4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

Representation Switch Smoothing for Adaptive HTTP Streaming

Michael Grafl[1], *Christian Timmerer*[1]

[1]Institute of Information Technology (ITEC), Alpen-Adria-Universität (AAU), Klagenfurt, Austria
michael.grafl@itec.aau.at, christian.timmerer@itec.aau.at

Abstract

When an adaptive media streaming system has to switch from one representation of the content to another, the switch causes viewer distraction. We introduce the concept of representation switch smoothing for alleviating the distraction and improving the overall quality of experience. As adaptive HTTP streaming systems typically deploy video buffers on the client side, the adaptation decision is known far enough ahead of playout time to perform a seamless transition between quality representations. We discuss implementation considerations for an adaptive HTTP streaming system with scalable video coding, present a subjective evaluation of the proposed approach, and identify factors that influence how smooth transitions are perceived.

Index Terms: adaptive streaming, representation switching, quality of experience

1. Introduction and Concept

Adaptive HTTP streaming has gained widespread adoption in delivering multimedia content within heterogeneous environments (i.e., networks, terminals, users). It is typically deployed on top of the existing (network) infrastructure without any quality guarantees and, thus, delivered on a best-effort basis. Various proprietary formats are available which may eventually converge to MPEG's standard on Dynamic Adaptive Streaming over HTTP (DASH) [1]. However, the way in which services, based on these formats, are perceived by the end user is deliberately not determined by the format itself but subject to research. In general, the multimedia content is provided in multiple versions – referred to as *representations* comprising different bitrates, resolutions, codecs, languages, etc. – and available in segments of about 2-10 seconds length. The client requests segments based on its context (e.g., current throughput which translates to available bandwidth) and may switch to different representations, typically at segment boundaries.

Frequent quality switches with high amplitudes in adaptive HTTP streaming sessions – e.g., switching from (very) high to (very) low bitrates – have been shown to annoy viewers and, thus, reduce the Quality of Experience (QoE) [2]. The disturbance can be reduced through intermediate quality levels [3] but in practice only very few levels (3-5) are deployed. Previous work focused only on quality switches at segment boundaries and viewers may still notice abrupt quality changes.

In this paper, we propose a more fine-grained approach, a smooth transition between representations, which we subsequently call *representation switch smoothing*. The goal of *representation switch smoothing* is to reduce the annoyance of quality switches even further. When the receiver is aware of an imminent switch to a lower representation, it can already reduce the playout quality of the current representation, enabling a smooth transition between the two representations. Frame by frame, the playout quality is slightly reduced. Vice versa, the playout quality is smoothly increased after a higher representation has been received. This concept is illustrated in Figure 1.

In client-driven streaming scenarios such as DASH, the adaptation decision is typically known at least one segment duration ahead of the playout time. While the current segment is played, the next segment has to be requested to ensure timely arrival. For the deployment of Scalable Video Coding (SVC) [4] in DASH, the time frame might be shorter, depending on whether enhancement layers of the segment are downloaded using HTTP pipelining [5][6]. Typical DASH clients already decide to adapt to a lower representation when still three or more 2-second segments are buffered [6]. If the adaptation logic pursues a conservative buffer management (e.g., [7]), the adaptation decision is taken even further ahead. In any case, the receiver is aware of pending representation switches ahead of playout time and can thus react by smoothing the quality transition.

Representation switch smoothing can be realized by an additional component in the decoding chain. This component is notified by the client's adaptation logic whenever the adaptation decision is changed. The amplitude of the switch has to be signaled as well. For SVC with medium-grain scalability (MGS) layers, this can be represented as the difference in MGS layers. In a more general system, the bitrates or the video qualities (e.g., PSNR) of the higher and lower representation may be signaled. If the first frame of the lower representation can already be decoded, its quality could be used by the representation switch smoothing component as reference to adjust the amount of noise it adds to frames of the higher representation. Depending on the amplitude of the representation switch, the smoothing component chooses the duration of the transition; higher amplitudes require longer durations.

In case of down-switching, the component adds increasing noise to the frames of the higher representation as detailed in Section 3 until it matches the quality of the lower representation just before the switching. In case of up-switching, the component adds noise with temporally decreasing intensity to the frames of the higher representation, such that the transition between representations becomes seamless.

