

SPATIAL FLICKER EFFECT IN VIDEO SCALING

Pengpeng Ni^{1,2}, Ragnhild Eg¹, Alexander Eichhorn¹, Carsten Griwodz^{1,2}, Pål Halvorsen^{1,2}

¹Simula Research Laboratory, Norway

²Department of Informatics, University of Oslo, Norway

ABSTRACT

Scalable video streaming may result in flicker effects – visual artifacts in video presentation due to adaptive layer switching. In our work, we have identified three types of flicker, noise, blur and motion flicker. Here, we investigate the blur and noise flicker, which are both related to the spatial domain. The perceptual impact of blur and noise flicker is explored through subjective assessments, focusing on acceptance of variations in the amplitudes and frequencies of the quality changes, across four content types. Our results indicate that the perception and acceptance of different layer variations are jointly influenced by multiple factors. When video scaling is required, frequency can be adjusted to relieve the annoyance of flicker artifacts, while amplitude still plays the dominant role in delivering satisfactory video quality. In addition, contents with complex details are more affected by blur flicker than other contents. This difference is not observed for noise flicker.

Index Terms— Subjective video quality, Video adaptation, Layer switching

1. INTRODUCTION

Video streaming over the Internet and wireless networks encounters varying network resource availability. To cope with bandwidth fluctuations, many video streaming systems employ scalable video coding techniques to facilitate their video transmission. Usually, these techniques structure the compressed video bit stream into layers. As a self-decodable sub-stream, each layer can easily be extracted from the full-scale video stream. Video streaming systems can then adaptively change the size of the streamed video (and thus the quality) to maintain continuous playback and avoid congestion.

However, the video layers contain different amounts of distortion and different types of visual artifacts due to coding and compression. For instance, current developments in modern scalable video coding techniques enable the extraction of layers in both spatial and temporal dimensions [1], which unavoidably results in visual artifacts such as blurriness, image noise and motion jerkiness. Adaption to available bandwidth causes frequent layer switching, which results in quality fluctuations in the streamed video. When quality is shifted frequently up and down, users can detect annoying *flicker* effects, i.e., visual artifacts in the video presentation

due to adaptive layer switching. This paper reports on a study that considers the following spatial scalability options:

- **Noise flicker** is a recurring transient change in noise, edge business, blockiness or other still-image artifacts in a video sequence. This is due to the variation of signal-noise-ratio (SNR).
- **Blur flicker** is a recurring transient blur that sharpens and unsharpens the overall details of some frames in a video sequence. This is due to repeated change of spatial resolution.

When a flicker effect is perceived, it usually harms the experienced subjective quality. Note, however, that noise, blur and motion flicker can not be considered deficient by themselves. Active adaptation to changes in available bandwidth is generally preferable to random packet loss or stalling streams, and not every layer change is perceived as a flickering effect. Furthermore, the perceptual effect of the flicker is also highly related to the *amplitude* and *frequency* of the layer changes. In figure 1, we illustrate four simple streaming example patterns. The figure 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, and 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 how fast the quality changes and repeats itself. In this example, sub-stream 1 changes quality every t frames, whereas sub-stream 2 changes every $3t$ frames.

To the best of our knowledge, very little work has been performed to evaluate the flicker effect in the video quality

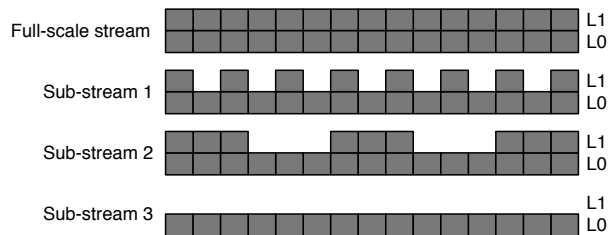


Fig. 1. Different streaming patterns for scalable video.

domain. In [2], 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. [3] 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 [4, 5]. The closest related work [6], points out that the frequency and amplitude of layer changes influence the perceived quality and should therefore be kept as small as possible. However, the current study limits itself to SNR scalability and does not take into account the influence of video content characteristics.

