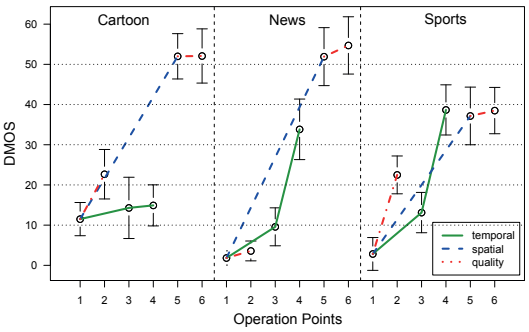


Figure 3. Subjective DMOS scores for selected sequences as means with 95% confidence intervals. QoE gradients for within-dimension scaling are shown as lines. Note that higher DMOS scores mean lower QoE and that the bitrate drops from 1.5 Mbit for OP1 to 128 Kbit for OP6.

A. Noticeable QoE Degradations

The objective of this analysis is to find out whether a degradation in a particular scaling dimension is visible, and if this depends on content or on another dimension that was previously downscaled. We assume at least for some sequences that if QoE is already poor, an additional reduction in another dimension is perceived as less severe.

For this analysis, we check if DMOS values of two operation points on the same scaling axes differ significantly. We perform directional Wilcoxon tests pair-wise for all operation points by expecting higher means for DMOS of lower-layer operation points, meaning they represent a lower QoE.

Table III shows that a QoE drop was noticed with a large effect size and sufficient power in almost all dimensions for almost all sequences. One exception is the Cartoon sequence, where no significant evidence for a noticeable QoE degradation for temporal scaling was found. Even at a very low frame rate our assessors seemed to regard the QoE as sufficient. The reason is that the content already is non-naturally jerky. We also observed that quality scalability seems to have a less noticeable effect, especially when applied to spatially downscaled content. At low spatial resolution we found no

significant degradation in most sequences and even at high spatial resolution the effects were small.

Figure 3 further clarifies the observed effects on three examples. Displayed are DMOS scores and QoE gradients for single-dimension scaling. We avoid speculations about absolute differences here, because scores are non-linear and ordinal only. However, some conclusions can still be drawn: First, detectability and severity of QoE degradations depend on scaling dimension and content. Second, QoE degradations may be non-monotonic to bitrate reduction.

Cartoon is almost unaffected by frame rate reduction due to its non-natural motion as demonstrated by the overlapping confidence intervals of OP1, OP3 and OP4. Our assessors were also less sensitive to further QoE reductions when the quality was already poor, such as shown for SNR scaling at low spatial resolution (OP5 – OP6). In the Sports sequence, initial spatial or quality scaling is perceived worse than temporal scaling. This is in line with results found in [9]. However, below a certain bitrate limit, further downscaling had no effect on QoE regardless of the scaling dimension.

While the News sequence shows a logistic relation between QoE and bitrate which was also found by [9], Cartoon and Sports display non-monotonic characteristics. At least the first temporal scaling stage got a better QoE score than quality scaling although the operation point has a lower bitrate. Moreover, despite the huge bitrate drop in the Sports sequence from 800 Kbit (OP4) to 128 Kbit (OP6) a further quality reduction was not found significant. Hence, monotony assumptions about the relation between bitrate and QoE should be reconsidered for multi-dimensional scaling.

B. Scaling Order Preferences

This analysis is supposed to identify quality-optimal ordering relations for SVC bitstream scaling. In particular, we want to find out (1) whether there exists a scaling dimension that is generally preferred to be scaled first and (2) whether optimal scaling paths depend on content.

We define a dominates relation $D_i \succeq D_j$, which expresses that scaling in one dimension D_i has a larger impact on QoE perception than scaling in another dimension D_j. Note that this is still possible for ordinal data. In order to determine domination from our data set, we select all operation point pairs (OP_k, OP_l) that differ in exactly two scaling dimensions, whereas OP_k contains more layers in dimension D_i and less in D_j and vice versa for OP_l. If OP_l has a significantly higher DMOS score than OP_k, an increase of layers in dimension D_j can obviously not compensate for a decrease of layers in dimension D_i. We then say that D_i dominates D_j or $D_i \succeq D_j$.

With the dominates relation we identify whether there is a positive effect $D_i \succeq D_j$ or a negative effect $D_j \succeq D_i$ between any two dimensions. Table IV displays the results for the five dimension pairs we covered with our OP selection. Spatial scaling is generally regarded worse compared to temporal and quality scaling, although it yields the largest bitrate variability. An adaptation scheme should therefore drop quality layers and some temporal layers first. The preferences are, however, content dependent as revealed by figure 3. Quality and temporal dimensions yield smaller bitrate variability, especially in OPs with higher spatial resolution. Fine granularity adaptation with a minimal QoE drop is possible here, but scaling options are rare due to a low scaling diversity. In contrast, the high scaling diversity at low spatial resolution is useless because QoE is already too bad to notice a significant difference there. Hence, reasonable relations between scaling paths and bitrate variability should already be considered during encoding.

Sequence	Dim / OP_k / OP_l	$T_{12} \succeq S$ OP3 OP5	$T_6 \succeq S$ OP4 OP5	$T_{12} \succeq Q$ OP1 OP3	$T_6 \succeq Q$ OP2 OP4	$S \succeq Q$ OP5 OP2	Pref. Order
Animation		---	+	○	+++	+++	1
Cartoon		---	---	-	-	+++	2
Documentary		---	-	○	+++	+++	3
Short Movie		---	---	○	+++	+++	3
News		---	---	++	+++	+++	2
Sports		---	○	--	+++	++	4
All		---	--	○	+++	+++	-

Table IV
SCALING ORDER PREFERENCES BETWEEN DIMENSIONS.
LEGEND: T - TEMPORAL, S - SPATIAL, Q - QUALITY (SNR) DIMENSION, --- LARGE NEGATIVE EFFECT, -- MEDIUM NEGATIVE EFFECT, - SMALL NEGATIVE EFFECT, ○ NOT SIGNIFICANT, + SMALL POSITIVE EFFECT, ++ MEDIUM POSITIVE EFFECT, +++ LARGE POSITIVE EFFECT. PREFERRED SCALING ORDERS: 1 ($Q - T_{12} - S - T_6$), 2 ($T_{12} - T_6 - Q - S$), 3 ($Q - T_{12} - T_6 - S$), 4 ($T_{12} - Q - T_6 - S$).

We also determined the preferred scaling order for each sequence which is easy because the dominates relation creates a partial order over dimensions. We found four different preferential orders for the six sequences in our test (see the last column of table IV). This clearly justifies that human perception of multi-dimensional QoE degradation is content-specific. An optimal SVC adaptation scheme should consider content characteristics.

We further observed that QoE perception is influenced by assessor expectations, rather than by technical content characteristics alone. Comparing the preferences of temporal and quality scaling for News and Sports in figure 3 it becomes clear that even for the low motion News sequence a lower frame rate was more annoying than a lower quality. The opposite happened to high-motion Sports sequence. Our assessors obviously expected less detail for News and more detail for Sports. Common metrics for textural detail and motion activity like the ones used in table I cannot model such situations well. We found no significant correlation to subjective preferences.

C. Objective Model Performance

In this section, we analyse the performance of some existing objective video quality assessment models. Among many existing models, we selected three popular ones: Y-PSNR, SSIM [6] and the NTIA General Model [7]. In addition, we implemented a recently proposed model which is specifically designed for video streams with multi-dimensional scalability [14]. For simplicity, we call this model SVQM.

Metric	CC	SROCC
Y-PSNR (copy)	-0.532	-0.562
Y-PSNR (skip)	-0.534	-0.555
SSIM (copy)	-0.271	-0.390
SSIM (skip)	-0.443	-0.451
NTIA GM	0.288	0.365
SVQM	-0.661	-0.684

Table V
CORRELATION RESULTS FOR OBJECTIVE QUALITY MODELS.
CC - PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT,
SROCC - SPEARMAN RANK-ORDER CORRELATION COEFFICIENT.

For each test sequence we compared the quality of all the extracted and decoded OPs with the original video sequence using the four objective models. We omitted temporal and spatial registration because all decoded OPs are perfectly aligned with the reference video. For those OPs with lower frame rate, the missing video frames were either skipped or the available frames were duplicated to replace the dropped frames. We performed skipping only for PSNR and SSIM to understand the influence of frame repetition and temporal scalability on those models. Finally, the video quality of each OP was quantified into a single value by averaging the quality values of each single or pair of frames. We measured the objective model performance using Pearson's and Spearman's correlation coefficients. Correlation was found to be significant with $p < 0.01$ at high power.

As table V reveals, SSIM and NTIA GM perform bad for scalable content on mobile screens. Although other studies reported good performance at television resolutions, both models are not tailored to multi-dimensional scalability and small screen sizes. PSNR performs only slightly better. SVQM achieved the best results of all examined models, but it is still far from being ideal. Although our version of SVQM is trained for the sequences used in [14] it still creates reasonable results for our content. This indicates that the general idea of considering motion, frame rate and spatial resolution in an objective model can yield some benefits. In contrast, a simple extension to traditional metrics like PSNR or SSIM which skips missing frames at low temporal resolutions does not create considerably better results.

VI. CONCLUSIONS

We performed a subjective field study to investigate the effects of multi-dimensional scalability supported by H.264/SVC on human quality perception. Our results reveal that visual effects of QoE degradations differ between scaling dimensions and scaling preferences are content dependent. None of the existing objective models works well on multi-dimensional scalable video, but the objective model with scalability-awareness performed slightly better than the others.

For optimal QoE and increased chances of adaptation tools to follow preferred scaling orders, video encoders should maximise the scaling diversity and granularity of bitstreams. MGS is generally recommended for increased scaling granularity and advanced signalling mechanisms are required to inform adaptation tools about content genre, recommended scaling paths, diversity and granularity of bitstreams.

ACKNOWLEDGEMENT

The authors would like to thank Pål Halvorsen, Carsten Griwodz and Dominik Strohmeier for the fruitful discussions as well as the numerous participants who volunteered in our field study.

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Extension of the H.264/AVC Video Coding Standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[2] S. Wenger, W. Ye-Kui, and T. Schierl, "Transport and Signaling of SVC in IP Networks," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1164–1173, 2007.

[3] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Trans. on Broadcasting*, vol. 54, no. 3, pp. 660–668, Sept 2008.

[4] M. H. Loke, E. P. Ong, W. Lin, Z. Lu, and S. Yao, "Comparison of Video Quality Metrics on Multimedia Videos," *IEEE Intl. Conf. on Image Processing*, pp. 457–460, Oct. 2006.

[5] M. M. et al, "A Study of Objective Quality Assessment Metrics for Video Codec Design and Evaluation," in *Proc. of the IEEE Intl. Symposium on Multimedia*, Washington, DC, USA, 2006, pp. 517–524.

[6] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[7] M. Pinson and S. Wolf, "A New Standardized Method for objectively Measuring Video Quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[8] N. Cranley, P. Perry, and L. Murphy, "User Perception of adapting Video Quality," *International Journal of Human-Computer Studies*, vol. 64, no. 8, pp. 637–647, 2006.

[9] J. D. McCarthy, M. A. Sasse, and D. Miras, "Sharp or Smooth?: Comparing the Effects of Quantization vs. Frame Rate for Streamed Video," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2004, pp. 535–542.

[10] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz, "Subjective Impression of Variations in Layer Encoded Videos," in *Proc. of Intl. Workshop on Quality of Service*, 2003, pp. 137–154.

[11] H. Wu, M. Claypool, and R. Kinicki, "On combining Temporal Scaling and Quality Scaling for Streaming MPEG," in *Proc. of NOSSDAV*, 2006, pp. 1–6.

[12] S. Jumisko-Pyykkö and J. Häkkinen, "I would like to see the subtitles and the face or at least hear the voice: Effects of Screen Size and Audio-video Bitrate Ratio on Perception of Quality in Mobile Television," in *4th European Interactive TV Conference*, 2006.

[13] M. Wien, H. Schwarz, and T. Oelbaum, "Performance Analysis of SVC," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1194–1203, Sept. 2007.

[14] C. S. Kim, S. H. Jin, D. J. Seo, and Y. M. Ro, "Measuring Video Quality on Full Scalability of H.264/AVC Scalable Video Coding," *IEICE Trans. on Communications*, vol. E91-B, no. 5, pp. 1269–1278, 2008.

[15] *ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television picture*, International Telecommunications Union - Radiocommunication sector, 2002.

[16] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of Local Edge Histogram Descriptor," in *Proc. of ACM workshops on Multimedia*, 2000, pp. 51–54.

[17] S. Jeannin and A. Divakaran, "MPEG-7 Visual Motion Descriptors," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 720–724, Jun 2001.

[18] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 3, no. 16, pp. 297–334, 1951.

[19] J. Cohen, *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, 1988.