The remainder of this paper is structured as follows. Related work is discussed in Section 2. In Section 3, implementation options for the representation switch smoothing component are explained. In Section 4, we conduct a subjective evaluation on whether representation switch smoothing has a positive impact on the QoE. Section 5 identifies several factors that influence how quality switches are perceived and discusses the potential impact of these factors on representation switch smoothing. The

10.21437/PQS.2013-32

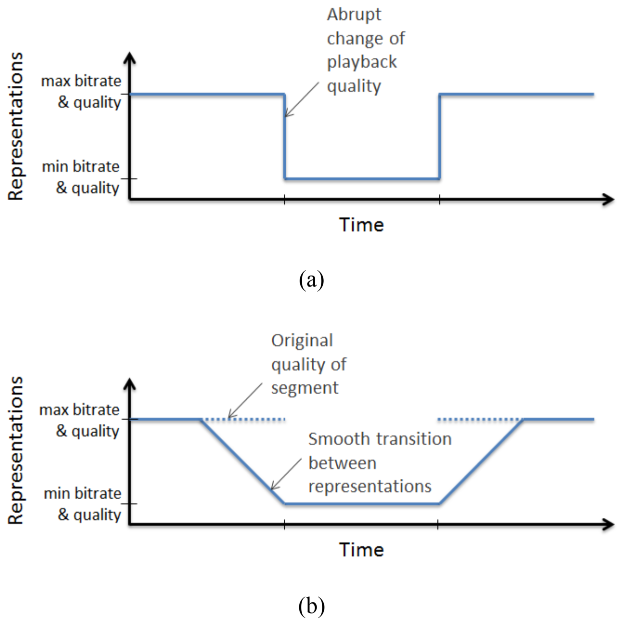


Figure 1: Adaptation with (a) traditional representation switching and (b) representation switch smoothing.

paper is concluded in Section 6 together with an outlook on future work.

2. Related Work

In adaptive HTTP streaming an important factor for QoE is flickering due to switches between representations. Ni et al. [2] have evaluated the impact of flickering on the video acceptance by the viewer on mobile devices. They have investigated the effects of changing video qualities (noise flicker), video resolutions (blur flicker), and frame rates (motion flickering) for SVC at various configurations with periodic flickering durations. Periodic flickering means that a switch from the higher to the lower representation, and vice versa, occurred periodically, e.g., every 2 seconds. Their results show that frequent noise flickering between two SNR representations with a period below 2 seconds impairs the viewing quality down to a point where viewers would prefer the lower video representation altogether. For blur flickering, viewers preferred the constant lower representation (at half the original resolution) for flickering periods up to 2 seconds.

Mok et al. [3] have proposed a QoE-aware DASH system based on AVC. As quality switches of high amplitude (e.g., from highest to lowest representation) are annoying to viewers, the proposed adaptation algorithm inserts intermediate steps to avoid abrupt quality changes. Thus, the reduced amplitude of quality switches seems to outweigh the additional number of quality switches in terms of QoE. This also confirms an earlier study on quality switches by Zink et al. [8] that has evaluated viewers' preferences of various quality switching patterns. General trends in those patterns are that high amplitudes in down-switches should be avoided and that switching up is preferred to switching down (i.e., it is better to start with a low quality and switch up than to start with a high quality and switch down).