To provide the best possible video quality, the streaming applications need to select the most suitable streaming options from several streaming patterns. Considering the alternatives in figure 1, 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 mobile devices. We have considered state-of-the-market encoding techniques represented by the H.264 series of standards. Our goals are to (1) identify the main influential factors, and (2) find the range of the factor levels to minimize the flicker effect in layer variation. We evaluated the effect of blur and noise flicker on four different types of video content. For each video type, we tested several levels of frequency and amplitude, and in total, we performed 192 individual assessments.

From our results, we observe that the perception of layer variation is jointly influenced by multiple factors. Amplitude and frequency have significant impact on subjective impression, and we find that flicker and content influence perceived quality as well. We show that decreasing the quality switching frequency contributes to a better perceived quality, but also that there are thresholds beyond which the quality is reduced.

2. EXPERIMENT DESIGN

2.1. Randomized block design

Human perception of flicker effect is subjective and influenced by many factors. The factors are flicker type, amplitude, frequency, content characteristics (see section 2.2) and streaming scenario context. The streaming context includes display device, viewing distance, peripheral vision, etc. We investigate flicker effects by subjective experiments with the purpose of factor screening and characterization. Each influential factor is modeled as an independent variable, and the dependent variable is the acceptance degree of the flicker ar-

tifact. Among the independent variables, the amplitude and frequency are the design factors selected for study in our experiments. We rely on randomization to balance out the effect of content and treat the flicker type and display device as nuisance factors.

We apply a repeated measures design [7] in this evaluation where each subject repeatedly provides responses to all different stimuli. This design is most often applied in perceptual evaluations since it excludes the variation due to difference between the subjects from the sources of random error. However, a full factorial experiment design that covers all possible combinations of influential factors at all levels would last for hours. Participants can easily become tired, bored and uncooperative. Their responses become increasingly unreliable with test duration. Particularly, in our field study, we randomly recruited participants who may not have previous experience with quality assessment. They may not be able to complete a test lasting for more than 15 minutes. Hence, techniques are needed to reduce the minimum number of stimuli that each participant receives. In this study, we use blocking strategies to divide the stimuli into subsets, where each subset reflects a block of relatively homogeneous experimental conditions. To make the test duration generally acceptable, we treat different streaming scenarios separately and focus on streaming to hand-held mobile devices. We also test the flicker type in blocks, where each block consists of the combination of the typical visual artifacts associated with the corresponding flicker effect. Finally, we block different amplitude levels. By using the blocking strategies, we limit the test duration of a block to about 12 minutes. Each block contains all the combinations of different levels of period and content. The display order of the stimuli within each block is completely randomized, and blocks of different flicker types are mixed together. Our participants are blind to the applied amplitude factor in each block. They can relax or leave after completing the tasks of one block. When they volunteer to continue the experiment, they select blocks randomly. Hence, the test order of different amplitude levels is also randomized.

This flexible experiment design relieves effectively the stress of the test participants. It creates the restriction that cross comparison between different flicker types and between different streaming scenarios is not permissible.

2.2. Content selection and preparation

The rate distortion performance of compressed video is highly related to its spatial and temporal complexity. A simple classification of the video is to classify video content into four groups with spatial and temporal complexity ranging from high to low level. We measure the spatial and temporal information of video as recommended in [8]. Based on the two metrics and observation, we select four test sequences for our subjective tests (figure 2).

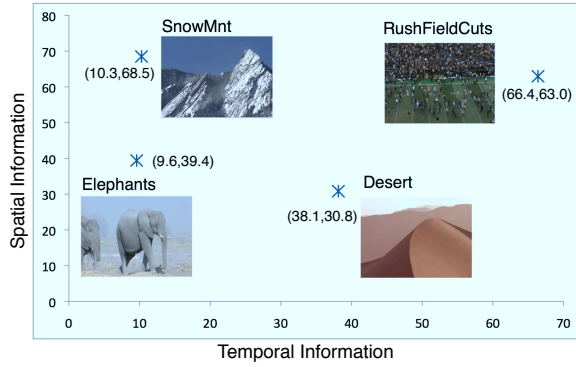


Fig. 2. Test sequences.

