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

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

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

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

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

I. Introduction

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

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

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

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

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

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

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

II. Related work

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

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

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

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

III. Field Study Design

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

A. Content Selection and Encoding

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

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

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

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

B. Scalable Operation Points

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

C. Subjective Assessment Procedures

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

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

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

Table II
SELECTED OPERATION POINTS.

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

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

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

D. Limitations

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

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

IV. BITSTREAM ANALYSIS

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

A. Scaling Granularity and Diversity

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

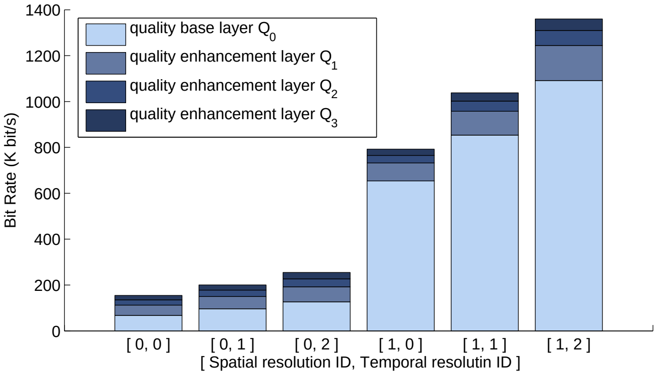

Figure 1. Bitrate allocation for scalable OPs $[S_m, T_n]$ in Sports sequence, where S_m represents m-th spatial resolution, T_n represents n-th temporal resolution. Each bar column can be additionally truncated into 4 quality layers identified by Q_i.

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

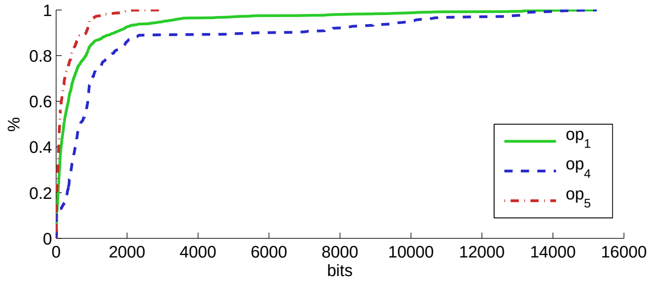

Figure 2. Cumulative distribution function (CDF) of NALU packet sizes for selected operation points of the Sports sequence.

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

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

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

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

B. Packet Statistics

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

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

V. SVC QUALITY ASSESSMENT

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

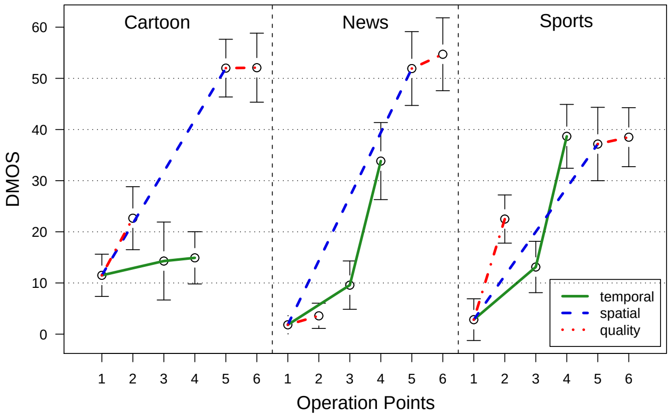

Figure 3. Subjective DMOS scores for selected sequences as means with 95% confidence intervals. QoE gradients for within-dimension scaling are shown as lines. Note that higher DMOS scores mean lower QoE and that the bitrate drops from 1.5 Mbit for OP1 to 128 Kbit for OP6.

A. Noticeable QoE Degradations

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

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

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

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

Table IV
Scaling order preferences between dimensions.
Legend: T - temporal, S - spatial, Q - quality (SNR) dimension, --- large negative effect, -- medium negative effect, - small negative effect, ○ not significant, + small positive effect, ++ medium positive effect, +++ large positive effect. Preferred scaling orders: 1 (Q – T_{12} – S – T_6), 2 (T_{12} – T_6 – Q – S), 3 (Q – T_{12} – T_6 – S), 4 (T_{12} – Q – T_6 – S).

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

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

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

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

B. Scaling Order Preferences

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

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

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

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

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

C. Objective Model Performance

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

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

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

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

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

VI. CONCLUSIONS

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

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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