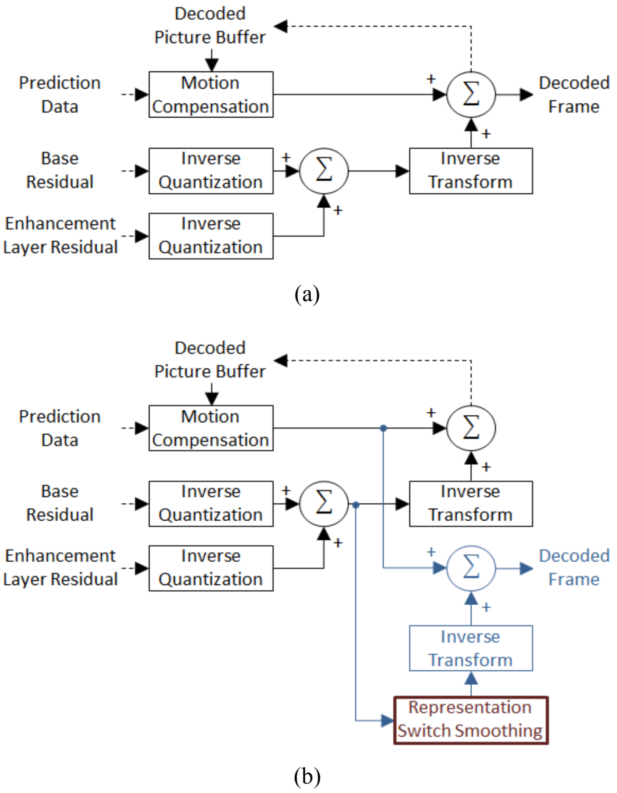


Figure 2: Simplified block diagram of the SVC decoding process for (a) traditional decoding, adopted from [9] and (b) decoding with representation switch smoothing.

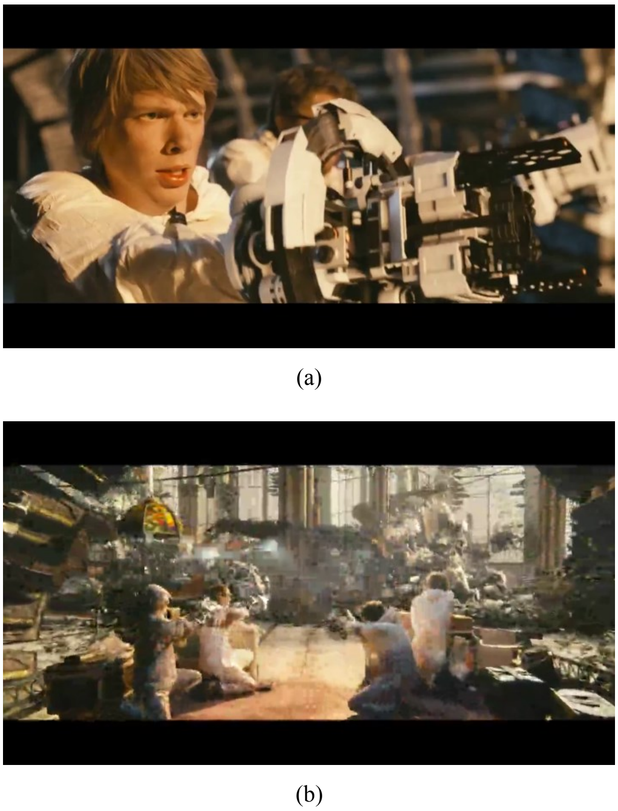
In a recent study, Sieber et al. [7] have proposed an SVC adaptation logic that reduces the number of quality switches by striving for a stable buffer level before increasing the number of consumed SVC layers. Their evaluations show a very high and stable overall playback quality of the proposed algorithm compared to other state-of-the-art SVC-DASH adaptation techniques. However, the comparison does not take the amplitude of quality switches into account.

3. Implementation Options

There are three options for implementing the smooth reduction of quality: either before, within, or after the decoder. As smoothing for down-switching is performed analog to up-switching, we only consider the former in our discussion.

The first option, denoted *pre-decoder* implementation, is to add a filter component before the decoder. This component alters the encoded bitstream by removing certain picture fidelity data. For SVC with MGS enhancement layers, a straight-forward implementation is to remove transform coefficients (i.e., set them to 0) from the enhancement layer. For the i^{th} frame in the smooth transition, n transform coefficients are removed as calculated in Equation (1).

$$\mathrm{n} = \left\lfloor \frac{i \cdot \Delta C}{s} \right\rfloor \qquad (1)$$


(a)

(b)

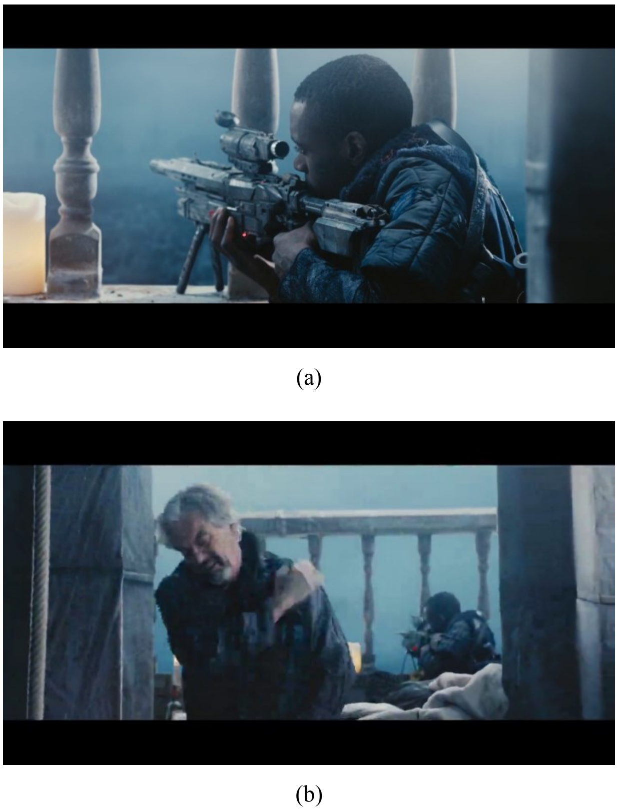
Figure 3: Snapshots of *Sequence 1* at (a) 2,000 kbps and (b) 400 kbps.