[20] A. B. Watson and L. Kreslake, "Measurement of Visual Impairment Scales for Digital Video," in *Proc. SPIE*, vol. 4299, 2001, pp. 79–89.

Appendix D

Paper III

Title: Frequent Layer Switching for Perceived Quality Improvements of Coarse-Grained Scalable Video

Authors: Ni, Pengpeng, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen

Published: In ACM Multimedia Systems Journal 16, 2010

Abstract: Scalable video is an attractive option for adapting the bandwidth consumption of streaming video to the available bandwidth. Fine-grained scalability can adapt most closely to the available bandwidth, but this comes at the cost of a higher overhead compared to more coarse-grained videos. In the context of VoD streaming, we have therefore explored whether a similar adaptation to the available bandwidth can be achieved by performing layer switching in coarse-grained scalable videos. In this approach, enhancement layers of a video stream are switched on and off to achieve any desired longer-term bandwidth. We have performed three user studies, two using mobile devices and one using an HDTV display, to evaluate the idea. In several cases, the far-from-obvious conclusion is that layer switching is a viable way of achieving bit-rate savings and fine-grained bit-rate adaptation even for rather short times between layer switches, but it does, however, depend on scaling dimensions, content and display device.

Multimedia Systems (2010) 16:171–182
DOI 10.1007/s00530-010-0186-9

REGULAR PAPER

Frequent layer switching for perceived quality improvements of coarse-grained scalable video

Pengpeng Ni · Alexander Eichhorn · Carsten Griwodz · Pål Halvorsen

Published online: 8 May 2010

Abstract Scalable video is an attractive option for adapting the bandwidth consumption of streaming video to the available bandwidth. Fine-grained scalability can adapt most closely to the available bandwidth, but this comes at the cost of a higher overhead compared to more coarse-grained videos. In the context of VoD streaming, we have therefore explored whether a similar adaptation to the available bandwidth can be achieved by performing layer switching in coarse-grained scalable videos. In this approach, enhancement layers of a video stream are switched on and off to achieve any desired longer term bandwidth. We have performed three user studies, two using mobile devices and one using an HDTV display, to evaluate the idea. In several cases, the far-from-obvious conclusion is that layer switching is a viable way of achieving bit-rate savings and fine-grained bit-rate adaptation even for rather short times between layer switches, but it does, however, depend on scaling dimensions, content and display device.

Keywords Quality of experience · Scalable video · Layer switching

1 Introduction

Streaming stored video to a large number of heterogeneous receivers over various networks introduces several challenges with respect to delivered rate and quality. Various layered video approaches that address this exist, including coarse-grained and fine-grained scalable video and multiple description coding. They can be used to choose a quality level whose bandwidth can be delivered to and consumed by a receiver with a limited amount of prefetching and buffering. They can also be used to adapt over time the amount of bandwidth that is delivered to a single receiver. Fine-grained scalable video is apparently meant for the latter approach in particular. However, since both fine-grained scalable video and multiple description coding suffer from a considerable overhead, the question arises whether more or less frequent switching between the layers of a coarse-grained scalable video could yield better bandwidth adaptation while providing similar or even better perceived quality.

In [10], we introduced the technique of frequent layer switching (FLS), a method for fine-grained bit-rate adaptation of scalable bitstreams with few scaling options. Here, we investigate the perceptual effects and usefulness of FLS in mobile and HDTV scenarios. Our aim is to provide recommendations on how to best incorporate FLS into practical streaming systems.

In general, we are interested in two central questions:

- Is FLS a useful alternative to downscaling in streaming scenarios with limited and fluctuating bandwidth?
- How do switching frequency and display size influence the subjective quality perception of human observers?

We used multiple assessment methods in different environments and investigated selected switching and scaling patterns systematically.

We performed our study on material that has been encoded in H.264 scalable video coding (SVC), an international video coding standard with multi-dimensional

P. Ni (✉) · A. Eichhorn · C. Griwodz · P. Halvorsen
Simula Research Laboratory, Lysaker, Norway
e-mail: pengpeng@simula.no

P. Ni · C. Griwodz · P. Halvorsen
Department of Informatics, University of Oslo, Oslo, Norway

scalability [13] supporting different temporal resolutions, spatial resolutions and qualities of a video sequence. SVC uses multiple enhancement layers and is designed for efficient and network-friendly operation [14]. Device heterogeneity and bandwidth variations can be supported by tuning resolution and bit-rate off-line to meet individual device capabilities and using adaptive downscaling of the compressed bitstream during streaming.

The granularity of the scaling options is determined by the bit rates of contained operation points, i.e., between the different encoded quality layers. Scaling options are predetermined at encoding time and the standard currently limits number of supported enhancement layers to 8 [13]. SVC's mid-grain scalability (MGS) feature is supposed to introduce higher adaptation granularity, but this comes again at the cost of increased signaling overhead. For better bit-rate efficiency, it is thus desirable to limit the number of layers and also the number of MGS partitions.

In previous work [3], we showed that, at low bit rates less than 200 kbps, a scalable stream with a fixed set of operation points (6) can have sufficient granularity. However, for higher bit-rate streams, the granularity becomes coarse and the diversity of scaling options is reduced. This results in a lack of alternative scaling options, either wasting resources or decreasing the quality of experience (QoE) more than necessary.

Layer switching can achieve a bandwidth consumption different from the long-term average of any operation point of a coarse-grained scalable video without the extra costs of MGS. This ability makes FLS suitable in several streaming scenarios:

- FLS can be used to achieve a long-term average target bit rate that differs from average bit rates of available operation points in coarse-grained scalable videos. This works even for variable bit-rate SVC streams. Every average target bit rate above the base layer's bandwidth demand can be achieved by switching enhancement layers on and off repeatedly, if necessary at different on and off durations.
- FLS can be used as an alternative means to exploit the temporary availability of bandwidth that exceeds the demands of the base layer, but does not suffice the bandwidth demands of an enhancement layer. Through variations of the retrieval speed (implicitly in pull mode, explicitly in push mode), receivers can use the excess bandwidth during a period of base-layer playout to prefetch data for a period of enhanced-quality playout. The period duration depends on the available space for a prefetching buffer, but it also depends on the perceived playout quality which forbids an arbitrary choice.
- FLS can be used for bandwidth sharing in fixed-rate channels, in particular, for multiplexing multiple scalable bitstreams over Digital Video Broadcasting channels. With FLS, a channel scheduler gains more selection options to satisfy quality and bit-rate constraints. In addition to coarse operation point bit rates, FLS can offer intermediate bit rates at a similar QoE.

In all the above scenarios, the choice of switching pattern and switching frequency is of central importance because they may considerably impact the perceived quality. To identify the feasibility of switching techniques and give advice on design constraints, we conducted a subjective quality assessment study asking human observers for their preferences when watching video clip pairs impaired with different switching and scaling patterns.

We have performed experiments in three different scenarios, i.e., mobile displays in private spaces, mobile displays in public spaces and HTDV displays in private spaces. Our results indicate that the perceived quality of different switching patterns may differ largely, depending on scaling dimensions, content and display device. In some cases, there are clear preferences for one technique while in other cases both, switching and downscaling, are liked or disliked equally. In several cases, FLS is a practical alternative for achieving fine-grained scalable streaming from coarse-grained videos, i.e., if the switching period is long enough to avoid flickering, then layer switching is even preferred over downscaling to a lower SVC quality layer.

The remainder of this paper is organized as follows: Sect. 2 discusses some relevant related work. Our study is further described in Sect. 3, whereas the experimental results are presented in Sects. 4, 5 and 6 for the three scenarios, respectively. We discuss our findings in Sect. 7, and in Sect. 8, we summarize the paper.

2 Related work

SVC increases perceptual uncertainty dramatically because of its multi-dimensional scaling possibility. There are a few published studies investigating the quality influence of different scaling options. In [2], a set of experiments was carried out to discover the Optimal Adaptation Trajectory that maximizes the user perceived quality in the adaptation space defined by frame rate and spatial resolution. It was shown that a two-dimensional adaptation strategy outperformed one-dimensional adaptation. Meanwhile, according to an objective video quality model [15] that multiplicatively combines the quantization distortion and frame loss, it was claimed that quality scaling worked better than temporal scaling under most circumstances. Additionally, the subjective tests presented in [8] showed that high frame

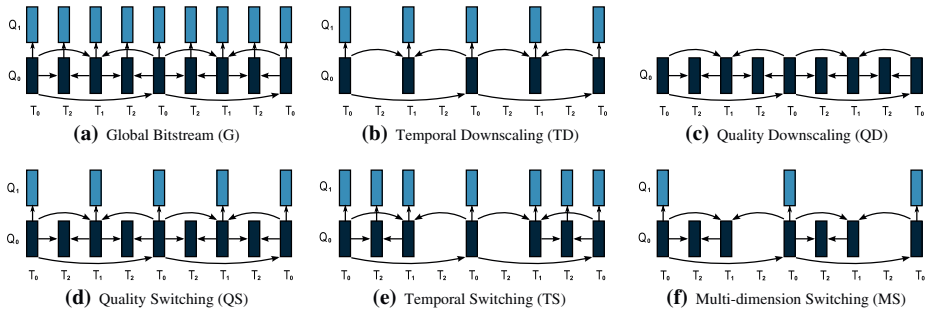


Fig. 1 Bitstream layout for downscaling and layer switching options used in the experiments. Q and T denote the quality and temporal dimensions, respectively

rate is not always more preferable than high image fidelity for high motion video. Probably closest to our work, Zink et al.'s evaluation has been performed to investigate quality degradation caused by variations in the amount of transmitted layers during streaming sessions [16]. The authors' results showed that the perceived quality of video is influenced by the amplitude and the frequency of layer switchings. In contrast to our work, they did not treat the layer switching and its related impairment in different dimensions separately. However, the resulting visual effects of quality impairment in temporal and quality dimensions are significant different and deserve an in-dept study. We identified the flickering and jerkiness as two peculiar effects caused by FLS in separate dimensions. Our work compare the two dimensional video impairment systematically and investigate how the visual effects are related to content, device and adaptation strategy.

The subjective tests by Cranley et al. [2] and Zink et al. [16] were conducted with regular monitors under lab conditions, which is different from our testing scenario defined for mobile video applications using iPods. Further, very few of previous studies performed subjective evaluation of the H.264 scalable extension. To the best of our knowledge, only in [3], a subjective field study about the H.264/SVC is introduced which also grounded our investigation presented in this paper.

3 Quality layer switching study

One of the main goals of our study is to see if our FLS can be used to achieve a more efficient fine-grained streaming solution compared to the high overheads of existing schemes. In this section, we show which operation points we have experimented with, identify possible quality reduction effects and describe the general subjective quality evaluation approach.

3.1 FLS

In contrast to adaptation approaches that downscale a SVC bitstream to a particular fixed operation point using fine-grain or mid-grain scalability, FLS alternates between two or multiple operation points in order to meet a given bitrate constraint over a short time-window without the extra overhead of defining additional operation points. For video with multi-dimensional scalability, layer switching is not limited to one single dimension. For instance, Fig. 1b, c shows two different approaches for downscaling. Moreover, Fig. 1d–f illustrates three different switching patterns, two that perform switching in a single dimension (temporal or quality) and one pattern that combines layer switching in the two multi-dimensions. Thus, FLS introduces intermediate scaling options, but it also causes two perceptible effects on the users QoE:

Flickering. Frequent switching between quality layers and spatial layers (at fullscreen resolution) can lead to a *flickering effect*. Flickering is characterized by rapid changes in edge blurriness and texture details or by repeated appearing of coding artifacts when a very low quality is displayed for a brief moment. Flickering is most visible in content with high details or when quality differences between operation points are large.

Jerkiness. Rapid changes in temporal resolution (frame rate) caused by temporal layer switching can be perceived as *jerkiness*. Jerkiness may even become visible if switching happens at frame rates that alone are regarded as sufficiently smooth [8]. Jerkiness is most visible in content with smooth global motion or low and natural local motion.

