SUBJECTIVE QUALITY OF SVC-CODED VIDEOS WITH DIFFERENT ERROR-PATTERNS CONCEALED USING SPATIAL SCALABILITY

Yohann Pitrey, Ulrich Engelke, Marcus Barkowsky, Romuald Pépion, Patrick Le Callet

LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597
(Institut de Recherche en Communications et Cybernétique de Nantes), Polytech NANTES, FRANCE
{first-name.last-name}@univ-nantes.fr

ABSTRACT

Degradation of network performance during video transmission may lead to disturbing visual artifacts. Some packets might be lost, corrupted or delayed, making it impossible to properly decode the video data on time at the receiver. The quality of the error-concealment technique, as well as the spatial and temporal position of the artifacts have a large impact on the perceived quality after decoding. In this paper, we use the spatial scalability feature of Scalable Video Coding (SVC) for error-concealment. This enables the transmission of a lower resolution video with a higher robustness, for example using unequal error protection. Under the assumption that only the higher resolution video would be affected, we evaluated the visual impact of packet losses in a large scale subjective video quality experiment using the Absolute Category Rating method. The number of impairments, the duration, and the interval between impairments as well the quality of the encoded lower resolution video are varied in a systematic evaluation. This allows for analyzing the influence of each factor both independently and jointly.

Index Terms— Subjective quality assessment, video transmission, error-concealment, scalable video coding, packet-loss

1. INTRODUCTION

Current communication networks are subject to impairments, either due to traffic, network topology or to hardware malfunction. Wireless networks additionally have to face signal reception problems, which is one more factor of impairment. These impairments can lead to packet loss, delay or corruption at the receiver. Nowadays, all competitive video compression techniques such as MPEG-4 AVC/H.264 [1] use predictive lossy coding. As a result, packet errors (assimilated to packet loss in the following) can lead to error propagation and have a strong impact on the visual quality.

Error-concealment techniques are commonly deployed to reduce the impact of packet loss on the quality of the video [2]. Several methods have been used, relying on the remaining data itself or adding redundancy inside the data Frame skipping and freezing (also denoted as frame copy) are two of the most commonly used techniques to perform error-concealment from the stream itself. They are mostly suited for sequences with low motion patterns and low bitrate applications. Using these techniques, the impact of a loss might be dramatic if it occurs during a scene change or in a scene containing high motion patterns. In such a situation, the temporal prediction relies on information that is severely different from the lost data and the consequent spatial and temporal distortions have a strong impact on the visual quality. Nonetheless, frame skipping and freezing do not require any special encoder configuration or extra data to conceal the lost areas. Therefore any encoded video sequence might be concealed using these techniques.

Another type of error-concealment techniques uses part of the payload to add redundancy into the bit-stream and use it as a backup against data loss. Usually this kind of techniques perform better than frame skipping and freezing because they are able to reproduce some part of the lost data and thus to reduce error propagation artifacts. Some approaches are based on conventional video coding techniques such as MPEG-2 or MPEG-4 AVC/H.264 [3], while others use MPEG-4 Scalable Video Coding (SVC) [4] and its ability to encapsulate several versions of a single coded video in one stream [5, 6]. A version of the video with lower resolution and/or quality called the base layer is incorporated inside the video stream, and used whenever the decoder is unable to decode the full-size version. This reduced size video layer requires less data than the full-size version and can easily be protected against transmission errors. This way, it is possible to protect the video stream against transmission errors with a reduced cost in terms of bit-rate.

Apart from the network behavior, lossy video coding has an impact on the visual quality, such as demonstrated in [7, 8]. Hence in a classical video transmission process, the encoding parameters together with the network behavior and the error-concealment technique all cause distortions that can affect the opinion of quality of human viewers. Several contributions evaluate the impact of the three aforementioned elements in different contexts. In [9], the impact of MPEG-4

AVC/H.264 coding, combined with frame copy error concealment on subjective quality is evaluated for different transmission error-patterns. The error length and the number of impairments are taken into consideration on heterogeneous encoded video contents. The results show that user opinion is linked to the length of an impairment if the drop in quality caused by frame freezing exceeds a certain threshold, showing that the impact of one factor can not be easily isolated from the others. In [10], the authors evaluate the joint impact of coding artifacts and transmission errors in MPEG-4 AVC/H.264 with macro-block copy error-concealment (similar to frame copy, at the macro-block level). Using a singular subjective quality assessment methodology, they show that users have a preference for coding artifacts over packet loss artifacts. The proposed explanation is that coding artifacts introduce a relatively consistent loss of quality, whereas transmission errors produce bursts of artifacts and therefore temporal distortions which are more visible. In [11], the authors use frame freeze as an error-concealment technique and study the impact of loss ditribution and of the duration of an impairment is presented. A model is proposed using a sort of logistic function for the impact of the duration of an impairment. The influence of what the authors call the impairment density is also evaluated. They affirm that several bursts are more annoying to the viewers than a single long burst.