Let d be the duration of the smooth transition and ΔC be the total difference of transform coefficients between the higher and the lower representation.

This approach is easy to implement and independent of the decoder. However, a drawback is that changes from one frame are propagated within the group of pictures (GOP) due to motion compensation drift [4], causing unwanted artifacts.

The second option is an implementation inside the decoder referred to as *in-decoder* implementation. Again, some picture fidelity data is removed from the coded frames, but without affecting the motion compensation of other frames. For SVC, this implies that inverse transform of the residual data has to be performed twice. The number of transform coefficients to be removed per frame is the same as in the first implementation option. A simplified block diagram of the decoding process is given in Figure 2.

Figure 2 (a) shows the original SVC decoder structure adopted from [9] with handling of base layer and enhancement layer residual data. Figure 2 (b) highlights the additional steps necessary for maintaining the original decoded picture buffer when performing representation switch smoothing. In contrast to the first implementation option the representation switch smoothing is performed after the inverse quantization. The operations are commutative; setting a transform coefficient to 0 has the same result before and after inverse quantization.


(a)

(b)

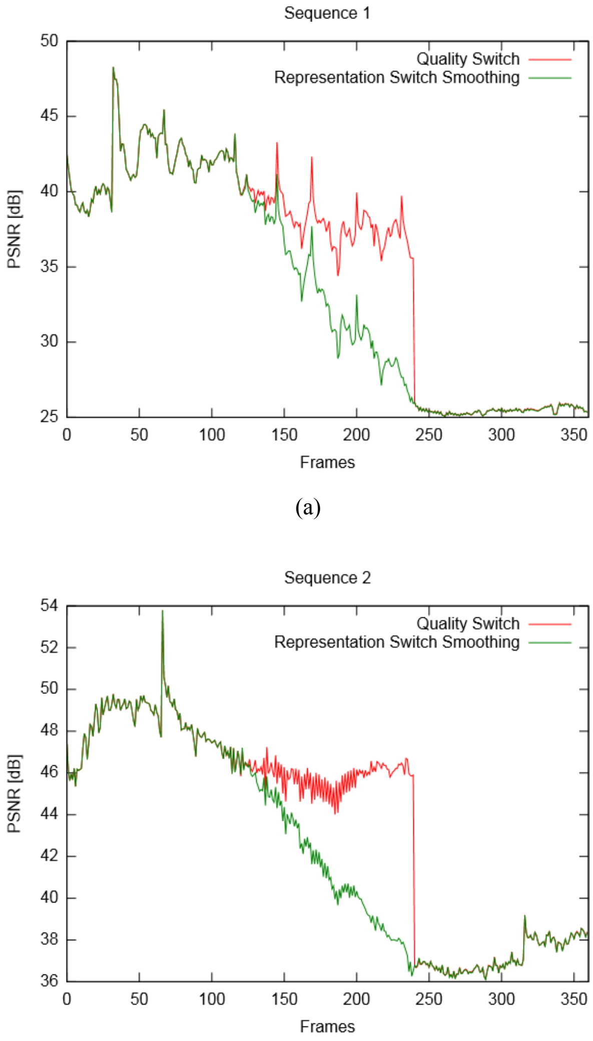
Figure 4: Snapshots of *Sequence 2* at (a) 2,000 kbps and (b) 250 kbps.