We use the H.264/SVC reference software JSVM 9.19¹ to encode the raw material. For each type of flicker, we generate streams with only two layers as shown in figure 1. The amount of impairment differs between the two layers, which decides the amplitude level of the layer variation. In table 1, we show our selection of factor levels for the independent variables introduced in section 2.1. In noise flicker, constant quantization parameters (QP) were used to encode a base layer L0 and an enhancement layer L1. We encode L1 of all test sequences with QP24. Quality variation is due to the use of different QPs in L0. Hence, we use the QPs in L0 to represent the amplitude levels in noise flicker. In blur flicker, L0 is downscaled to a lower spatial resolution. We use the lower resolution to represent the amplitude levels in blur flicker.

To simulate the flicker effect, 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 a periodic duration to represent the frequency factor. We use the periods of 6, 10, 30, 60, 90 and 180 frames for both noise and blur flicker.

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	

Table 1. Selection of factor levels

2.3. Participants

The 28 participants (9 female, 19 male) were recruited at the University of Oslo and volunteered by responding to posters on campus. They were rewarded with a monetary compen-

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

sation. Ages ranged from 19 to 41 years (mean 24), and all participants reported normal or corrected to normal vision.

2.4. Procedure

The experiment was carried out as a field study in a University library. Videos were presented on iPod touch devices with 3.5-inch displays and resolutions of 480x320, frame rate of 30 fps. The brightness levels were consistently set to 50%. Participants sat in lounge chairs, away from any sunlight, and were asked to hold the device at a comfortable viewing distance. They chose one of the experiment blocks to commence the experiment. In accordance with the ITU-T Absolute Category Rating method [8], the participants were presented 12-seconds videos as single-stimulus events. Each stimuli was presented once. To respond, participants had to evaluate the stimulus according to the statement “I accept the overall quality of the video” and indicate their agreement on a balanced 5-point Likert scale, by tapping the appropriate option on-screen.

3. DATA ANALYSIS

3.1. Method of Analysis

The current study evaluates the impact of two types of flicker stimuli, blur and noise, on user acceptance, where stimuli vary in period, amplitude and content. To establish baselines for rating behavior, we include clips with video quality displayed constantly at either the highest or the lowest level. Rating responses from the different blocks are combined and grouped according to flicker stimulus type, with control stimuli included in both groups. Responses are transformed to scores by assigning them with values ranging from -2 to 2, so that the score “2” represents the highest acceptance, “0” the neutral element and “-2” the lowest acceptance. Separate analyses are carried out for blur and noise flicker stimuli.

The coherence of results is controlled by each participant’s acceptance scores for the high and low quality control videos. Individual outliers are identified by checking for large numbers of conflicting responses - rating scores that are higher for low quality control videos compared to high quality ones. A tolerable number of conflicting responses lies within 1.5 times the interquartile range. Exceeding this, a participant’s responses are regarded as unreliable and thus excluded from further analysis. As a result, 2 participants are excluded from the analyses pertaining to blur flicker stimuli. None are excluded for noise flicker stimuli.

Due to the ordinal nature of the data, we deem it inappropriate to apply parametric statistics that assume data to be continuous and normally distributed. Results are instead analyzed using Friedman’s chi-square test [9]. This is a non-parametric test for repeated measures designs that assesses the difference between means for multiple variables, across individuals. To explore main effects, multiple Friedman’s chi-square tests are applied to data sets collapsed across factors.

Simple effects are also investigated using Friedman’s tests. For this, the relation between two factors will be assessed by averaging across levels of the third. Although multiple comparisons require adjustments to significance levels, such as the Bonferroni correction, these tend to increase the occurrence of Type II errors, that is, increase the chances of rejecting a true difference [10]. Therefore, we have avoided these and instead report significant results without corrections. Furthermore, we refrain from drawing definite conclusions, leaving our results open to interpretation.