The choice of switching pattern and switching frequency is therefore of central importance due to the possible high impact on the perceived quality. Questions such as under

Table 1 Sequences used in the experiments

Genre	Content	Detail	Motion	Audio	CGS bit rate		MGS bit rate	
					Max	Min	Max	Min
Animation	BigBuckBunny	3.65	1.83	Sound	530.8	136.1	823.6	175.5
Cartoon	South Park	2.75	0.90	Speech	533.8	158.8	767.5	199.7
Docu	Monkeys & River	3.64	1.61	Sound	1,156.1	192.1	1,244.3	208.7
Movie	Dunkler See	1.85	0.58	Sound	255.2	67.9	419.9	92.4
News	BBC News	2.92	0.69	Speech	268.6	74.0	453.1	101.0
Sports	Free Ride	3.32	1.90	Music	734.8	121.1	745.9	129.1
HD-Animation	BigBuckBunny	2.88	4.13	Sound	10,457.0	1,032.4	14,210.0	1,021.7
HD-Docu	Canyon	3.09	3.33	Sound	25,480.0	2,407.0	28,940.0	2,394.0

Detail is the average of MPEG-7 edge histogram values over all frames [11] and motion is the MPEG-7 motion activity [6], i.e., the standard deviation of all motion vector magnitudes. Bit rates are given in kbit for the SVC bitstream at the highest enhancement layer (max) and the base layer (min)

which conditions (e.g., viewing context, display size and switching frequency) these effects become noticeable and how they influence the perceived quality impression are therefore important research issues, and to identify the feasibility of switching techniques and advice design constraints, we were interested in answering the following questions:

- Do people perceive a difference in quality between scaling and switching techniques?
- Is there a general preference of one technique over the other?
- Does a preference depend on genre, switching frequency or the scaling dimension?
- Are there frequencies and dimensions that are perceived as less disturbing?
- How general are our observations, i.e., do location, device type, display size and viewing distance influence the results?

3.2 Subjective quality evaluation

To answer the above question finding appropriate switching and scaling patterns, we have performed a set of subjective video quality evaluation experiments. In this study, we asked human observers for their preferences when watching video clip pairs.

To test different kinds of content with varying detail and motion, we selected eight sequences from different genres (see Table 1), i.e., six for the small mobile devices and two for the HDTV. We obtained the content from a previous study on scalable coding [3] which allowed for a comparison with earlier results. From each sequence, we extracted an 8-s clip without scene cuts. After extraction, the texture complexity and motion activity are measured according to MPEG-7 specification.

We encoded the SVC bitstreams with version 9.16 of the JSVM reference software.[1] The encoder was configured to generate streams in the scalable high profile with one base layer and one coarse-grained scalable or MGS enhancement layer, a GOP size of 4 frames with hierarchical B-frames, an intra period of 12 frames, inter-layer prediction and CABAC encoding. Note that SVC defines the set of pictures anchored by two successive key pictures together with the first key picture as a group of picture, where key pictures are usually encoded as P-frames within an intra period, see [13]. Due to the lack of rate control for quality enhancement layers in JSVM, we used fixed quantization parameters.

From the encoded SVC bitstreams, we extracted three scalable operation points with high variability in the bit rates (see Fig. 1a–c). The 'G' operation point (Fig. 1a) contains the full bitstream including the base layer (Q_0) and the quality enhancement layer (Q_1) at the original frame rate, while the other two operation points are each downscaled in a single dimension to the low-quality base layer at full temporal resolution (QD) or a lower temporal resolution T_1 (12 fps), but with quality enhancement (TD). These operation points were then used to generate streams with different switching patterns and to compare the switched streams' quality. Note that we only focused on quality scalability and temporal scalability in this study. We did not consider spatial scalability, because it is undesirable for FLS due to the large decrease in perceived quality as shown in previous subjective studies [3].

Next, we have performed experiments in three different scenarios: mobile displays in both private and public spaces and HDTV displays in private spaces trying to find suitable switching patterns from the downscaling operation

[1] Available at http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm.

points (Fig. 1b, c) resulting in patterns like the ones shown in Fig. 1d–f, i.e., better and more efficiently matching the available bit rates between the downscaling operation points giving better video quality than the lower base layer only.

4 Mobile scenario: field study 1

In our first experiment, we were interested in how user perception over FLS compared to static layer scaling. The experiment is performed in a private, in-door environment (lab), and each participant evaluated all the video content.

4.1 Experiment design

Three types of video quality assessment methodologies have been introduced in international recommendations such as ITU BT.500-11 [5] and ITU-T P.910 [4], namely Double Stimulus (DS), Single Stimulus (SS) and Pair Comparison (PC) methods. In DS method, assessors are asked to rate the video quality in relation to an explicit reference. In SS method, assessors only see and rate the quality of a single video with an arbitrary length. Both DS and SS methods use an ordinal grade scale and require assessors to give a rating from Bad (very annoying) to Excellent (imperceptible). In the PC method, a pair of video clips containing the same content in two different impairment versions is presented, and the assessor provides a preference for one version in each pair. The rating procedure of this method is simpler than that of DS and SS methods, and the comparative judgment can be easily verified by examining the transitivity of the ratings. In this paper, the comparison between layer switching and downscaling is of the most interest. Hence, the PC method suits best the context of our studies. We based our first experiment design on the standardized full factorial PC method (F/PC).

In our studies, we always compared one layer switching pattern against one static operation point. Each pair of patterns was presented twice during a test sequence, once in each possible order to assess the reliability of votes from each participant and detect inconsistent ratings. The order of all the pairs of a test sequence was a random permutation. Between subsequent pairs, there was a 6-s break, displaying a mid-grey image with black text that called for voting and announced the following clip. The participants were asked to judge whether they preferred the first or the second clip in the pair or whether they did not perceive a difference.

For each clip pair, we obtained a single measure about which clip a participant preferred to watch. If undecided, participants could also select that they had no preference. This resembles a repeated measurement design with three rating categories. We used all ratings from both clip-pair orders (AB, BA) in our analysis. We also included conflicting ratings, because they would just decrease significance, but not invalidate our results. For statistical analysis, we ran binomial tests to see if a significant majority of ratings for one of the preference categories existed.

4.1.1 Material

In this experiment, we tested video from all the six different genres listed in Table 1. The selected six sequences were downscaled and eventually cropped from their original resolution to QVGA (320 × 240) in order to fit the screen size of our display devices. Based on our previous experience and in order to obtain a perceivable quality difference, we selected quantization parameter 36 for the base layer and quantization parameter 28 for the enhancement layer. The switching periods that were chosen for this experiment were 0.08, 1 and 2 s.

4.1.2 Participants

Twenty-eight payed assessors (25% female) at mean age of 28 participated in the test. Among the assessors, 90% are at the age between 18 and 34 while 10% are at the age between 35 and 39. All of the assessors are college students with different education but no one has majored in multimedia technologies. All of the assessors are familiar with concepts such as digital TV and Internet video streaming while 75% of them claimed that media consumption is part of their daily life. We obtained a total of 2,016 preference ratings of which 44% indicated a clear preference (consistent ratings on both clip orders), 31% a tendency (one undecided rating) and 10% no difference (two undecided ratings). We observed 15% conflicting ratings, where participants gave opposite answers to a test pattern and its hidden check pattern. Participants with more than 1.5 times the inter-quartile range of conflicting ratings above the average were regarded as outliers. In total, we removed two outliers from our data set. Regardless of remaining conflicts we found statistically significant results.

4.1.3 Procedure

As mobile display devices, we used the iPod classic and the iPod touch from 2008. The two iPod models contain, respectively, a 2.5- and 3.5-in. display and have pixel resolutions of 320 × 240 and 480 × 320 at 163 pixel per inch. The selected display size is sufficient for depicting content at QVGA resolution according to [7]. All videos had an undistorted audio track to decrease the exhaustion of test participants.

Although quite a few of assessors have previous experience in watching video on handheld devices such as iPod, a brief introduction about how to operate the iPods during the experiments was given to the assessors prior to a test session. A whole test session lasted for about 1 h, including two short breaks. Each participant watched in total 144 clip pairs. During the session, the assessors were free to choose a comfortable watching position and to adjust the watching distance. For example, they could choose to sit on sofas or in front of a desk. They were also free to decide when they wanted to continue the test after a break.

4.2 Results

Results are reported as preference for layer switching or layer scaling with 0.01 confidence intervals. If a preference was found as not significant, we still give a weak tendency. Table 2 displays preferences between switching and scaling across genres, and Table 3 shows results for different period lengths. The 'all' line in Table 3 contains general results for all periods and all genres.

4.2.1 Temporal layer switching

Participant ratings indicate no clear preference when temporal switching (TS) is compared to temporal downscaling (TD). This is significant for all low-motion sequences where temporal resolution is less important to convey information, but not significant for other genres. Besides a weak tendency towards an undecided rating, a general conclusion is not possible.

Table 2 Private space mobile: quality preference per genre for layer switching versus downscaling (+ switching preferred, − downscaling preferred, ○ no preference, * not significant)

	TS		QS	
	TD	QD	TD	QD
Animation	○	+	−	(+)
Cartoon	(○)	+	−	(+)
Documentary	(+)	+	−	(−)
Short movie	○	+	−	(−)
News	○	+	−	(+)
Sports	(○)	+	−	(−)

Table 3 Private space mobile: quality preference over different switching periods for layer switching versus downscaling (symbols as in Table 2)

	TS		QS	
	TD	QD	TD	QD
80 ms			−	−
1 s	(○)	+	−	(+)
2 s	(○)	+	−	+
All	(○)	+	−	(+)

Empty cells are not covered by this study

One possible reason for this observation is that temporal resolution changes between 25 and 12 fps have a minor impact on quality perception. This confirms results of previous studies as reported in [3, 8]. Using more bandwidth for a temporally switched stream (92%) compared to a temporal downscaled stream (85%) is thus not justified by a significant increase in quality perception. We are currently investigating whether this observation also applies to switching to lower temporal resolutions (below 10 fps).

When layer switching in the temporal (TS) or quality dimension (QS) is compared to downscaling in the other dimension (QD and TD, respectively), the results indicate a clear preference towards decreasing the temporal resolution rather than the quality of a video. With high significance, our results are consistent across all genres and independent of the switching period. The result again confirms previous findings reported in [8]. People seem to be more sensitive to reductions in picture quality than to changes in frame rates when watching video on mobile devices. This clearly indicates that switching is a viable option for frequent temporal resolution changes. Although temporal base layers consume the main bit rate and potential savings are small, switching can still yield fine-grained adaptation in the upper bit-rate range of a stream.

For a fair comparison, it is noteworthy that the TS (92%) had a considerably higher bit rate than the low-quality operation point QD (28%). However, the quality of switching pattern QS (89%) compared to the lower temporal resolution TD (85%) shows that a lower bit-rate stream can yield a higher subjective quality regardless of the content.

4.2.2 Quality layer switching

When quality switching (QS) is compared to downscaling in the same dimension (QD), the combined results over all period sizes are not significant. There is also no general tendency towards a single adaptation technique that can be attributed to content characteristics alone. However, we observed a significant preference for quality layer switching at long periods while for shorter periods a preference for quality scaling exists.

We attribute this observation to a flickering effect that was perceived as disturbing by almost all participants. Flickering is caused by fast switching between high- and low-quality encodings which leads to rapid iteration of high- and low-frequency textures. At longer switching periods, this effect becomes less annoying and disappears

eventually. We call the limit at which flickering disappears the *flickering threshold*. Interestingly, long switching periods above the flickering threshold are also preferred to a constant low quality.

We just conducted tests with equally long intervals of high and low quality. Hence, the bit-rate demand of a QS scheme is still much higher than that of the low-quality operation point (89 vs. 28%). Asymmetric patterns with longer low-quality intervals will have a much lower bit-rate consumption and offer a wider range of bit-rate adaptation. We will investigate whether such patterns can also yield a better visual quality. We assume, however, that the flickering threshold plays an important role for asymmetric patterns as well.

5 Mobile scenario: field study 2

The second field study discussed in this section was conducted to verify the validity of the conclusions drawn in Sect. 4 by changing the way in which the user study itself was performed. We made a new method for performing the tests, and we moved the experiment location from a lab setting to a more realistic public space environment.

5.1 Experiment design

The primary concern that had arisen from the first study (Sect. 4) was the long duration of each participant's viewing time, about 1 h. Although participants had been allowed to change location and even to take breaks, they were generally annoyed with the test itself, and we were concerned that this can have had unpredictable effects on the quality of their evaluation. Furthermore, the video quality tests in our first study were mostly performed at Simula and on the university campus.

In order to perform tests with people with a more varied background in more realistic environments, we designed an evaluation method that is easy-to-use and less demanding to the participants. We named this test method as randomized PC (R/PC). R/PC is a flexible and economic extension to traditional pair comparison designs. Conventionally, it presents stimuli as pairs of clips. In contrast to traditional PC design that collects a full data sample for all pairs from every participant, R/PC uses random sampling to select small subsets of pairs and thus creates a shorter but unique experiment session for each participant. The randomization procedure in R/PC guarantees that all pairs get eventually voted for.

We designed our second field study with the R/PC method. In this study, participants were allowed to stop at anytime, viewing and evaluation were better integrated, and the test was performed in an everyday environment.

5.1.1 Material

In this field study, we used the same video material to generate our test sequences as in Sect. 4. We used only iPod touch devices from 2008 to perform the tests and used encoding settings that were similar to those of the first field study, except that the resolution was changed. Instead of scaling the video on the devices itself, all sequences were downscaled and cropped from their original resolution to 480×272 pixels in order to fit the 3.2-in. screen size of iPod touch and keep the 16:9 format.