Since motion compensation is still based on the original, unimpaired coded video data, we expect the reconstructed frame to slightly differ from the case where the respective transform coefficients had been set to 0 in the encoding process. The assessment of the resulting video quality is subject to future work. Nevertheless, an implementation within the decoder is more accurate and robust than the *pre-decoder* implementation option as it avoids error propagation. Of course, it requires a specialized decoder.

Note that the first two implementation options will have to consider that SVC allows for custom scaling matrices, which even may change between frames. The scaling matrix provides the values by which the transform coefficients of a macroblock are inversely quantized. Full support for custom scaling matrices might increase the computational complexity of the implementation.

The third implementation option is to add a video filter component after the decoder for inserting additional noise into the decoded frames. We denote this as *post-decoder* implementation. This noise mimics the degrading quality to enable a smooth transition to the lower representation. The computational complexity is still slightly higher than for the first two implementation options.

This third implementation option is independent of the decoder and the video coding format and also avoids drift.



(a)

(b)

Figure 5: Per-frame PSNR results for quality switching and representation switch smoothing for (a) *Sequence 1* and (b) *Sequence 2*.

While the third implementation option is video coding format independent, it has to know the extent to which the quality changes with the representation switch, and, subsequently, how the new quality can be approximated by the synthetic distortion. Such a general model for video quality approximation remains an open research challenge.

4. Evaluation

We have performed an initial evaluation of the *representation switch smoothing* approach for down-switching scenarios through subjective tests. As up-switching might be perceived differently from down-switching, the combination of both up- and down-switching in a single test sequence could bias the results. For example, viewers might experience a sudden increase in video quality as a positive event. Thus, we only considered down-switching in our first evaluation in order to decide whether the approach is worth pursuing.

We used two test sequences, both extracted from the open-content short film *Tears of Steel* [10]. Both sequences have durations of 15 seconds at resolution 1280x720 and 24 fps frame rate. *Sequence 1* has high-motion content and was extracted starting at time point 7:43; *Sequence 2*, with low-motion content, was extracted starting at 1:57. The sequences were selected such as to avoid confusing scene changes, although both contain cuts.

The 15-second sequences were split into 5-second segments. We simulated a quality switch from a high bitrate (2,000 kbps) to a low bitrate (400 kbps for *Sequence 1* and 250 kbps for *Sequence 2*) after 10 seconds. As *Sequence 1* has higher temporal information, it was harder to compress for the encoder, causing already strong visible artifacts at 400 kbps. Snapshots of the high and low bitrate encodings are shown in Figure 3 for *Sequence 1* and in Figure 4 for *Sequence 2*.

Each sequence was encoded once with a quality switch (after 10 seconds), and once with a smooth downward transition (between seconds 5 and 10). For the purpose of this test, the sequences were encoded to AVC at constant target bitrates with the FFmpeg encoder.

We observed that the encoder badly allocates bitrates for the first few frames, especially at low target bitrates. In per-segment encoding, this caused unwanted distortion at segment boundaries. We thus decided to always encode the entire sequences and to split them into segments after encoding. In the absence of a working implementation of any of the aforementioned options, the smooth transition was realized by encoding the sequences at predetermined target bitrates (one per frame in the transition segment) and stitching the respective frames to a continuous segment. Thus, 120 encodings were used to obtain the 5-second transition.

The bitrates for the smooth transition were determined as follows. The sequence was first encoded at 5 sample bitrates (from 2,000 kbps to the lowest bitrate). The PSNR for the transition segment was calculated to obtain the rate-distortion performance. As the RD performance typically follows a logarithmic curve, a logarithmic curve fitting $f(br)$ was computed as shown in Equation (2) in order to approximate the video quality $q \approx f(br)$ for bitrate br and model parameters a and b.

$$f(br) = a \cdot \ln(br) + b \qquad (2)$$

The inverse function $f^{-1}(q)$ of this curve fitting is shown in Equation (3).

$$f^{-1}(q) = e^{\left(\frac{-b}{a}\right)} \cdot e^{\left(\frac{q}{a}\right)} \qquad (3)$$

Based on this inverse function, the 120 bitrates were calculated that predicted a linear decrease of PSNR over the entire transition duration. The per-frame PSNR results for both versions are shown in Figure 5 for the two test sequences.

One drawback of the applied solution is that the encoder uses different blocks for motion (and intra-) prediction at each bitrate. With low bitrates, blocking artifacts become increasingly visible. Due to the different predictions, the positions of the blocking artifacts change randomly for the extracted frame of each

Table 1: Subjective test results for the evaluation of representation switch smoothing.