3.2. Noise flicker effects

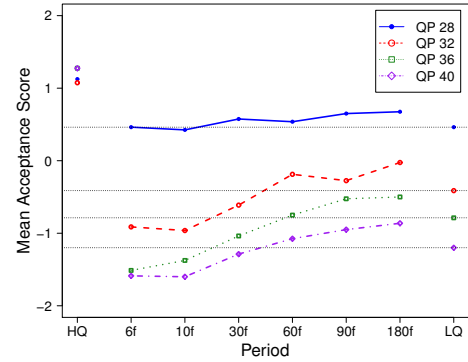
For the noise flicker stimuli, Friedman’s Chi-Square reveals main effects for period ($\chi^2(5) = 69.25, p < .001$), amplitude ($\chi^2(3) = 47.98, p < .001$) and content ($\chi^2(3) = 27.75, p < .001$). Simple effects are found between two or more levels of period at amplitudes QP 32, QP 36 and QP 40, and between two or more levels of amplitude for all content types. As for content, simple effects are only evident at amplitudes QP 28 and QP 32. As seen in figure 3, mean acceptance scores vary across content types and tend to decrease with higher amplitudes and increase with higher period levels. Moreover, the scores hint at potential interactions, particularly between period and amplitude.

3.3. Blur flicker effects

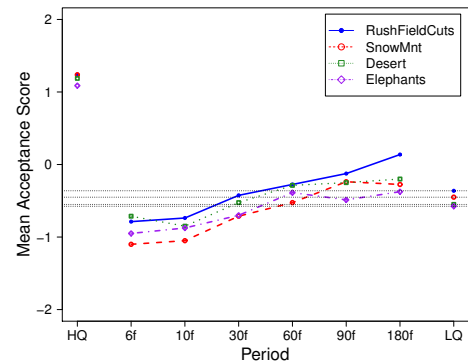
Friedman’s Chi-Square unveils main effects for all variables also for blur flicker stimuli, including period ($\chi^2(6) = 41.79, p < .001$), amplitude ($\chi^2(1) = 14.00, p < .001$) and content ($\chi^2(3) = 33.80, p < .001$). Here, simple effects are found between two or more levels of period, and between two or more content types, for both amplitude levels. Simple effects are also uncovered between the two amplitude levels for the desert and elephants contents, with higher mean acceptance scores for amplitude 240x160 compared to 120x80. The variations in mean acceptance scores across period levels, amplitudes and content types are illustrated in figure 4. Again, the tendency is for acceptance to increase with higher levels of period and to vary across content types. However, the scores seem to collect in two content groups, with Rush-Field and SnowMountain in one, and Desert and Elephants in the other.

4. DISCUSSION

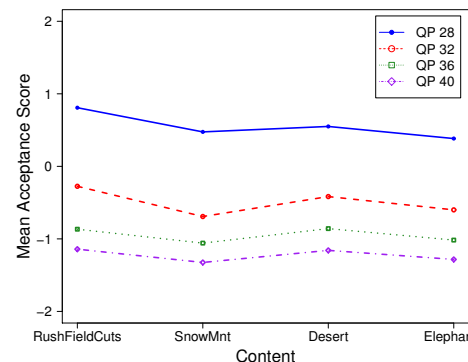
The five-point Likert scale which we used to collect responses has a neutral element in the center and two opposite extreme values at both ends. In figures 3 and 4, the mean value of all responses to a particular test item is displayed. A positive value can be interpreted as an acceptable quality level, a neutral score means undecidedness, while a negative score indicates an unacceptable quality level. The scale is certainly



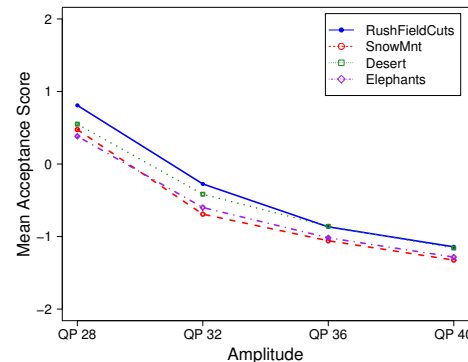
(a) Effects of period according to levels of amplitude



