

Flicker Effects in Adaptive Video Streaming to Handheld Devices

Pengpeng Ni^{1,2}, Ragnhild Eg^{1,3}, Alexander Eichhorn¹,
Carsten Griwodz^{1,2}, Pål Halvorsen^{1,2}

¹Simula Research Laboratory, Norway

²Department of Informatics, University of Oslo, Norway

³Department 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'11, November 28-December 1, 2011, Scottsdale, Arizona, USA.

Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...\$10.00.

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 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 (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.

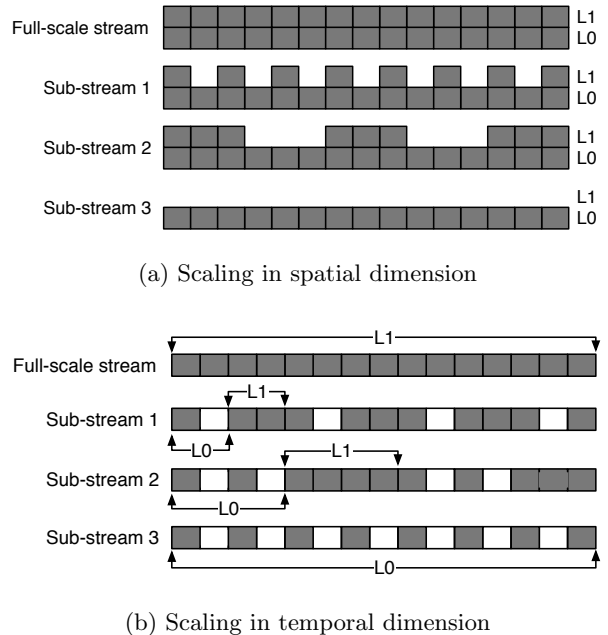


Figure 1: Illustration of streaming patterns for scalable video.

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

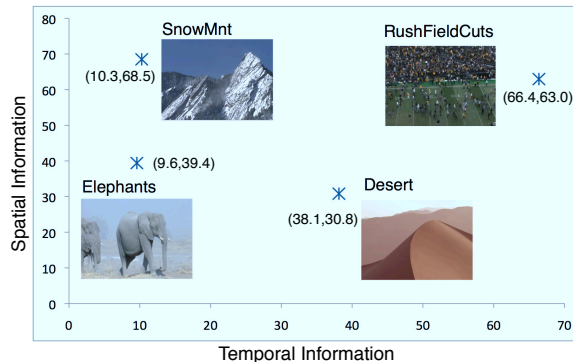


Figure 2: Test sequences.

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 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

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).

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.

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 to 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).

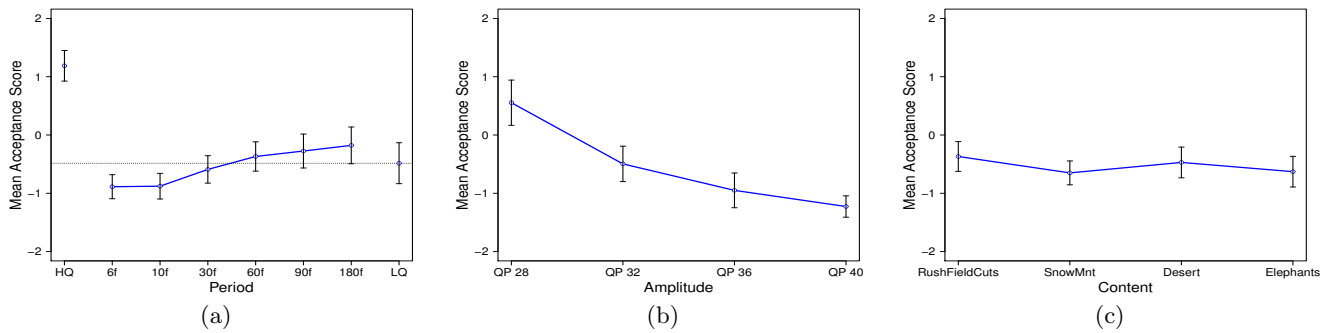


Figure 3: Effects of period, amplitude and content on Noise flicker stimuli. Error bars represent 95% confidence intervals.