We simulated layer switching in quality dimension (QS) and temporal dimension (TS) according to the patterns illustrated in Fig. 1d, e. The switching periods that were chosen for this experiment were 0.5, 1.5 and 3 s.

5.1.2 Participants

The field study was performed under conditions that differ from the first one in several ways. Participants were approached by students in public locations in Oslo in the summer and autumn. They were approached in situations that we considered realistic for the use of a mobile video system. We had 84 respondents, who had mostly been approached when they were idle, e.g., waiting for or sitting on a bus. They were asked for 15 min of their time.

Among the participants, 74% are between the age of 18 and 34, 20% are between the age of 35 and 59 and 6% are at the age under 18. 96% of the participants have normal visual acuity with or without glasses while 4% have limited visual acuity in spite of glasses. The field study was mostly conducted indoors (95%) in different locations (restaurant, bus station, cafeteria), while three participants were en-route and one person was outdoors. Using the same criterion introduced in Sect. 4, we gathered in total 2,405 ratings of which 30% indicated a clear preference (consistent ratings on both clip orders), 36.3% a tendency (one undecided rating), 24.4% no preference (two undecided ratings) and 8% conflicting ratings. Using the same criterion introduced in Sect. 4, we filter out three unreliable participants.

5.1.3 Procedure

Consistently with an experiment that was as close to the real world, we did not control lighting or sitting conditions. Participants were not protected from disturbances that are consistent with those that a user of a mobile video service would experience. They experienced distractions by passersby, or the urge to check departure times or the station for the next stop. In case of such a short-term disturbances, they were allowed to restart watching the same pair of clips.

Participants were not shown training sequences, but they received a brief introduction by the student, explaining that clips might look identical. The expected number of clips watched by a participant was 30, but considering the experience of fatigue and annoyance with the first experiment design and the situation of the participants, they could terminate the experiment at any time. The downside of this possibility was that the consistency of an individual participant's answers could not be checked, and that every vote for a clip pair needed to be considered an independent sample. Lacking the control mechanism, we required 20 or more votes for each clip pair. Following this method, participants were asked to assess the quality of two sequentially presented clips. A subset of clip pairs was randomly chosen for each participant from a base of 216 clip pairs (including reverse order for each pair). The quality changed once in each clip, either increasing or decreasing. The changes occurred at 2, 4 or 6 s.

The evaluation procedure was changed from the paper questionnaire approach taken in Sect. 4. This field study integrated both testing and evaluation into the iPod. Thus, users were given the opportunity to first decide whether they had seen a difference between the clips after each pair of clips that they had watched. If the answer was yes, they were asked to indicate the clip with higher quality.

5.2 Results

The results of the second field study are presented in the same way as those for the first study. Confidence intervals are reported as 0.01. Table 4 displays preferences between switching and scaling across genres, and Table 5 shows results for different period lengths. The 'all' line in Table 5 contains general results for all periods and all genres.

5.2.1 Temporal layer switching

Two series of ratings provided by the participants yielded results that were identical independent of genre. In the comparison of TS and TD in Table 4, our random, untrained participants did not favor either option for any type of content independent of motion speed in the clip. This makes it very clear that a frame rate difference of 25 versus 12 fps on a mobile device has minimal impact to the casual viewer. Additionally, TS is given a clear preference for all types of content of quality downscaling (QD). This repeats the equally clear findings of the first field study. Both of these comparisons stay the same when different switching periods are considered.

Table 4 Public space mobile: quality preference per genre for layer switching versus downscaling (symbols as in Table 2)

	TS		QS	
	TD	QD	TD	QD
Animation	○	+	−	+
Cartoon	○	+	(○)	(+)
Documentary	○	+	(○)	(○)
Short movie	○	+	(○)	○
News	○	+	−	(+)
Sports	○	+	(○)	(○)

Table 5 Public space mobile: quality preference over different switching periods for layer switching versus downscaling (symbols as in Table 2)

	TS		QS	
	TD	QD	TD	QD
500 ms	○	+	(○)	(+)
1.5 s	○	+	(○)	(+)
3 s	○	+	(−)	○
All	○	+	(−)	(○)

5.2.2 Quality layer switching

The preference that is given to TD over QS is detected much less clearly in the second field study than in the first. While TD was clearly preferred in the first study, the result is only clear for the animation clip with its sharp edges, and the news clip that has very little motion. For all other content types, the results are not statistically significant, but answers tend not to prefer either clip.

The comparison of QS and QD was similarly undecided for each of the different genres of clips as in the first field study. It can be mentioned that QD was never the preferred answer for any of the clips. QS was clearly preferred for the three contents that gave the participants the opportunity of focusing on quality rather than motion: the sharp-edged animation, the cartoon clip and the fairly static news clip. For the three clips with faster motion, participants tended not to prefer any clip.

Considering the different switching period for this series of tests, it is remarkable that participants did not prefer any clip when the switching period reached 3 s. This seems to indicate that users ignore quality changes at this longer time-scale.

6 HDTV scenario: field study 3

With respect to both environment and device, there are large differences between small mobile devices such as iPods and large, high-resolution devices like a 42-in. HDTV. The goal of our third experiment was to validate

whether the results obtained in the mobile scenarios are general observations or whether the results depend on the screen size and viewing distance.

6.1 Experiment design

As we did in the first experiment described in Sect. 4, we used the pair comparison method to test whether either the downscaling or the switching video adaptation options did significantly affect whether a user perceived the one or the other as better. The assessors could select if they preferred layer switching or layer downscaling, or if they had no preference. After gathering enough votes, we ran binomial tests to see if a significant majority of the ratings exist among the three rating categories.

6.1.1 Material

We prepared the test sequences in a similar way to our previous experiments. We encoded one base layer and one MGS enhancement layer using fixed quantization parameters of 36 and 20, respectively. The original spatial resolution of 1,920 × 1,080 was preserved in the selected two HD video sequences (see Table 1). The HD-Animation test sequence had the same content as the animation movie in the mobile tests. The HD-Docu sequence was extracted from the same documentary movie as the one used in our mobile scenario. But to fit the visual characteristics and potential for HDTV presentation, we selected a different part of the movie.

6.1.2 Participants

The study was conducted with 30 non-expert participants in a test room at Oslo University. All of them were colleagues or students between the age of 18 and 34. 3 of them claimed to have limited visual acuity even with glasses. In total, we gathered 720 preference ratings of which 49% indicated clear preference, 29% a tendency and 12% no preference. In the results, there were 10% conflicting ratings. We removed three outliers from our data set using the same criterion as that introduced in Sect. 4.1.2

6.1.3 Procedure

The visual setup was a 32-in., 1080p HDTV monitor. Our assessors were seated directly in line with the center of the monitor with a distance of about three monitor screen heights (3H distance). Since we conducted the test as a field study, we did not measure the environmental lighting in the test room, but the lighting condition was adjusted to avoid incident light being reflected from the screen. We displayed the video clip pairs in two different randomized orders. The duration of a whole continuous test session was 20 min and none of the accessors requested break during the test.

6.2 Results

In a similar way as in the two previous sections, the results of this study are reported with 0.01 confidence intervals. We demonstrate the correlations between the preferences, content genres and switching period lengths in Tables 6 and 7.

6.2.1 Temporal layer switching

Similar to what we found in mobile test scenarios, participant ratings do not indicate a clear preference when comparing temporal layer switching (TS) to TD. There is an indication that neither is preferred, but it is not possible to make a general conclusion.

When temporal layer switching (TS) is compared with downscaling in the other dimension (QD), preferences differ between genres.

The majority of our assessors preferred TS over QD when watching the animation video. Watching the Canyon clip, on the other hand, they indicated the opposite preference, which contradicts also all the results from the two mobile field studies. Also the combined results over all period length indicate a preference towards QD than TS. This preference is significant for shorter switching periods, while it weakens when the period reaches 3 s. This observation differs significantly from what we found out in mobile scenarios.

Table 6 HDTV scenario: quality preference per genre for layer switching versus downscaling (symbols as in Table 2)

	TS		QS	
	TD	QD	TD	QD
Animation	(○)	+	(+)	+
Canyon	(+)	−	+	(○)

Table 7 HDTV scenario: quality preference over different switching periods for layer switching versus downscaling (symbols as in Table 2)

	TS		QS	
	TD	QD	TD	QD
500 ms	(○)	−	+	(−)
1.5 s	(○)	−	+	(+)
3 s	(○)	(−)	+	(○)
All	(○)	−	+	(+)

6.2.2 Quality layer switching

In the HDTV scenario, people seem to be more sensitive to frame rate changes than quality loss at the picture level. When QS is compared to TD, participant ratings indicate a clear preference towards QS instead of TD, which is again different than the test results obtained from mobile scenarios. The results are consistent across genres and the preference of QS applies for different switching periods.

When layer switching is compared with downscaling in the single quality dimension (QS against QD), we do not find any significant results except for the animation content genre. However, the results show that the length of switching period affects the psychophysical video quality in a similar way both in HDTV and mobile scenarios. Namely, more people preferred QD than QS at short period because of the flickering effect. In the HDTV scenario, the period would be less than 500 ms. When the period was extended to a certain length such as 1.5 s, the flickering effect became less annoying. However, when the period was extended beyond a certain length such as 3 s in our experiments, most people became uncertain of their preference. One possible reason for this uncertainty is that people are able to detect video impairment that last longer than a certain interval, and they evaluate video quality by their worst experience within memory.

7 Discussion

In this section, we provide an analysis of the perceived quality of FLS and its usefulness to adapt to a given average bandwidth. We also take a critical look at the assessment methods itself.

7.1 Range of experiments

We have performed three field studies in order to understand whether people who watch video consider it beneficial to increase and decrease video quality frequently, and whether the answer to this question changes with the switching frequency. That it is beneficial to exploit available bandwidth to its fullest and adapt video quality quickly to use it, is an assumption that has frequently been made in the past. Through prefetching or buffering on the client side, even course- and medium-grained scalable video codecs would be able to come close to exploiting all available bandwidths in the long-term average.

Our investigations considered only options that are available in the toolset of SVC as implemented by the reference encoder. We considered bandwidth changes through temporal quality adaptation and through quality adaptation separately. We investigated only switching patterns where half of the frames belong to an upper and half to a lower operation point. A finer adaptation granularity can be achieved by adaptively turning this ratio, but the 8-s clip length used in our tests in accordance with the PC approach prevents an exploration of other ratios. When analyzing the results from all three studies, we found that preference indicators depend highly on the scenario.

7.2 Mobile devices

In our two field studies that examined mobile devices, we found that TS and also TD down to 12 fps result in better subjective quality than any type of quality layer reduction. When directly comparing switching versus downscaling in the temporal domain alone, no preference became apparent. Hence, temporal adaptation could be employed at any desired ratio in the observed range between 25 and 12 fps. The reason for this is that human observers regard all frame rates above a margin of 10 fps as sufficiently smooth, when they watch videos on small devices at typical viewing distances. These observations have been reported in earlier studies [3, 8] and were confirmed by us. The obvious conclusion from this observation is that it is not meaningful to encode videos for mobile devices at a higher frame rate than 12 fps.

For QS, the period length is a crucial design criteria. Very short periods (less than 0.5 s) should be avoided, because they introduce flickering at edges and in high-frequency textures. This observation strengthens the assumption that per-frame scaling decisions result in bad visual quality and should be avoided. QS above a period of 2 s, on the other hand, is perceived as having a similarly bad quality as downscaling. This implies that long periods of low quality are identified with constant bad quality by many viewers, meaning that there is either no significant preference or that undecidedness prevails.

7.3 Small versus large screens

The mobile test scenarios reveal a clear preference of TS over quality scaling regardless of content and switching period. In our investigation of HD screens, we found nearly the opposite picture. Therefore, people prefer a regular quality reduction over temporal jerkiness which, interestingly, becomes apparent on large screens even when the frame rate is reduced from 25 to 12 fps. The explanation for this can be found in the human visual system. Mobile devices are best viewed from 7 to 9.8 screen heights distance, which keeps the entire screen inside the visual focus area. HDTV screens, on the other hand, are best viewed from 3 screen heights distance, where the display still covers most of the human field-of-vision. This difference

influences the minimal required angular resolution of the human eye and foveal field-of-vision [7, 8].

Visual acuity in human's foveal field-of-vision decreases from the center towards the outside while sensitivity to motion effects increases [1, 9, 12]. On mobile screens, the complete screen is in the central high acuity region and therefore detail is resolved throughout the displayed image at almost the same fidelity. Frame rate is less important here. On HDTV screens, the image covers a larger region of the field-of-vision. Hence, humans focus on particular details within the image, which are seen with high acuity, while outer regions of the image cover the temporally sensitive area perceived in peripheral vision. Temporal abnormalities (jerkiness, jumping objects, flickering) are detected much easier and may even be annoying for the viewer.