So far, no work has been published evaluating the impact of network behavior and encoder configuration on the visual quality using SVC-based error concealment. Although, it has been demonstrated that this type of error concealment outperforms usual techniques such as frame freeze and should be considered as an interesting candidate for practical video transmission applications [12]. In this paper, we present the results of a large scale subjective test evaluating the impact of network behavior and coding artifacts using SVC-based error-concealment. Several packet loss patterns are studied in order to evaluate the impact of three network-based distortions: 1) the length of one impairment, 2) the number of impairments and 3) the interval between impairments. In order to evaluate the impact of the error-concealment technique as a fourth factor of distortions, we use two different settings for the SVC streams: a high-quality base layer scenario and a low-quality base layer scenario. For each of these four factors, we use several representative configurations in order to draw a systematic analysis of the influence of each factor, both independently and jointly.

The remaining of this paper is organized as follows. Section 2 describes the error-concealment method we use after packet-loss. Section 3 presents the design of our subjective experiment. The results are analyzed in section 4. Section 5 concludes the paper.

2. SCALABLE VIDEO CODING-BASED ERROR-CONCEALMENT

Scalable Video Coding allows for several video layers with increasing resolution and/or quality to be embedded in a single stream [4]. This feature has already been used to perform error-concealment after packet-loss distortion [5, 6]. A base layer with lower resolution and/or quality is transmitted together with the full-size video sequence. The base layer can be used as a backup when packet loss makes it impossible to decode the full-size video. In this section we describe the encoding configuration for the scalable streams, as well as the error-concealment process.

In our experiment, we use scalable streams encoded with the MPEG-4 SVC Reference Encoder Software version 9.18 [13]. The streams contain two layers: a base layer in QVGA format (320x240 pixels) at 30 frames per second (fps), and a spatial enhancement layer in VGA format (640x480 pixels) also at 30 fps. The enhancement layer is encoded with a constant QP equal to 32, which corresponds to a good trade-off between quality and bit-rate. The base layer is encoded using a constant QP equal to either 38 or 44, leading to two different scenarios. The scenario in which the base layer is encoded with a QP of 38 is considered as a high-quality base layer scenario. The coding artifacts are limited, but the bit-rate needed to transport the base layer together with the enhancement layer is relatively high. The scenario in which the base layer is encoded with a QP of 44 is comparably considered as a low-quality base layer scenario. The coding artifacts are more visible in the base layer, but the bit-rate needed to transport it is lower. In the following, the two scenarios will be referred to according only to the QP used for the base layer, as the enhancement layer is encoded with an equal QP of 32 in both scenarios. In both scenarios the base layer and the enhancement layer are encoded with one IDR frame every 64 frames, one I frame every 32 frames and one P frame every 16 frames. The size of a Group-of-Pictures is therefore equal to 16. Inter-layer prediction is enabled between the base layer and the enhancement layer, in order to reduce the bit-rate of the whole encoding. We make the assumption that an error-protection technique is applied on the base layer, so that it is possible to decode it in any case.

When packet loss makes the decoding of the enhancement layer impossible, the corresponding frames from the base layer are decoded and displayed instead. As the base layer does not have the same size as the higher layer, switching from one layer to the other might create abrupt quality changes, which are particularly noticeable from a user point of view. In order to reduce these quality variations, some post-processing is performed before displaying the frames from the base layer. This post-processing step is referred to as *upscaling*. Generally speaking, two kinds of upscaling are to be distinguished, namely spatial and temporal upscaling. Spatial upscaling is required when the frames from the base layer