(b) Effects of period according to content types

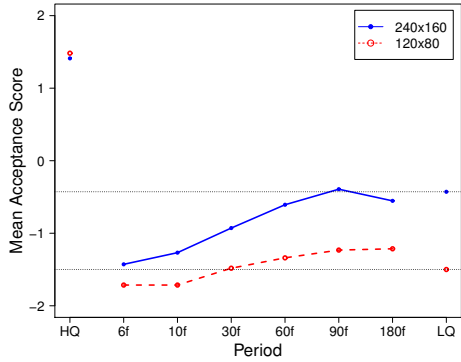


(c) Effects of content type according to levels of amplitude

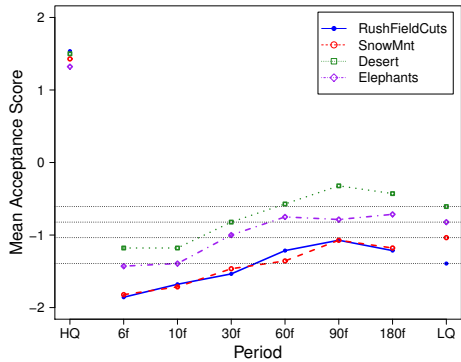


(d) Effects of amplitude according to content types

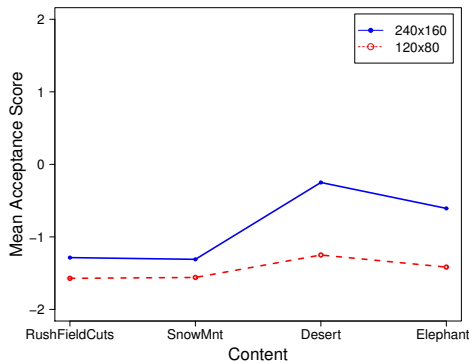
Fig. 3. Mean acceptance scores for Noise flicker. (HQ = constant high quality, LQ = constant low quality)



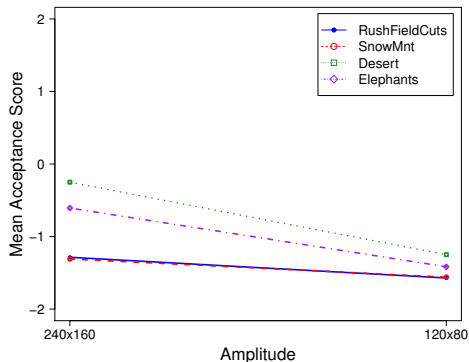
(a) Effects of period according to levels of amplitude



(b) Effects of period according to content types



(c) Effects of content type according to levels of amplitude



(d) Effects of amplitude according to content types

Fig. 4. Mean acceptance scores for Blur flicker. (HQ = constant high quality, LQ = constant low quality)

not linear, so our conclusions consider rank relations rather than absolute differences.

4.1. Period Effect

The period or frequency of the flicker is a major influential factor. We can actually identify three frequency ranges that influence the overall quality acceptance differently (see figures 3(a), 3(b), 4(a) and 4(b)). In figure 3(b), the mean values over all amplitudes are shown, making the pure period effect visible. 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 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. This is significant across amplitudes (figure 3(a)), content (figure 3(b)) and flicker types (figure 4(b)), but counter examples exist (see the top line in figure 4(a)).

The conclusion is that an adaptive video streaming can outperform constant low quality streams, but the frequency of switching must be considered in relation to the magnitude of flicker amplitudes.

4.2. Amplitude Effect

The amplitude is the most dominant factor influencing the perception of flicker. This seems reasonable since, with increasing amplitude, the visual artifacts become more apparent when alternating between two quality versions. Our statistical results, presented in section 3 and displayed in figures 3(a) and 3(d), indicate that there is no interaction between amplitude and content for noise flicker (figure 3(d)), but a weak interaction exists for blur flicker (figure 4(d)).