7.4 Applicability of findings

The layer switching pattern must be supported by the SVC encoding structure and synchronized to the decoder operation to avoid prediction errors. The switching patterns used in our study assumed short GOP sizes and frequent intra-updates to allow for short switching periods. Due to inter-frame prediction, switching may not be possible at every frame boundary. FLS points are usually in conflict with practical encoder setups that use multiple reference pictures, long GOPs and rare intra-updates for increased coding efficiency. This requires a trade-off at encoding time.

The results of our studies are not limited to layer switching in the coarse-grain encoded versions of H.264/SVC streams alone. Any adaptation strategy in streaming servers, relaying proxies and playout software that can alternate between different quality versions of a video may benefit from our findings.

7.5 Usefulness of testing methods

For our tests, we used two different assessment methods, standardized full factorial PC (F/PC) and randomized PC (R/PC). Both have their particular problems. F/PC requires that test participants sit through long test sessions, which leads to fatigue and annoyance with the test itself. Test subjects are also experiencing learning effects; since the method requires the frequent repetition of the same content at different qualities, participants learn to focus on spots in the video that show quality differences best. The overall quality impression of the video clips is then no longer evaluated. Long test duration results in often high ratio of conflicting rating. For example, there are 15% conflicting ratings in our first study that lasted for about 1 h. Our second study was conducted in more interferential environments. But only 8% conflicting ratings were found due to shorter test duration at maximum 15 min.

R/PC avoids these problems and has many practical benefits. However, it requires a much larger number of participants who watch each pair clip. Through our intentional use in a noise and disruptive (but realistic) environment, R/PC test results did also tend towards undecidedness.

Finally, the explanatory power of both tests suffers from the requirement to use short clips to avoid memory effects. Especially when trying to answer questions about change frequency as we did in this paper, this is a strong limitation. We do therefore believe that we need new test methods that are suited for longer durations without increase in memory effects and fatigue.

8 Conclusion

We have investigated whether we can achieve fine-grained video scalability using coarse-grained H.264 SVC without introducing the high overhead of MGS in different streaming scenario including mobile TV and HDTV. This was tested by switching enhancement layers on and off to achieve the target bit rate between CGS operation points. We tested different switching patterns against different downscaling patterns, and our subjective tests indicate:

- Switching patterns with sufficient perceptual quality exist.
- Human perception of quality impairment in FLS is content and context specific.

For mobile devices, TS is shown to perform better than QD, but not better than TD. Hence, when bandwidth adaptation is required, the streamed video can select to first downscale its temporal resolution to an extent without introducing perceptual quality degradation. After that, QS and QD alone can be compared to determine whether FLS should be applied for additional bandwidth saving. The comparison of QS and QD on mobile devices shows that QS with an 80-ms period leads to a visually disturbing flickering effect, while QS above a 3-s period is not clearly preferable than QD. Between these points, however, QS, and thus FLS, has a beneficial effect that grows until a period length of 2 s.

For large screens, frequent temporal layer switching is generally undesirable, while the conclusions for QS are genre-dependent. At a switching period above 1 s, FLS is shown to improve perceptual quality for content with clear edges and little visual change, while FLS provides no clearly proven improvement for clips with fast visual changes.

In terms of resource consumption, both the TS (Fig. 1e) and QS (Fig. 1d) can achieve bit rates between the encoded SVC base layer and the enhancement layer. Both switching patterns were preferred over the quality downscaled operation point (QD, Fig. 1c). Thus, we claim that such fine-grained adaption is possible in different scenarios.

However, based on our preliminary tests, we cannot say which switching pattern will give the *best* possible result. This requires additional subjective studies. For example, we must further investigate the flickering threshold and the different ratios between high and low switching points. We need also understand how the detectability of jerkiness is related to content and context variations. In practice, popular HD videos are not only streamed to large display, but also can be watched on displays with smaller size. Additional studies can be done to verify if the same TD strategy also applies to HD video on smaller screens. At this point, we have also only tested clips without scene changes. To further limit the perceived quality degradation of switching techniques, scene changes can for example be used as switching points.

References

1. Beeharee, A.K., West, A.J., Hubbold, R.: Visual attention based information culling for distributed virtual environments. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST '03, ACM, New York, NY, USA, pp. 213–222 (2003)
2. Cranley, N., Perry, P., Murphy, L: User perception of adapting video quality. Int. J. Human–Computer Stud. **64**(8), 637–647 (2006)
3. Eichhorn, A., Ni, P.: Pick your layers wisely—a quality assessment of H.264 scalable video coding for mobile devices. IEEE Int. Conf. Commun., pp. 1019–1025 (2009)
4. International Telecommunications Union. ITU-T P.910. Subjective video quality assessment methods for multimedia applications (1999)
5. International Telecommunications Union—Radiocommunication sector. ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television picture (2002)
6. Jeannin, S., Divakaran, A.: MPEG-7 visual motion descriptors. IEEE Trans. Circuits Syst. Video Technol. **11**(6), 720–724 (2001)
7. Knoche, H.O., Sasse, M.A.: The sweet spot: how people trade off size and definition on mobile devices. In: Proceeding of the 16th ACM International Conference on Multimedia, MM '08, ACM, New York, NY, USA, pp. 21–30 (2008)
8. McCarthy, J.D., Sasse, M.A., Miras, D.: Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 535–542 (2004)
9. Nadenau, M.J., Winkler, S., Alleysson, D., Kunt, M.: Human vision models for perceptually optimized image processing—a review. Proc. IEEE (2000) (submitted)
10. Ni, P., Eichhorn, A., Griwodz, C., Halvorsen, P.: Fine-grained scalable streaming from coarse-grained videos. In: Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video, NOSSDAV '09, ACM, New York, NY, USA, pp. 103–108 (2009)
11. Park, D.K., Jeon, Y.S., Won, C.S.: Efficient use of local edge histogram descriptor. In: Proceedings of ACM Workshops on Multimedia, pp. 51–54 (2000)
12. Rix, A.W., Bourret, A., Hollier, M.P.: Models of human perception. BT Technol. J. **7**(1), 24–34 (1999)
13. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable extension of the H.264/AVC video coding standard. IEEE Trans. Circuits Syst. Video Technol. **17**(9), 1103–1120 (2007)
14. Wenger, S., Ye-Kui, W., Schierl, T.: Transport and signaling of SVC in IP networks. IEEE Trans. Circuits Syst. Video Technol. **17**(9), 1164–1173 (2007)
15. Wu, H., Claypool, M., Kinicki, R.: On combining temporal scaling and quality scaling for streaming MPEG. In: Proceedings of NOSSDAV, pp. 1–6 (2006)
16. Zink, M., Künzel, O., Schmitt, J., Steinmetz, R.: Subjective impression of variations in layer encoded videos. In: Proceedings of International Workshop on Quality of Service, pp. 137–154 (2003)

Appendix E

Paper IV

Title: Flicker Effects in Adaptive Video Streaming to Handheld Devices

Authors: Ni, Pengpeng, Ragnhild Eg, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen

Published: In ACM International Multimedia Conference (ACM MM), 2011

Abstract: Streaming video over the Internet requires mechanisms that limit the streams' bandwidth consumption within its fair share. TCP streaming guarantees this and provides lossless streaming as a side-effect. Adaptation by packet drop does not occur in the network, and excessive startup latency and stalling must be prevented by adapting the bandwidth consumption of the video itself. However, when the adaptation is performed during an ongoing session, it may influence the perceived quality of the entire video and result in improved or reduced visual quality of experience. We have investigated visual artifacts that are caused by adaptive layer switching – we call them *flicker effects* – and present our results for handheld devices in this paper. We considered three types of flicker, namely noise, blur and motion flicker. The perceptual impact of flicker is explored through subjective assessments. We vary both the intensity of quality changes (*amplitude*) and the number of quality changes per second (*frequency*). Users' ability to detect and their acceptance of variations in the amplitudes and frequencies of the quality changes are explored across four content types. Our results indicate that multiple factors influence the acceptance of different quality variations. Amplitude plays the dominant role in delivering satisfactory video quality, while frequency can also be adjusted to relieve the annoyance of flicker artifacts.

Flicker Effects in Adaptive Video Streaming to Handheld Devices

Pengpeng Ni[1,2], Ragnhild Eg[1,3], Alexander Eichhorn[1],
Carsten Griwodz[1,2], Pål Halvorsen[1,2]
[1]Simula Research Laboratory, Norway
[2]Department of Informatics, University of Oslo, Norway
[3]Department of Psychology, University of Oslo, Norway

ABSTRACT

Streaming video over the Internet requires mechanisms that limit the streams' bandwidth consumption within its fair share. TCP streaming guarantees this and provides lossless streaming as a side-effect. Adaptation by packet drop does not occur in the network, and excessive startup latency and stalling must be prevented by adapting the bandwidth consumption of the video itself. However, when the adaptation is performed during an ongoing session, it may influence the perceived quality of the entire video and result in improved or reduced visual quality of experience. We have investigated visual artifacts that are caused by adaptive layer switching – we call them *flicker effects* – and present our results for handheld devices in this paper.

We considered three types of flicker, namely noise, blur and motion flicker. The perceptual impact of flicker is explored through subjective assessments. We vary both the intensity of quality changes (*amplitude*) and the number of quality changes per second (*frequency*). Users' ability to detect and their acceptance of variations in the amplitudes and frequencies of the quality changes are explored across four content types. Our results indicate that multiple factors influence the acceptance of different quality variations. Amplitude plays the dominant role in delivering satisfactory video quality, while frequency can also be adjusted to relieve the annoyance of flicker artifacts.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems – *Human factors*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *Video*

General Terms

Experimentation, Human Factors

Keywords

Subjective video quality, Video adaptation, Layer switching

*Area Chair: Wu-chi Feng

MM'11, November 28-December 1, 2011, Scottsdale, Arizona, USA.

1. INTRODUCTION

To cope with the Internet's varying bandwidth, many video streaming systems use adaptive and scalable video coding techniques to facilitate transmission. Furthermore, transfer over TCP is currently the favored commercial approach for on-demand streaming [1, 11, 14, 19] where video is progressively downloaded over HTTP. This approach is not hampered by firewalls, and it provides TCP fairness in the network as well as ordered, lossless delivery. Adaptation to the available bandwidth is controlled entirely by the application.

Several feasible technical approaches for performing adaptation exist. One frequently used video adaptation approach is to structure the compressed video bit stream into layers. The based layer is a low-quality representation of the original video stream, while additional layers contribute additional quality. Here, several scalable video codec alternatives exist, including scalable MPEG (SPEG) [6], Multiple Description Coding (MDC) [4] and the Scalable Video Coding (SVC) extension to H.264 [17]. The other alternative is to use multiple independent versions encoded using, for example, the advanced video coding (AVC) [8], which supports adaptation by switching between streams [1, 11, 14, 19]. Thus, video streaming systems can adaptively change the size or rate of the streamed video (and thus the quality) to maintain continuous playback and avoid large start-up latency and stalling caused by network congestion.

Making adaptation decisions that achieve the best possible user perception is, on the other hand, an open research field. Current video scaling techniques allow adaptation in either the spatial or temporal domain [17]. All of the techniques may lead to visual artifacts every time an adaptation is performed. An algorithm must take this into account and, in addition, it must choose the time, the number of times, and the intensity of such adaptations.

This paper reports on our investigation of the types of visual artifacts that are specific for frequent bandwidth adaptation scenarios:

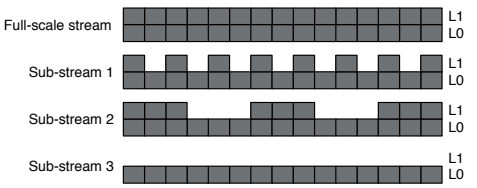
- **Noise flicker** is a result of varying the signal-to-noise-ratio (SNR) in the pictures. It is evident as a recurring transient change in noise, ringing, blockiness or other still-image artifacts in a video sequence.
- **Blur flicker** is caused by repeated changes of spatial resolution. It appears as a recurring transient blur that sharpens and unsharpens the overall details of some frames in a video sequence.
- **Motion flicker** comes from repeated changes in the

video frame rate. The effect is a recurring transient judder or jerkiness of naturally moving objects in a video sequence.

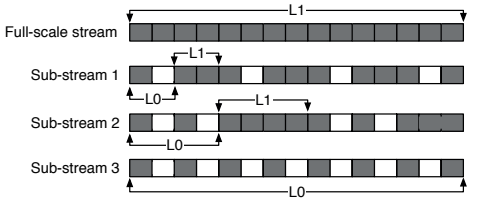
When the frequent quality fluctuations in the streamed video are perceived as flicker, it usually degrades the experienced subjective quality. However, noise, blur and motion flicker as such can not be considered deficient. Active adaptation to changes in available bandwidth is generally preferable to random packet loss or stalling streams, and not every quality change is perceived as a flicker effect. Essentially, the perceptual effect of flicker is closely related to the *amplitude* and *frequency* of the quality changes. This paper explores the acceptability of flicker for a handheld scenario.