Preferred Version Sequence	Quality Switching	Representation Switch Smoothing	No Difference
Sequence 1	5	**7**	6
Sequence 2	3	**12**	3

respective bitrate. When stitching the frames from these encodings, this causes some temporal noise. This noise is particularly visible in low-motion areas of the picture. In contrast, the low-bitrate segment at the end of a sequence has blocking artifacts that continuously move through the scene. So, even though the blocking artifacts are clearly visible, their movements correlate with the actual motions in the scene. Due to the temporal noise in the transition, the actual visual quality might be lower than what is reported by PSNR. As this effect was only recognized after time-consuming encoding of the transition segments, and due to the lack of a more accurate short-term solution, the subjective tests were performed with the described transition segments. This means that representation switch smoothing based on one of the implementation options discussed in Section 3 may even provide better results than our evaluation.

The subjective tests were performed with 18 participants (13 male, 5 female) of age 23 to 45 adopting pair-wise comparison. The participants were told that the test concerned changes in video quality. No further indication as to the nature of the quality changes was given. The participants were presented with the two versions of each sequence (labeled *Version a* and *Version b*). One version contained the quality switch, the other the smooth transition. The attribution to *Version a* and *Version b* was changed between the two sequences (i.e., representation switch smoothing was shown in *Version b* of *Sequence 1* and in Version a of *Sequence 2*). The participants were instructed that they may start with either version and may watch each version as often as they wanted. The videos were shown in full-screen mode on a Dell 1907FPc LCD monitor having a native display resolution of 1280x1024. The videos were shown without audio. The participants were asked to rate whether they preferred *Version a*, *Version b*, or saw no difference.

The results of the subjective tests are provided in Table 1. We performed the Kruskal Wallis test [11] for both sequences to test for significance of our results. The Kruskal Wallis test is the non-parametric counterpart of the one-way analysis of variance. For Sequence 1, the ρ-value is 0.8479 ($H = 0.33$), which means that the null hypothesis (i.e., viewers voting equally often for each of the three samples, thus being generally indifferent towards the transition technique) cannot be rejected. For Sequence 2, the ρ-value is 0.012 ($H = 8.84$), which means that the null hypothesis has to be rejected for $\alpha = 0.05$.

Representation switch smoothing performed significantly better for *Sequence 2* than for *Sequence 1*. Several participants reported that the high motion of *Sequence 1* made the two versions look indifferent. Many participants viewed each version at least two or three times before making a decision. There were no significant differences in the test results between male and female participants, although male participants tended to prefer representation switch smoothing slightly more than female participants. While the overall results show only a slight preference towards representation switch smoothing, we argue that further tests should be conducted, investigating the effects of smooth transitions on various configurations. Note also that the aforementioned temporal noise in the smooth transitions may have affected the test results.

5. Discussion

For future subjective tests, the following evaluations should be performed. Main influence factors to test are the amplitude of the quality switch (e.g., measured as the bitrate difference), the duration of the smooth transition, as well as the amount of spatial and temporal information. Based on our experiences and feedback from test participants, we assume representation switch smoothing to achieve the highest gain for scenes with high spatial and low temporal information. Furthermore, we speculate that longer transition durations (e.g., 10 seconds) will better mask the quality changes.

Other possible influence factors that we identified in our evaluations are the base quality (in contrast to just the bitrate differences), the presence of cuts, the resolution, and the duration for which only low quality segments are available (e.g., only 2 seconds of low quality might not justify two 10-second transitions).

Furthermore, a comparison to the intermediate switching qualities of the *QoE-aware DASH* system [3] should be considered in future evaluations. A hybrid solution between QoE-aware DASH and representation switch smoothing might lead to a simpler implementation. That is, instead of a continuous decrease of video quality, several small discrete steps could be used, all being below the just-noticeable difference (JND) [12]. However, this would require an on-the-fly estimation of JND in order to set the number and amplitude of switches accordingly.

It has to be investigated whether smooth transitions are also useful for up-switching scenarios. As evaluated by Zink et al. [8], viewers prefer to watch low-quality segments followed by high-quality segments rather than the other way around. Thus, we infer that up-switching is perceived to be less annoying than down-switching. Furthermore, Seshadrinathan and Bovik [13] have reported that viewers give poor quality ratings to sharp video quality drops but do not increase ratings as eagerly when the video quality resumes to its previous high state. From those results, we reason that up-switching is noticed less than down-switching. These two effects may diminish the benefits of a smooth transition for up-switching.

