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

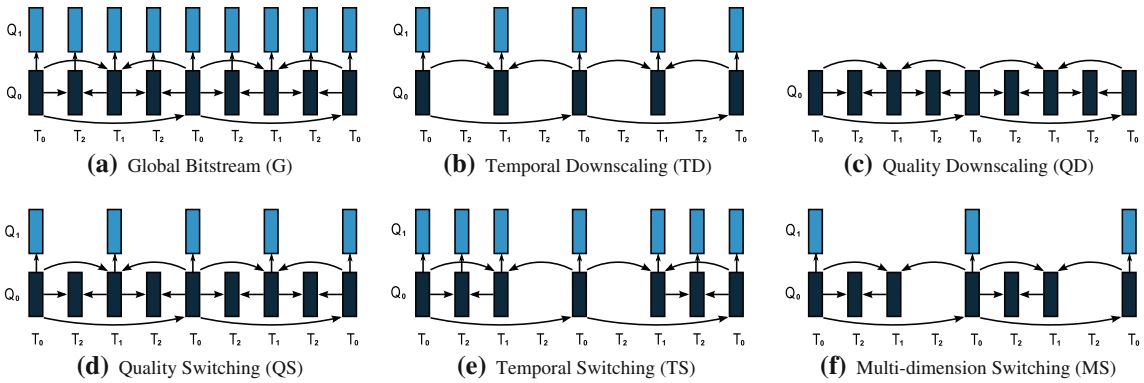

Fig. 1 Bitstream layout for downscaling and layer switching options used in the experiments. Q and T denote the quality and temporal dimensions, respectively

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.

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.

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

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)