In figure 1, we show sketches of simple streaming patterns for both spatial and temporal scaling. Figure 1(a) depicts a video stream encoded in two layers; it consists of several subsequent segments, where each segment has a duration of t frames. The full-scale stream contains two layers (L0 and L1), and the low quality stream (sub-stream 3) contains only the lower layer (L0), it is missing the complete L1 layer. For these, the number of layers remains the same for the entire depicted duration, meaning that neither of the two streams flickers. The other two examples show video streams with flicker. The *amplitude* is a change in the spatial dimension, in this example the size of the L1 layer (in other scenarios, this may be the number of layers). The *frequency* determines the quality change period, i.e., how often the flicker effect repeats itself. In this example, sub-stream 1 changes its picture quality every t frames (2 blocks in the figure), whereas sub-stream 2 changes every $3t$ frames (6 blocks in the figure). Figure 1(b) shows a similar example of how the amplitude and frequency affect the streaming patterns in the temporal dimension. Here, the *amplitude* is a change in the temporal dimension. In this example, we index video segments by their temporal resolutions since only temporal scalability is in our concern. The full-scale stream can be displayed at a normal frame rate. Sub-stream 3 drops frames regularly and can be displayed at a constant low frame rate. Neither of the two streams flickers in the temporal dimension. Hence, we say that the full-scale stream contains layer L1, whereas sub-stream 3 contains only layer L0. Sub-stream 1 and 2 halve the normal frame rate at a regular interval of $2t$ and $4t$ time units, respectively. Therefore, the layer variations in sub-streams 1 and 2 have the same amplitude, but the changes appear at different frequencies.

To provide the best possible video quality for a given available bandwidth, the applications need to select the most suitable options from several streaming patterns. Considering the alternatives in figures 1(a) and 1(b), three substream alternatives can be used if the full quality stream cannot be provided. Therefore, to get a better understanding of human quality perception of *flicker*, we have performed a subjective field study with a special focus on handheld devices. We have considered state-of-the-market encoding techniques represented by the H.264 series of standards. Our goals are (1) to evaluate the influence of the main influential factors on acceptability, and (2) to find the range of these factors' levels. With these answers we hope to minimize the flicker effect in layer variation. We evaluate the effect of noise, blur and motion flicker on four different types of video content. For each video type, we tested several levels of frequency and amplitude. In total, we performed 5088 individual assessments.



(a) Scaling in spatial dimension



(b) Scaling in temporal dimension

Figure 1: Illustration of streaming patterns for scalable video.

From our results, we observe that the perception of quality variation is jointly influenced by multiple factors. Amplitude and frequency have significant impact on subjective impression. Most notably, when decreasing the quality switching frequency for flicker in the spatial domain, including noise and blur flickers, users' acceptance scores of the video quality tend to be higher. Moreover, the different flicker and content types are found to influence perceived quality in their own ways.

The paper is structured as follows. The experiment design is presented in section 3. Section 4 analyzes user responses and reports the analytical results. In section 5, we discuss our findings. Finally, section 6 concludes the paper.

2. RELATED WORK

To the best of our knowledge, very little work considers the flicker effect in the video quality domain. In [16], the National Telecommunications and Information Administration General Model (NTIA GM) introduced combined measures for the perceptual effects relating to different types of impairments, such as, blurriness, blockiness, jerkiness, etc. Kim et al. [9] proposed a scalability-aware video quality metric, which incorporated spatial resolution with frame rate and SNR distortion into a single quality metric. However, none of these objective metrics have considered the temporal variation of different impairments. Some subjective tests evaluate the visual quality of scalable video; for instance, the effect of quality degradation in the temporal and spatial dimensions is explored in [10, 12, 13]. The closest related work [20], points out that the frequency and amplitude of layer changes influence the perceived quality and should therefore be kept as small as possible. However, that user study limits itself to SNR scalability and does not take the influence of video content characteristics into account.

3. EXPERIMENT DESIGN

3.1 Randomized Block Design

We conduct subjective experiments to explore the impact of noise, blur and motion flicker on the perception of video quality. In addition to the three different adaptation domains (SNR for noise flicker, spatial resolution for blur flicker and temporal resolution for motion flicker), the overall video quality is influenced by other factors including amplitude, frequency and content characteristics (see section 3.2). All of these are design factors studies in our experiment. We do not limit ourselves to a single genre of video content, but we do not aspire to cover all semantic categories. We explore four content types, which are selected as representatives for extreme values of low and high spatial and temporal information content. In our experiments, the subjects are asked to rate their acceptance of the overall video quality. Due to the fluctuating state of videos that flicker, we predict flicker to be perceived differently than other artifacts. We add a Boolean score on perceived stability, which we expect to provide us with more insight into the nature of the flicker effects (see section 3.4). Finally, we measure participants' response time, which is the time between the end of a video presentation and the time when they provide their response.

The repeated measures design [2] of these experiments ensures that each subject is presented with all stimuli. The repeated measures design offers two major advantages: First, it provides more data from fewer people than, e.g., pairwise comparison studies. Second, it makes it possible to identify the variation in scores due to individual differences as error terms. Thus, it provides more reliable data for further analysis. This study employs an alternative to the traditional full factorial repeated-measures design that is called Randomized Block Design. It blocks stimuli according to flicker type and amplitude level. A stimuli block consists of a subset of test stimuli that share some common factor levels and can be examined and analyzed alone. Stimuli are randomized within each block and blocks are randomized to an extent that relies solely on the participant, as they are free to choose which block to proceed with.

The randomization of stimuli levels ensures that potential learning effects are distributed across the entire selection of video contents and frequency levels, and, to a degree, also amplitudes and flicker type. Moreover, we hope to minimize the effect of fatigue and loss of focus by dividing stimuli into smaller blocks and allowing participants to complete as many blocks as they wish, with optional pauses between blocks.

3.2 Content Selection and Preparation

As the rate distortion performance of compressed video depends largely on the spatial and temporal complexity of the content, the flicker effect is explored across four content types at different extremes. Video content is classified as being high or low in spatial and temporal complexity, as recommended in [7] and measured by spatial information (SI) and temporal information (TI) metrics, respectively. Four content types with different levels of motion and detail are selected based on the metrics (figure 2). To keep the region of interest more global and less focused on specific objects, we avoid videos with dictated points of interest, such as a person speaking. It is beyond the scope of the current investigation to generalize the results to all video content.

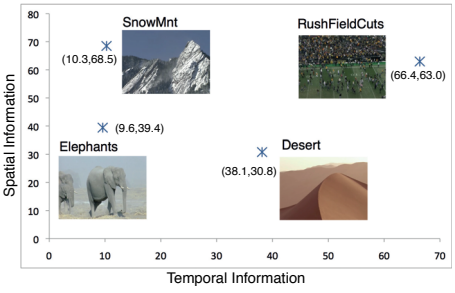


Figure 2: Test sequences.

Raw video material is encoded using the H.264/SVC reference software, JSVM 9.19, with two-layer streams generated for each type of flicker, as portrayed in figure 1. The amplitude levels of the layer variations are thus decided by the amount of impairment that separates the two layers. Table 1 summarizes the factor levels of amplitude, frequency, and content, according to the different flicker stimuli, noise, blur, and motion. For noise flicker stimuli, constant quantization parameters (QP) are used to encode a base layer L0 and an enhancement layer L1. Since the latter is encoded with QP24 for all test sequences, the amplitude levels and variations in video quality are represented by the QPs applied to L0 for noise flicker stimuli. Similarly, with blur flicker stimuli, amplitude is represented by spatial downscaling in L0, and temporal downscaling in L0 defines the amplitude for motion flicker stimuli.

To simulate the different flicker effects that can arise in streamed video, video segments from the two layers are alternately concatenated. Different frequencies of layer variation are obtained by adjusting the duration of the segments. For simplicity, we use only periodic duration. Corresponding to six frequency levels, six periods in terms of the L1 frame rate are selected, which include 6, 10, 30, 60, 90 and 180 frames for both noise and blur flicker stimuli. Since short durations for changes in frame rate are known to lead to low acceptance scores [13], the periods for motion flicker stimuli are limited to 30, 60, 90 and 180 frames.

a) Noise flicker

Amplitude	L1	QP24
	L0	QP28, QP32, QP36, QP40
Period	6f, 10f, 30f, 60f, 90f, 180f	
Content	RushFieldCuts, SnowMnt, Desert, Elephants	

b) Blur flicker

Amplitude	L1	480x320
	L0	240x160, 120x80
Period	6f, 10f, 30f, 60f, 90f, 180f	
Content	RushFieldCuts, SnowMnt, Desert, Elephants	

c) Motion flicker

Amplitude	L1	30fps
	L0	15fps, 10fps, 5fps, 3fps
Period	30f, 60f, 90f, 180f	
Content	RushFieldCuts, SnowMnt, Desert, Elephants	

Table 1: Selection of factor levels

3.3 Participants

In total, 28 participants (9 female, 19 male) were recruited at the University of Oslo, with ages ranging from 19 to 41 years (mean 24). They volunteered by responding to posters on campus with monetary compensation rewarded to all. Every participant reported normal or corrected to normal vision.

3.4 Procedure

This field study was conducted in one of the University of Oslo's library with videos presented on 3.5-inch iPhone of 480x320 resolution and brightness levels at 50%. Participants were free to choose a seat among the available lounge chairs but were asked to avoid any sunlight. They were told to hold the device at a comfortable viewing distance and to select one of the video blocks to commence the experiment. The 12-second long video segments were presented as single-stimulus events, in accordance with the ITU-T Absolute Category Rating method [7]. Each video stimulus was displayed only once. Video segments were followed by two response tasks, with responses made by tapping the appropriate option on-screen. For the first, participants had to evaluate the perceived stability of the video quality by answering "yes" or "no" to the statement "I think the video quality was at a stable level". The second involved an evaluation of their acceptance of the video quality, where they had to indicate their agreement to the statement "I accept the overall quality of the video" on a balanced 5-point Likert scale. The Likert scale includes a neutral element in the center and two opposite extreme values at both ends. A positive value can be interpreted as an acceptable quality level, a neutral score means undecidedness, while a negative score indicates an unacceptable quality level. Upon completion of a block, participants could end their participation, have a short break, or proceed immediately to the next block. Participants spent between 1.5 and 2 hours to complete the experiment.

4. DATA ANALYSIS

4.1 Method of Analysis

The current study explores the influence of amplitude and frequency of video quality shifts for three types of flicker stimuli, noise, blur and motion, as well as video content characteristics, on the perception of stability, the acceptance of video quality and response time. Control stimuli with constant high or low quality are included as references to establish baselines for the scores provided by participants. Stability scores and rating scores are processed separately, grouped according to flicker type. Thus responses are analyzed in six different groups, with control stimuli included in all of them. Since the perception of stability relies on detection, scores are binary and are assigned the value "1" for perceived stability of quality, and the value "0" for the opposite. Rating scores are assigned values ranging from -2 to 2, where "2" represents the highest acceptance, "0" the neutral element, and "-2" the lowest acceptance.

Consistency of acceptance scores is evaluated by comparing scores for control stimuli of constant high or low quality. Whenever a low quality stimulus scores better than the corresponding high quality stimulus, this is counted as a conflict. Conflicts are added up for each participant. If the acceptable number of conflicting responses is exceeded, the participant is excluded as an outlier. An acceptable number of conflicts stays within 1.5 times the interquartile range around the mean as suggested by [2, 3]. For the blur stimuli group this excluded two participants (12.5%), two for the motion stimuli group (10.5%), and none for the noise stimuli group.

Consistency of response times is also evaluated in order to eliminate results that reflect instances in which participants may have been distracted or taken a short break. Thus, any response time above three standard deviations of a participant's mean is not included in the following analyses.

Stability scores are analyzed as ratios and binomial tests are applied to establish statistical significance. As for acceptance scores, these are ordinal in nature and are not assumed to be continuous and normally distributed. They are therefore analyzed with the non-parametric Friedman's chi-square test [18]. The Friedman test is the best alternative to the parametric repeated-measures ANOVA [5], which relies on the assumption of normal distribution; it uses ranks to assess the differences between means for multiple factors across individuals. Main effects are explored with multiple Friedman's chi-square tests, applied to data sets that are collapsed across factors. Confidence intervals are calculated in order to further investigate the revealed main effects, assessing the relations between factor levels. Multiple comparisons typically require adjustments to significance levels, such as the Bonferroni correction. Yet, such adjustments can increase the occurrence of Type II errors, thus increasing the chances of rejecting a valid difference [15]. In light of this, we avoid the use of adjustments and instead report significant results without corrections. This procedure requires caution; we avoid drawing definite conclusions and leave our results open to interpretation. Repeated-measures ANOVA tests are finally introduced when analyzing response times.