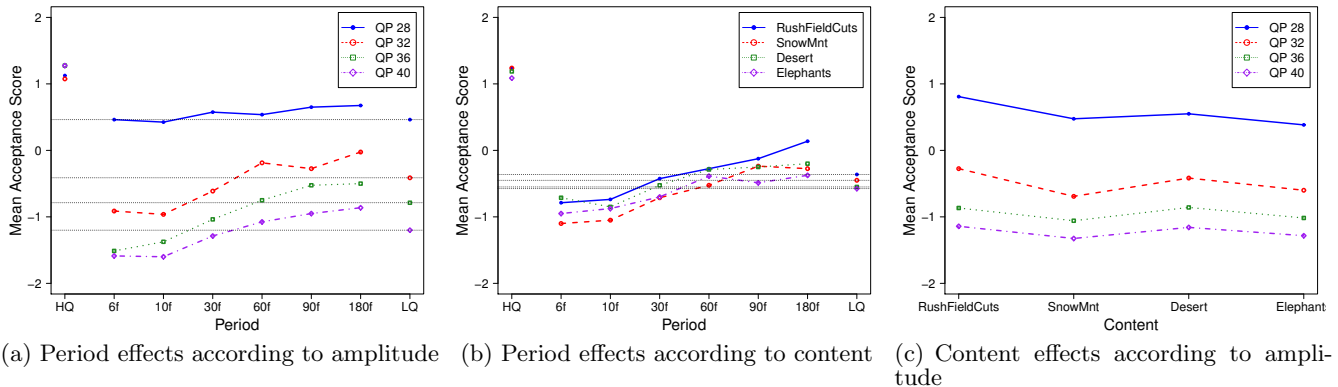


Figure 4: Explored interactions for Noise flicker. (HQ = constant high quality, LQ = constant low quality)

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 differ-

ent 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.

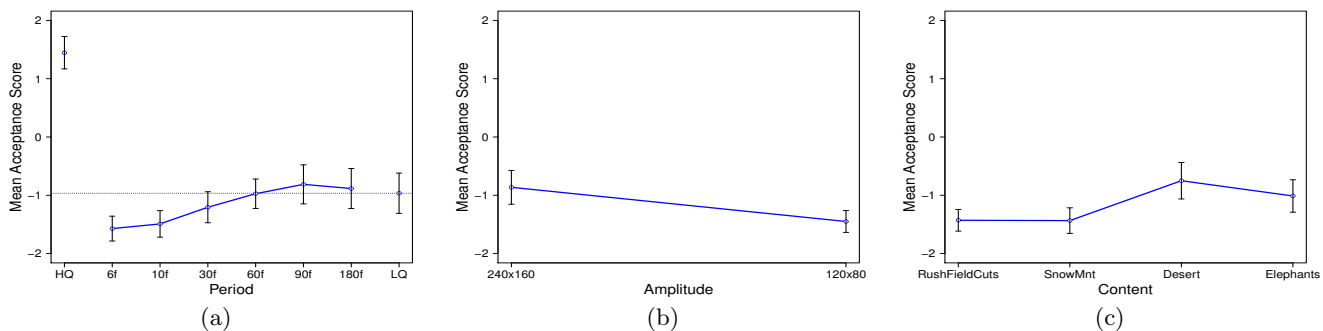


Figure 5: Effects of period, amplitude and content on Blur flicker. Error bars represent 95% confidence intervals.