At low flicker amplitudes, where visual artifacts are least obvious (see Q28 in figure 3(a)), the flicker effect is almost not detectable. An increase in flicker amplitude above a threshold of 8 QPs in our experiment makes artifacts clearly detectable so the effect of flicker at different period lengths become significant.

For blur flicker, we have tested two amplitude levels only (see figure 4(a)). Although the effect is significant here, the range we have selected does not cover the entire acceptance score spectrum, which makes it hard to draw further conclusions. The user acceptance of watching up-scaled video that was originally half or a quarter of the native display resolution of a mobile device has turned out to be below neutral scale values. Further investigations to find acceptability thresholds for amplitude levels of blur are necessary.

4.3. Content Effect

Content seems to play a minor role for flicker. For noise flicker, we observe weak interaction effects between period

and content (figure 3(b)), but no interaction between amplitude and content (figure 3(d)). However, for blur flicker, slightly stronger interactions between amplitude and content can be found, i.e., in figure 4(c) and figure 4(d), the Desert and Elephant sequences get significant different scores compared to RushFieldCuts and SnowMnt. The observation means different content characteristics can influence the perception of flicker.

Comparing the ranges of the curves in figures 3(c) to figure 4(c), we see that for noise flicker the acceptance scores vary only slightly between content although the chosen amplitudes cover a large part of the scale. For blur flicker, where we unfortunately have only two amplitudes (at half and at a quarter of the native display resolution), the scores are in the negative part of the scale only. 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 be larger than half the original size in order to deliver a general acceptable quality.

5. CONCLUSION

To understand the human perception of video quality adaption in fluctuating bandwidth scenarios like streaming to mobile devices over wireless networks, we have performed a series of subjective assessment experiments using iPhones/iPods. We looked at blur and noise flicker effects when the user experiences 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, but below a threshold (on the scale of few seconds), decreasing further frequency does not bring significant effect. On the other hand, the amplitude has dominant effect and should be kept as low as possible for satisfactory visual quality. Finally, blur flicker effect on different content types varies even for the same amplitude. Videos with complex spatial details are particularly affected.

We have in this study focused on noise and blur flicker, but there are still numerous questions to answer and experiments to perform which is ongoing work. In this respect, we are currently expanding our experiments to also include motion flicker. We are also performing experiments using HD displays to see if there are differences in the conclusions compared to the performed iPhone experiments. Finally, we plan to add other types of videos and other switching patterns, amplitudes and frequencies.

6. 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).

7. 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] M.H. Pinson and S. Wolf, "A New Standardized Method for objectively Measuring Video Quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.
- [3] Cheon Seog Kim, Sung Ho Jin, Dong Jun Seo, and Yong Man 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.
- [4] Pengpeng Ni, Alexander Eichhorn, Carsten Griwodz, and Pål Halvorsen, "Frequent layer switching for perceived quality improvements of coarse-grained scalable video," *Springer Multimedia Systems Journal*, vol. 16, no. 3, pp. 171–182, 2010.
- [5] John D. McCarthy, M. Angela Sasse, and Dimitrios Miras, "Sharp or Smooth?: Comparing the Effects of Quantization vs. Frame Rate for Streamed Video," in *Proc. CHI*, 2004, pp. 535–542.
- [6] Michael Zink, Oliver Künzel, Jens Schmitt, and Ralf Steinmetz, "Subjective Impression of Variations in Layer Encoded Videos," in *Proc. IWQoS*, 2003, pp. 137–154.
- [7] Hugh Coolican, *Research Methods and Statistics in Psychology*, Hodder Arnold, 4 edition, 2004.
- [8] International Telecommunications Union, *ITU-T P.910. Subjective video quality assessment methods for multimedia applications*, 1999.
- [9] Michael R. Sheldon, Michael J. Fillyaw, and W. Douglas Thompson, "The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs," *Physiotherapy Research International*, vol. 1, no. 4, pp. 221–228, 1996.
- [10] Thomas V. Perneger, "What's wrong with Bonferroni adjustments," *British Medical Journal*, vol. 316, no. 7139, pp. 1236–1238, 1998.