4.2 Response Times

None of the repeated-measures ANOVA tests reveals any effect of amplitude, frequency or content on response time, for any type of flicker. In fact, response times seem to vary randomly across most stimuli levels. Possibly, this may be related to individual effort in detecting stability. If so, the video quality variation did not increase the decision-making effort. We may even surmise that participants evaluated the stability of video quality with a fair degree of confidence.

4.3 Noise Flicker Effects

The perceived stability of noise flicker stimuli is generally low and varies little over the different periods, as seen in table 2(a). However, the response percentage reflecting stable video quality is slightly higher for video segments of 180 frames. A significantly larger share of responses for the control stimuli reports video quality to be stable, as opposed to unstable, refer to the top and bottom lines in table 2(a). Due to the small difference between layers for QP28, it is plausible that the vast majority of participants do not perceive the flicker effect, which would explain why two thirds report stable quality, see the top line in table 2(b). Meanwhile, the higher rate of reported stability for non-flicker stimuli fits well with predictions. It indicates that participants detect and identify flicker as instability, whereas constant quality is experienced as stable, even when it is poor.

a) Period

Options	Stable	Unstable	P-value	Signif.
HQ	95.3%	04.7%	2.04e-71	+
6f	30.6%	69.4%	3.32e-12	–
10f	30.0%	70.0%	6.18e-13	–
30f	30.3%	69.7%	1.44e-12	–
60f	31.6%	68.4%	3.71e-11	–
90f	32.5%	67.5%	3.65e-10	–
180f	41.2%	58.8%	0.002	–
LQ	71.3%	28.7%	1.80e-14	+

b) Amplitude

Options	Stable	Unstable	P-value	Signif.
QP28	65.8%	34.2%	3.66e-12	+
QP32	27.7%	72.3%	4.49e-23	–
QP36	21.7%	78.3%	3.51e-37	–
QP40	15.6%	84.4%	8.74e-56	–

Table 2: Perceived quality stability for Noise flicker (+ Stable, - Unstable, (*) not significant), HQ = constant high quality, LQ = constant low quality.

Main effects are found with Friedman's chi-square tests for period ($\chi^2(5) = 69.25, p < .001$), amplitude ($\chi^2(3) = 47.98, p < .001$) and content ($\chi^2(3) = 27.75, p < .001$). The means and confidence intervals presented in figure 3(a) show that acceptance scores become increasingly higher than the constant low quality controls for periods of 60 frames and above. Figure 3(b) displays the decrease in acceptance with larger amplitudes, while figure 3(c) shows only small variations in acceptance scores depending on content type. As for potential interactions, figure 4 illustrates how mean acceptance scores vary across content types, with a tendency to increase as amplitude decreases or period increases. Moreover, the scores point to possible interactions, particularly between period and amplitude.

4.4 Blur Flicker Effects

For blur flicker stimuli, perceived video quality stability is again low across the different periods, accompanied by high perceived stability ratios for control stimuli, summarized in table 3(a). Furthermore, participants tend to judge the video quality as unstable at both amplitude 240x160 and amplitude 120x80, see table 3(b). This is also consistent with expectations, suggesting again that flicker is detectable and perceived to be unstable.

a) Period

Options	Stable	Unstable	P-value	Signif.
HQ	100%	00.0%	3.85e-34	+
6f	11.6%	88.4%	1.50e-17	–
10f	11.6%	88.4%	1.50e-17	–
30f	11.6%	88.4%	1.50e-17	–
60f	13.4%	86.6%	7.12e-16	–
90f	12.5%	87.5%	1.08e-16	–
180f	17.0%	83.0%	6.75e-13	–
LQ	81.2%	18.8%	1.42e-11	+

b) Amplitude

Options	Stable	Unstable	P-value	Signif.
240x160	19.3%	80.7%	4.89e-31	–
120x80	06.6%	93.5%	2.57e-67	–

Table 3: Perceived quality stability for Blur flicker (+ Stable, - Unstable, (*) not significant).

Friedman's chi-square tests reveal main effects for period ($\chi^2(6) = 41.79, p < .001$), amplitude ($\chi^2(1) = 14.00, p < .001$) and content ($\chi^2(3) = 33.80, p < .001$). As seen in figure 5(a), the mean acceptance scores are generally low across periods, only at 60 frames and above do they approach the acceptance of constant low quality. Moreover, there are little variations in acceptance according to amplitude and content, see figures 5(b) and 5(c). However, figure 6 illustrates how the differences in acceptance scores become greater when considering interactions. Similar to noise flicker, acceptance tends be higher for longer periods, but more markedly for the amplitude 240x160. Also acceptance scores for the Desert and Elephants clips appear to be higher than the RushFieldCuts and SnowMnt clips.

4.5 Motion Flicker Effects

Low perceived stability ratios are evident across all periods for motion flicker stimuli, presented in table 4(a). As expected, the vast majority of participants think that the video quality is stable for constant high quality control stimuli but not for constant low quality; there are more responses that correspond to perceived instability for low quality control stimuli. This is potentially explained by the lack of fluency of movement that occurs at lower frame rates. The stability scores for amplitude may also reflect a bias towards reporting jerkiness as instability, as listed in table 4. However, stability is reported more frequently for larger periods and better frame rates; this indicates influences from both period and amplitude on perceived quality stability.

Friedman's chi-square tests uncover main effects for all factors, including period ($\chi^2(3) = 7.82, p < .05$), amplitude ($\chi^2(3) = 41.62, p < .001$), and content ($\chi^2(3) = 27.51, p < .001$). However, the main effect for period is very close to the significance threshold (p=0.0499), which is likely the reason for the relatively flat distribution of acceptance scores observed in figure 7(a). Amplitude and content type, on the other hand, have larger effects on quality acceptance, as seen in figures 7(b), 7(c) and 8.

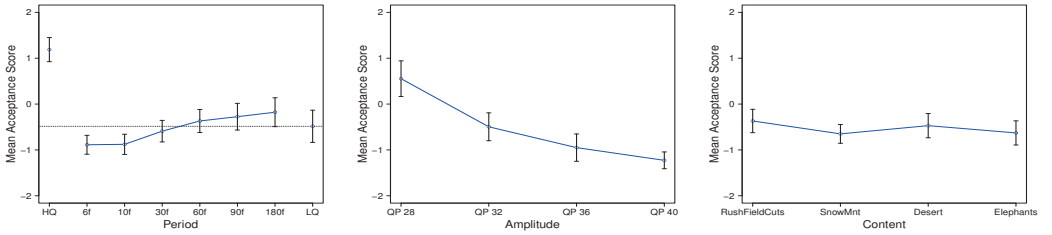
a) Period

Options	Stable	Unstable	P-value	Signif.
HQ	90.8%	09.2%	4.43e-47	+
30f	14.3%	85.7%	7.85e-35	–
60f	16.2%	83.8%	4.08e-31	–
90f	18.0%	82.0%	1.08e-27	–
180f	20.6%	79.4%	2.44e-23	–
LQ	40.8%	59.2%	0.0029	–

b) Amplitude

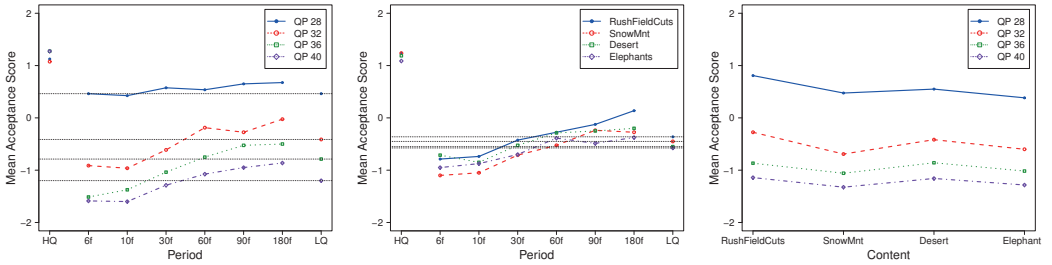
Options	Stable	Unstable	P-value	Signif.
15fps	43.8%	56.2%	0.045	(*)
10fps	15.1%	84.9%	2.62e-33	–
5fps	07.4%	92.6%	2.82e-52	–
3fps	02.9%	97.1%	1.82e-67	–

Table 4: Perceived quality stability for Motion flicker (+ Stable, - Unstable, (*) not significant).



(a) (b) (c)

Figure 3: Effects of period, amplitude and content on Noise flicker stimuli. Error bars represent 95% confidence intervals.



(a) Period effects according to amplitude (b) Period effects according to content (c) Content effects according to amplitude

Figure 4: Explored interactions for Noise flicker. (HQ = constant high quality, LQ = constant low quality)

5. DISCUSSION

5.1 Period Effect

The period of flicker is a major influential factor for flicker in the spatial dimension. Significant differences between acceptance scores given to different periods in noise flicker can be found in figure 3(a), and for blur flicker in figure 5(a). In figures 6(a) and 6(b), we can highlight three period ranges that influence the overall quality acceptance: low acceptance for short periods, acceptance higher than the low-quality control stimuli for moderate periods, and stagnating for long periods. Stagnation is less pronounced in figures 4(a) and 4(b).

In figure 4(b), the average across all amplitudes is shown for individual contents, reinforcing that the effect is independent of the content. At high frequencies ($< 30f$ or $< 1sec$ respectively), the flicker is perceived as more annoying than constant low quality for all different content types. Starting at moderate frequencies ($30 \sim 60f$ or $1 \sim 2s$), the quality is considered as better than a constant low quality for some content types. At low frequencies ($> 60f$ or $> 2s$), the quality is in most cases regarded as better than a constant low quality. For both flicker types in the spatial dimension, this is significant across amplitudes (figures 4(a) and 6(a)), content (figures 4(b) and 6(b)), but counter-examples exist (see the top line in figure 6(a)).

In the temporal dimension, the period does not seem to have a significant influence on the motion flicker. There are only small differences between acceptance scores for different periods, ranging from $30f$ to $180f$ (see figures 7(a), 8(a) and 8(b)). When the amplitude of temporal downscaling is small, scores are higher than for the low-quality control stimuli (figures 8(a), 10(a)). No period ranges can be highlighted.

A general observation for all three flicker types is that adaptive video streaming can outperform constant low quality streams, but the switching period must be considered in relation to the flicker amplitudes.

5.2 Amplitude Effect

The amplitude is the most dominant factor for the perception of flicker. This seems reasonable since the visual artifacts become more apparent with increasing amplitude when alternating between two quality versions. Our statistical results, presented in section 4, show this and evaluate the strength of the influence. The noise flicker effect is not detectable for the majority of our participants (see Q28 in table 2(b)) at low flicker amplitudes, where visual artifacts are less obvious. In the case of motion flicker, close to 50% of the responses show that changes between frame rates of 15fps and 30fps are not detectable. When the amplitude grows, meaning that the lower frame rate is reduced further, the detectability of quality fluctuation grows as well (see table 4(b)). The detectability shows the same changing trend for noise and blur flicker. The effect of flicker at different period lengths becomes significant only if the flicker artifacts are clearly detectable from the increase of flicker amplitude.

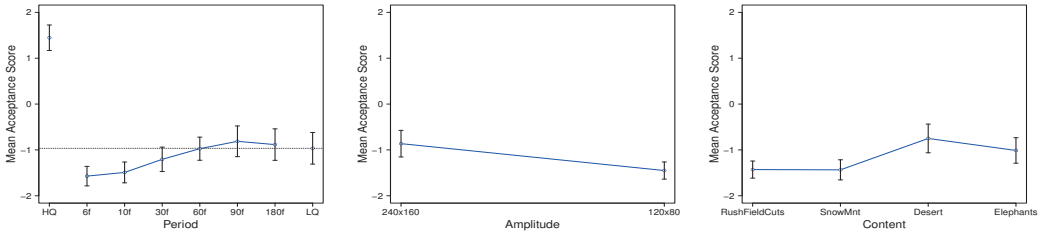
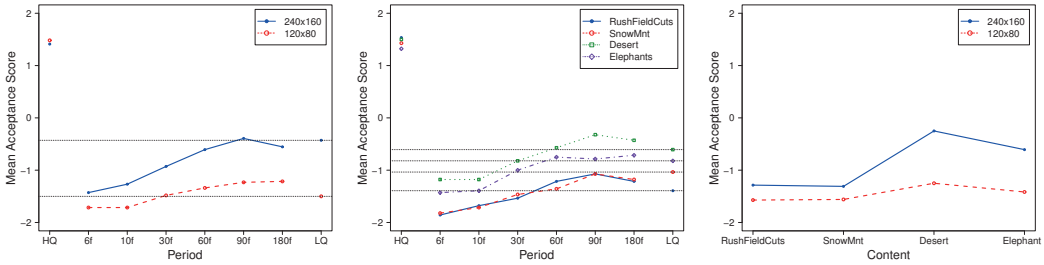


Figure 5: Effects of period, amplitude and content on Blur flicker. Error bars represent 95% confidence intervals.