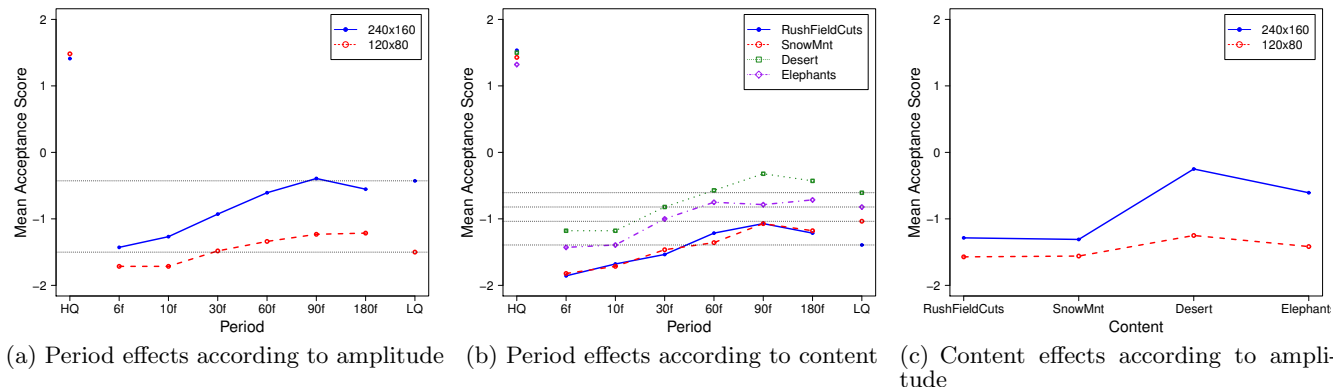


Figure 6: Explored interactions for Blur flicker. (HQ = constant high quality, LQ = constant low quality)

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 ob-

serve 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 than 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

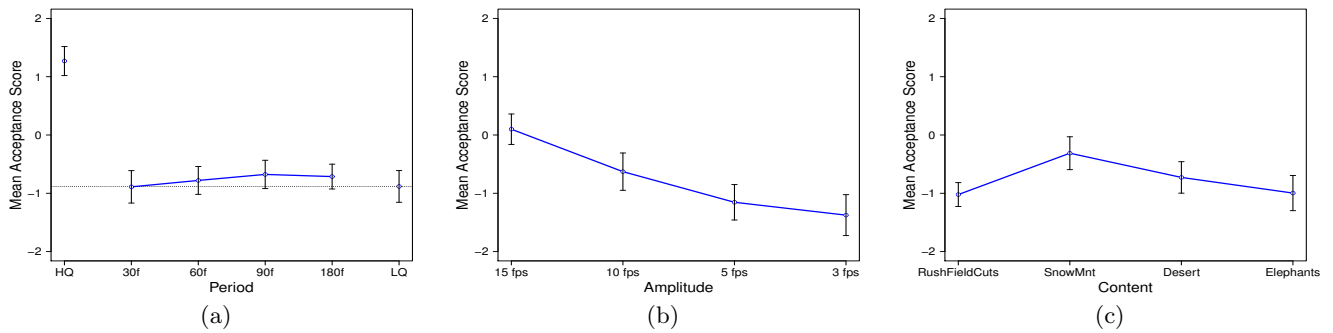


Figure 7: Effects of period, amplitude and content on Motion flicker stimuli. Error bars represent 95% confidence intervals.