For test content generation, the aforementioned temporal noise should be avoided by implementing one of the suggested implementation options from Section 3. Instead of allowing participants to watch versions as often as they like, the test material could contain around 3-5 quality switches and be shown only once to create the same conditions for all participants. Additionally, a 5-point Likert scale could be used to better distinguish preferences between the tested versions.

6. Conclusions

In this paper, we have introduced the concept of representation switch smoothing. The approach avoids abrupt quality switches by smoothly reducing the video quality on a per-frame basis. This avoids unnecessary viewer distraction in adaptive HTTP streaming. We have discussed three implementation options for the smoothing component in an SVC-based DASH system.

While down-switching is generally considered annoying, abrupt up-switching might even increase the QoE as viewers might be happy to notice visual improvements in the video quality. It has to be evaluated whether representation switch smoothing is beneficial for up-switching at all.

Our initial evaluations indicate a tendency towards the benefit of representation switch smoothing compared to hard quality switches. So far, we have only evaluated down-switching scenarios with very few configurations. Based on these evaluations, we have identified parameters and test methods for future subjective tests on the impact of representation switch smoothing on the QoE. Future work shall derive a model from these evaluations for configuring the duration of a quality transition against the amplitude of the representation switch.

7. Acknowledgements

This work was supported in part by the EC in the context of the ALICANTE project (FP7-ICT-248652).

8. References

[1] I. Sodagar, “MPEG-DASH: The Standard for Multimedia Streaming Over Internet”, *IEEE Multimedia*, vol.18, no.4, pp.62-67, October-December 2011.

[2] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, “Flicker effects in adaptive video streaming to handheld devices,” in Proc. *19th ACM International Conference on Multimedia*, New York, NY, USA, pp. 463–472, 2011.

[3] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, “QDASH: a QoE-aware DASH system,” in Proc. *3rd Multimedia Systems Conference*, New York, NY, 2012.

[4] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the Scalable Video Coding Extension of the H. 264/AVC Standard,” *IEEE Trans. on CSVT*, vol. 17, no. 9, 2007.

[5] C. Müller, “libdash supports now persistent connections and pipelining,” blog entry, URL: “http://dash.itec.aau.at/?p=553”, March 1, 2012. Accessed July 7, 2013.

[6] C. Müller, D. Renzi, S. Lederer, S. Battista, and C. Timmerer, “Using Scalable Video Coding for Dynamic Adaptive Streaming over HTTP in Mobile Environments,” in *Proc. 20th European Signal Processing Conference* (EUSIPCO), Bucharest, Romania, August 2012.

[7] C. Sieber, T. Hoßfeld, T. Zinner, P. Tran-Gia, C. Timmerer, “Implementation and User-centric Comparison of a Novel Adaptation Logic for DASH with SVC,” in *Proc. IFIP/IEEE International Workshop on Quality of Experience Centric Management* (QCMan), Ghent, Belgium, May 2013.

[8] M. Zink, O. Künzel, J. Schmitt, and R. Steinmetz, "Subjective Impression of Variations in Layer Encoded Videos,” in *Quality of Service — IWQoS 2003*, K. Jeffay, I. Stoica, and K. Wehrle, Eds. Springer Berlin Heidelberg, pp. 137–154, 2003.

[9] A. Segall and Jie Zhao, “Bit stream rewriting for SVC-to-AVC conversion,” in Proc. *15th IEEE International Conference on Image Processing*, San Diego, CA, pp. 2776–2779, 2008.

[10] Tears of Steel, “Tears of Steel | Mango Open Movie Project,” Home Page, URL: "http://mango.blender.org/", accessed July 7, 2013.

[11] R. Lowry, “Concepts and Applications of Inferential Statistics,” R. Lowry, 1998-2013. Available online: "http://vassarstats.net/textbook/", accessed June 21, 2013.

[12] Y. Jia, W. Lin, and A. A. Kassim, “Estimating just-noticeable distortion for video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 7, pp. 820–829, 2006.

[13] K. Seshadrinathan and A. C. Bovik, “Temporal hysteresis model of time varying subjective video quality,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP '11), Prague, Czech Republic, pp. 1153–1156, May 2011.