(a) Period effects according to amplitude (b) Period effects according to content (c) Content effects according to amplitude

Figure 6: Explored interactions for Blur flicker. (HQ = constant high quality, LQ = constant low quality)

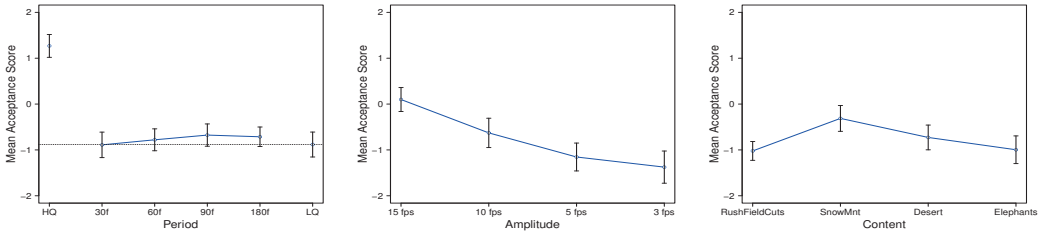
In noise and motion flicker, we find an amplitude threshold below which the flicker is considered better than the low-quality control stimuli for all content types. Figures 9 and 10 show amplitudes above and below the threshold. In our experiments, an increase of the amplitude above 8 QPs for noise flicker or 10 fps (one third of the original frame rate) for motion flicker brings significant flicker effect that may make frequent adaptation worthless to perform (see figures 9(b) and 10(b)). While it is possible to obtain a benefit by choosing a suitable period for SNR variation, only the amplitude is critical for frame rate variation.

For blur flicker, we have tested only two amplitude levels (see figure 6(a)). Although the difference between them significant, the range we have selected does not cover enough amplitudes to draw further conclusions. The user experience of watching up-scaled video that was originally half or a quarter of the native display resolution of a handheld device turned out to yield low acceptance. Given the fact that our content is chosen from a wide range of spatial and temporal complexities (figure 2), this indicates that the change of spatial resolution should not exceed half the original size in order to deliver a generally acceptable quality. Further investigations are necessary to find acceptability thresholds for amplitude levels of blur.

5.3 Content Effect

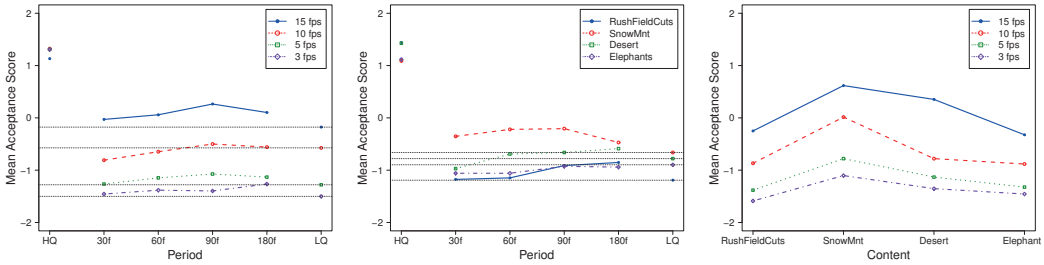
Content seems to play a minor role for flicker, but its effect varies across different flicker types. For noise flicker, the effect of content is not significant (figure 3(c)). We observe weak interaction effects between period and content (figure 4(b)), but no interaction between amplitude and content. In figure 4(c), we see that the acceptance scores vary only slightly between content for the noise flicker although the chosen amplitudes cover a large part of the scale. However, a significant effect of content can be found in both blur and motion flicker (figures 5(c) and 7(c)). Content interacts slightly with amplitude as well. For blur flicker, the Desert and Elephant sequences get significantly different scores than RushFieldCuts and SnowMnt, see figure 6(c). For motion flicker, the SnowMnt sequence is least influenced by the loss of frame rate and always has significantly higher scores, see figures 8(b), 8(c) and 10. The observation means different content characteristics can influence the perception of flicker.

The SnowMnt and RushFieldCuts sequences have more complex texture details then the other two content types and are therefore more strongly affected by the loss of spatial resolution. Additionally, SnowMnt contains significantly less motion; half of the sequence moves slowly around the snow mountain at fairly constant distance. The lack of relative movement between objects in the scene may limit the visible effect of frame dropping. However, video classification based only on two simple metrics of spatial and temporal information does not cover enough content features that are related to human perception. Region of interest, the scope and direction of motion etc. may also have influences on visual experience. In our experiments, 15fps has the effect that the scores for two test sequences are on the



(a) (b) (c)

Figure 7: Effects of period, amplitude and content on Motion flicker stimuli. Error bars represent 95% confidence intervals.



(a) Period effects according to amplitude (b) Period effects according to content (c) Content effects according to amplitude

Figure 8: Explored interactions for Motion flicker. (HQ = constant high quality, LQ = constant low quality)

negative part of the scale (see figure 10(a)), while the two sequences have quite different temporal complexity according to the TI metric, introduced in section 3. More advanced feature analysis is needed for further explanation of these phenomena.

5.4 Applicability of the Results

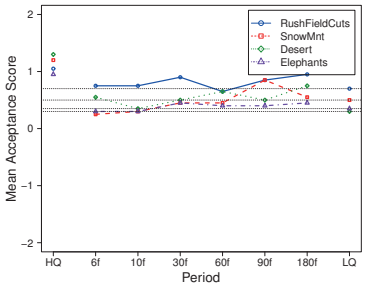
The results of our study can help improve the adaptation strategy in streaming systems or bit-rate controller for processing scalable video. Among three dimensions, SNR scalability is the most recommended adaptation option. When switching SNR layer, quality differences should be limited to less than 4 QPs to avoid additional visual artifacts. However, if larger quality shift is necessary, a quality level should be kept stable for at least 2 seconds.

The knowledge is applicable for both SVC-type and AVC-type systems – We have used SVC, but the results should be equally important/relevant for AVC-type systems like those used in modern HTTP streaming systems. For SVC, this knowledge helps to schedule the different enhancement layers and decide which to drop in case of congestion. For AVC, it helps determining how to code the different layers in order to increase quality if congestion forces the application to choose another quality layer.

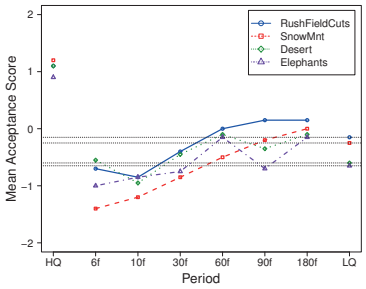
6. CONCLUSION

To understand the human perception of video quality adaptation in fluctuating bandwidth scenarios, like streaming to handheld devices over wireless networks, we have performed a series of subjective assessment experiments using iPhones and iPods. We have identified three types of visual artifacts caused by adaptive bit-rate variations, the noise, blur and motion flicker effects. Furthermore, for these flicker effects we investigated how users experience quality changes at different amplitudes and frequencies, using several content types. Our results show that multiple factors influence the quality with respect to flicker effects in different scenarios. Among the influential factors, low frequency can relieve the annoyance of flicker effect in spatial dimension, but below a threshold (on the scale of a few seconds), decreasing frequency further does not have any significant effect. On the other hand, the amplitude has a dominant effect across spatial and temporal dimensions and should be kept as low as possible for satisfactory visual quality. Finally, blur and motion flicker effect on different content types varies even for the same amplitude. Videos with complex spatial details are particularly affected by blur flicker, while videos with complex and global motion require higher frame rate for smooth playback effect.

There are still numerous questions to answer and experiments to perform which is ongoing work. We are currently expanding our experiments to HD displays to see if there are differences in the findings as compared to the performed iPhone experiments. We are also interested in other content features and their influences on user perceived quality. We will consider in particular whether content with a unique focus point (e.g. speaking person) in a scene leads to different

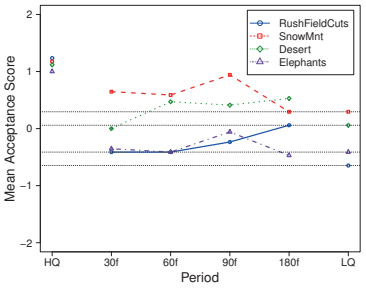


(a) Amplitude = QP28

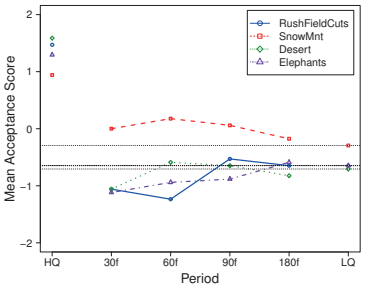


(b) Amplitude = QP32

Figure 9: Mean acceptance scores for two top amplitude levels in Noise flicker. (HQ = constant high quality, LQ = constant low quality)



(a) Amplitude = 15fps



(b) Amplitude = 10fps

Figure 10: Mean acceptance scores for two top amplitude levels in Motion flicker. (HQ = constant high quality, LQ = constant low quality)

results, whether connecting temporal and spatial complexity to regions of interest makes a difference, and how camera motion vs. content motion affects results.

7. ACKNOWLEDGMENTS

The authors would like to thank the volunteer participants. This work is sponsored by the Norwegian Research Council under the Perceval project (project number 439838), the Verdione project (project number 187828) and the iAD centre for Research-based Innovation (project number 174867).

8. REFERENCES

[1] Adobe. HTTP dynamic streaming on the Adobe Flash platform. http://www.adobe.com/products/-httpdynamicstreaming/-pdfs/httpdynamicstreaming_wp_ue.pdf, 2010.

[2] Coolican, H. *Research Methods and Statistics in Psychology*, 4 ed. Hodder Arnold, 2004.

[3] Frigge, M., Hoaglin, D. C., and Iglewicz, B. Some implementations of the boxplot. *The American Statistician 43*, 1 (Feb. 1989), 50–54.

[4] Goyal, V. K. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine 18*, 5 (September 2001), 74–93.

[5] Howell, D. C. *Statistical Methods for Psychology*, 5 ed. Duxberry, 2002.

[6] Huang, J., Krasic, C., Walpole, J., and Feng, W. Adaptive live video streaming by priority drop. In *in Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance* (2003), pp. 342–347.

[7] International Telecommunications Union. *ITU-T P.910. Subjective video quality assessment methods for multimedia applications*, 1999.

[8] ITU-T and ISO/IEC JTC 1. *Advanced Video Coding for Generic Audiovisual services, ITU-T Recommendation H.264*, Apr. 2003. ISO/IEC 14496-10(AVC).

[9] Kim, C. S., Jin, S. H., Seo, D. J., and Ro, Y. M. Measuring video quality on full scalability of H.264/AVC scalable video coding. *IEICE Trans. on Communications E91-B*, 5 (2008), 1269–1278.

[10] McCarthy, J. D., Sasse, M. A., and Miras, D. Sharp or smooth?: Comparing the effects of quantization vs. frame rate for streamed video. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004), pp. 535–542.

[11] Move Networks. Internet television: Challenges and opportunities. Tech. rep., Move Networks, Inc., November 2008.

[12] Ni, P., Eichhorn, A., Griwodz, C., and Halvorsen, P. Fine-grained scalable streaming from coarse-grained videos. In *Proceedings of the 18th International Workshop on Network and Operating*

Systems Support for Digital Audio and Video (NOSSDAV) (2009), pp. 103–108.

[13] Ni, P., Eichhorn, A., Griwodz, C., and Halvorsen, P. Frequent layer switching for perceived quality improvements of coarse-grained scalable video. *Springer Multimedia Systems Journal 16*, 3 (2010), 171–182.

[14] Pantos, R., Batson, J., Biderman, D., May, B., and Tseng, A. HTTP live streaming. http://tools.ietf.org/html/draft-pantos-http-live-streaming-04, 2010.

[15] Perneger, T. V. What's wrong with Bonferroni adjustments. *British Medical Journal 316*, 7139 (1998), 1236–1238.

[16] Pinson, M., and Wolf, S. A new standardized method for objectively measuring video quality. *IEEE Trans. on Broadcasting 50*, 3 (Sept. 2004), 312–322.

[17] Schwarz, H., Marpe, D., and Wiegand, T. Overview of the scalable video coding extension of the h.264/avc standard. *Circuits and Systems for Video Technology, IEEE Transactions on 17*, 9 (sept. 2007), 1103–1120.

[18] Sheldon, M. R., Fillyaw, M. J., and Thompson, W. D. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International 1*, 4 (1996), 221–228.

[19] Zambelli, A. Smooth streaming technical overview. http://learn.iis.net/page.aspx/626/smooth-streaming-technical-overview, 2009.

[20] Zink, M., Künzel, O., Schmitt, J., and Steinmetz, R. Subjective impression of variations in layer encoded videos. In *Proceedings of International Workshop on Quality of Service* (2003), pp. 137–154.