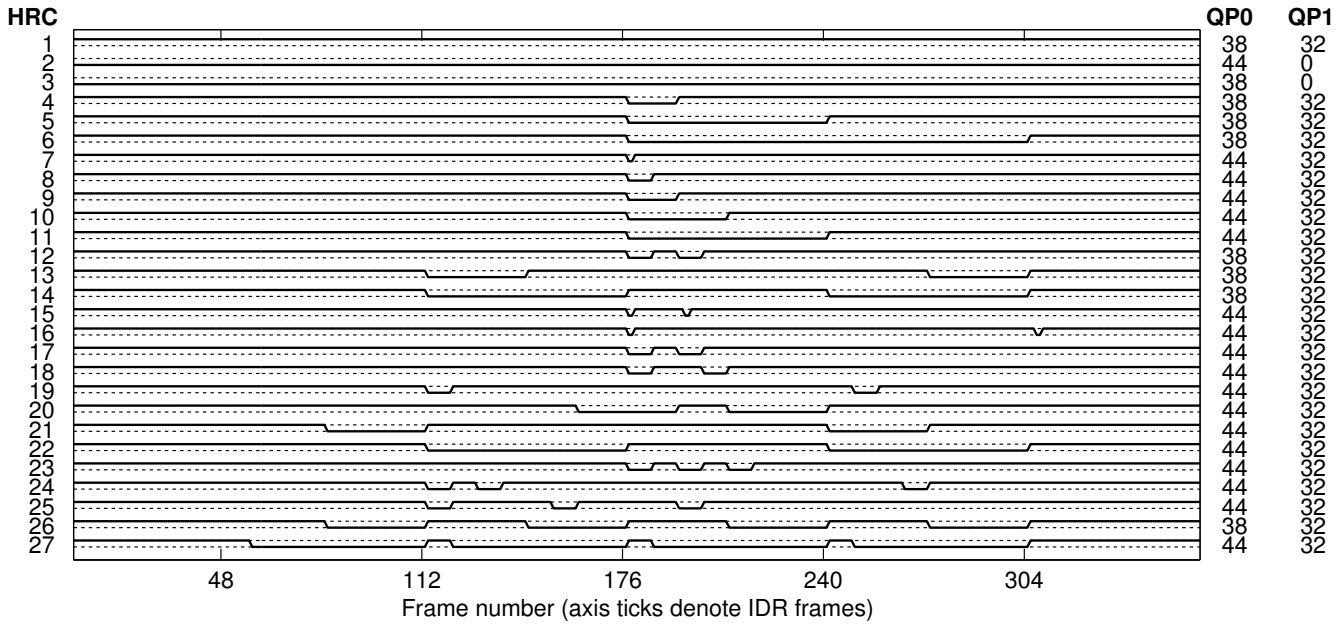

Fig. 1. Impairment patterns with associated Mean Opinion Scores and QP values. For each HRC, the two dotted lines denote the base and enhancement scalable layers. The solid line represents the layer from which the current frame is displayed (base layer means impairment). We vary the number of impairments, the total length of the impairments, the interval between impairments and the quality of the base layer (QP values appear in the two columns on the right.

have lower resolution than the frames from the higher layer. Several techniques exist, from direct pixel replication to more complex upscalers. The spatial upscaling technique has an impact on the quality of the upscaled base layer and thus it is important to choose it wisely. In [12], the impact on the subjective quality of two well known techniques, namely Bilinear interpolation and Lanczos upscalers have been studied in the context of High-Definition video coding using H.264. The authors demonstrated that these two techniques were well accepted by the viewers and they could lead to comparable quality between a full-HD 1080p stream encoded at 6 Mbit/s and an upscaled HD-ready 720p stream encoded at 3 Mbit/s. According to these results, we use the Lanczos upscaler from the JSVM Reference Software Suite [13] in order to convert the frames from the base layer from QVGA to VGA. Temporal upscaling is needed when the base layer has a lower temporal frequency (number of frames per second) than the enhancement layer. While the problem of temporal upscaling can be quite complex if the two frequencies are not simply related to each other, it is broadly simplified in SVC as the temporal frequency usually increases by two from one layer to the higher layer. Therefore, temporal upscaling in the context of SVC can be performed by duplicating each frame in order to obtain a video stream with twice as many frames as the layer to upscale.

Both spatial and temporal upscaling introduce distortions in the upscaled video. Therefore, the choice of the spatial and temporal resolution of the base layer is an important step for the design of our experiment. In [12], two SVC encoding configurations in terms of visual quality in an error-concealment context were compared. In both configurations, the higher SVC layer was in VGA format at 30 fps. In the first scenario the base layer was in QVGA format at 30 fps, while in the second scenario the base layer was in QVGA format at 15 fps. In both scenarios the base layer was encoded with the same constant bit-rate using a rate control tool. Thus each frame in the second scenario receives twice as much bit-rate as one frame in the first scenario. The error-concealment method was similar to the one we are using in this paper. In order to get a VGA stream at 30 Hz, the first scenario only required spatial upscaling, while the second scenario required both spatial and temporal upscaling. The subjective quality ratings obtained after displaying some videos encoded according to these two scenarios showed that human viewers tend to prefer the second scenario, in which the base layer had the same temporal frequency as the higher layer. This preference probably holds in the fact that temporal discontinuities are more visible than spatial discontinuities, such as earlier stated in [10]. Thus, switching from the higher layer to the base layer is more annoying for the viewers if the temporal frequency changes at the time of the impairment. As a result, we choose to use the same number of frames per second in the two layers in our experiment.

The next section describes the experimental design of

our subjective test, presenting the error patterns, the encoded video contents as well as the viewer population involved and the subjective quality assessment protocol we used.