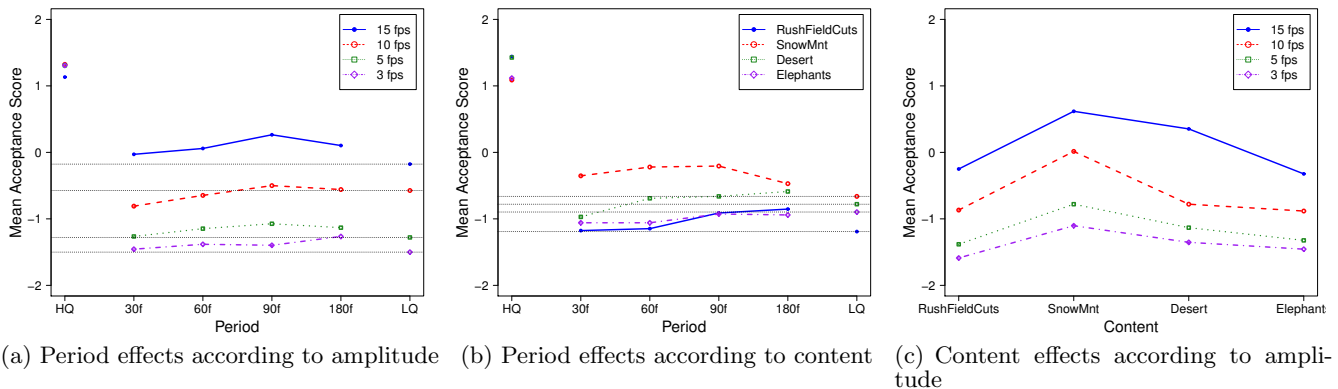


Figure 8: Explored interactions for Motion flicker. (HQ = constant high quality, LQ = constant low quality)

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.

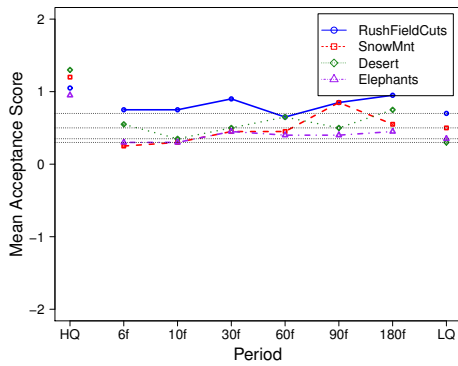
The knowledge is applicable for both SVC-type and AVC-type systems – We have used SVC, but the results should be equally important/relevant for AVC-type systems like those used in modern HTTP streaming systems. For SVC, this knowledge helps to schedule the different enhancement layers and decide which 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

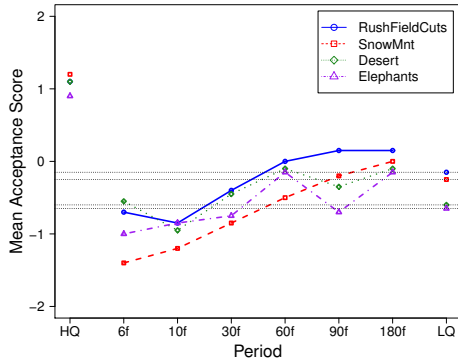
To understand the human perception of video quality adaptation in fluctuating bandwidth scenarios, like streaming to

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)

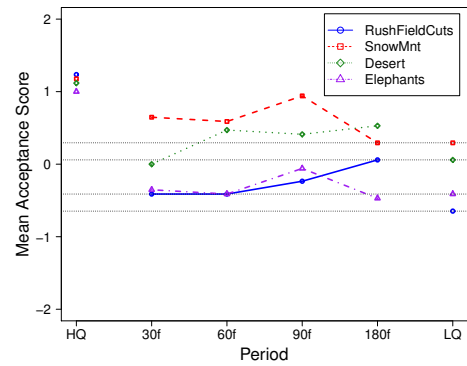
results, whether connecting temporal and spatial complexity to regions of interest makes a difference, and how camera motion vs. content motion affects results.

7. ACKNOWLEDGMENTS

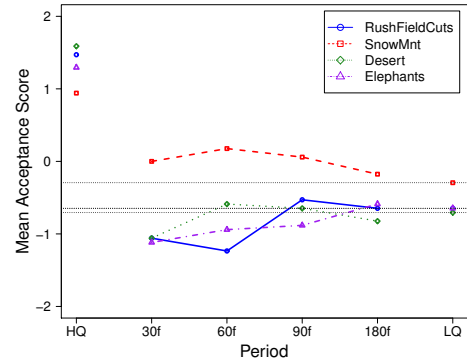
The authors would like to thank the volunteer participants. This work is sponsored by the Norwegian Research Council under the Perceval project (project number 439838), the Verdione project (project number 187828) and the iAD centre for Research-based Innovation (project number 174867).

8. REFERENCES

- [1] ADOBE. HTTP dynamic streaming on the Adobe Flash platform. http://www.adobe.com/products/httpdynamicstreaming/-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.



(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)

- [6] HUANG, J., KRASIC, C., WALPOLE, J., AND FENG, W. Adaptive live video streaming by priority drop. 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.