3. EXPERIMENTAL DESIGN

In order to evaluate the influence of network behavior on the visual quality, we simulate packet loss in the scalable video streams by dropping intervals of frames with various time distributions. Using frames in VGA format, packet loss bursts usually affect an area that is large enough in the frame to consider it entirely lost, such as presented in [14]. As a result, we consider the whole frame is lost if a packet loss affects its decoding.

Figure 1 depicts all the error patterns we included in our experiment. We include the QP values for each layer (QP0 for the base layer, QP1 for the enhancement layer). Each error pattern is combined with values of QP0 and QP1 in order to create a Hypothetical Reference Circuit (HRC). Intuitively, one HRC simulates the whole process of encoding an original video sequence, transmitting it on a lossy network with a specific error pattern and eventually performing error concealment after decoding. For each HRC, we display two time lines (dotted-lines): the upper line represents the enhancement layer, while the lower line represents the base layer. The plain line represents the actual layer displayed at a given time. For example, HRC 1 represents the best-case scenario, when no impairment occurs and only the frames from the enhancement layer are displayed. In the following, HRC 1 will be referred to as the high-quality reference. Symmetrically, HRC 2 and 3 represent the worst-case scenarios, when only frames from the base layer can be displayed. In HRC 2 the base layer is encoded with a QP of 38, whereas in HRC 3 a QP of 44 is used. These two HRC are used as low-quality references. Apart from the three reference HRC, the range of an impairment varies between 2 and 128 frames. The number of impairments varies from 1 to 4. The interval between impairments varies between 8 and 128 frames.

We would like to stress that our experiment is designed in such a way that the influence of each factor can be precisely analyzed both independently and jointly to the others. This will be demonstrated in the experimental results section, with a systematic analysis of the visual impact of the four aforementioned factors. Moreover, the location of the errors are calculated so that switching from the base layer to the enhancement layer occurs when a IDR or a I frame is received for the enhancement layer. This makes our error simulation more realistic as in practical application only these frames can be decoded independently and be used as "reset" frames to start displaying the enhancement layer again.

For each HRC, we encoded 11 different video contents, each lasting 12 seconds. These video were taken from the VQEG Multimedia test plan [7] and represent a good variety of spatial and temporal activities. The input VGA and QVGA videos are in YUV 4:2:0 format and do not contain any coding artifact before encoding. Each video source is processed by each presented HRC to generate a Processed Video Sequence (PVS). The subjective experiment setup follows the recommendation in ITU-R BT.500 [15]. The test environment has correct illumination, display calibration and viewing distance. The display device is a 40 inch TV-Logic LMV401 reference LCD screen with a native resolution of 1920x1080 pixels at 60 frames per second. As the test videos are in VGA format, a uniform gray border is displayed around the frames corresponding to 25% of the display maximum brightness. The subjective quality assessment protocol is the 5-level Absolute Category Rating (ACR) with Hidden Reference, described in the ITU-T P.910 recommendation [16]. During a test session, each PVS is displayed once to the viewer. The display order is randomized before the beginning of the session. The viewer is asked to give his opinion about the quality of the video using a 5-level discrete scale ranging from 1 (Bad) to 5 (Excellent). A total of 28 viewers were presented the test, with ages comprised between 18 and 57. A total of 390 video sequences were presented to each viewer during four sessions of 30 minutes, in order to limit visual fatigue.

A detailed analysis of the impact of the different error patterns is presented in the next section.

4. EXPERIMENTAL RESULTS

To conduct the data analysis on the results of our subjective test, we compare the Mean Opinion Score (MOS) of different HRC groups in Figure 2(a) to (d). The 95% intervals of confidence are displayed as vertical error bars, for statistical verification of the results. They reflect the variability of the opinion score among observers. Practically, the difference in quality score between two configurations can be considered significant if their intervals of confidence do not overlap. In order to make the results easy to interpret, we repeat the impairment pattern on each MOS resctangle. An impairment pattern appears as a string of numbers, listing the number of frames displayed alternatively from he base layer and the enhancement layer. For instance, HRC 16 is associated with the string "2-128-2" in Figure 2(d), which means a first impairment of 2 frames, followed by 128 frames from the enhancement layer and a second impairment of 2 frames.

The three reference HRC are included for comparison in all figures. It can be observed that the high-quality reference HRC obtains the highest quality score, equal to 4.3. On the other hand, the two low-quality reference HRC reach very low quality scores. The upscaled base layer encoded with a QP of 44 (HRC 2) obtains a score close to 1, while the upscaled base layer encoded with a QP of 38 (HRC 3) reaches a MOS of 1.6. As expected, all the other HRC obtain quality scores comprised between the high and the low quality reference HRC.

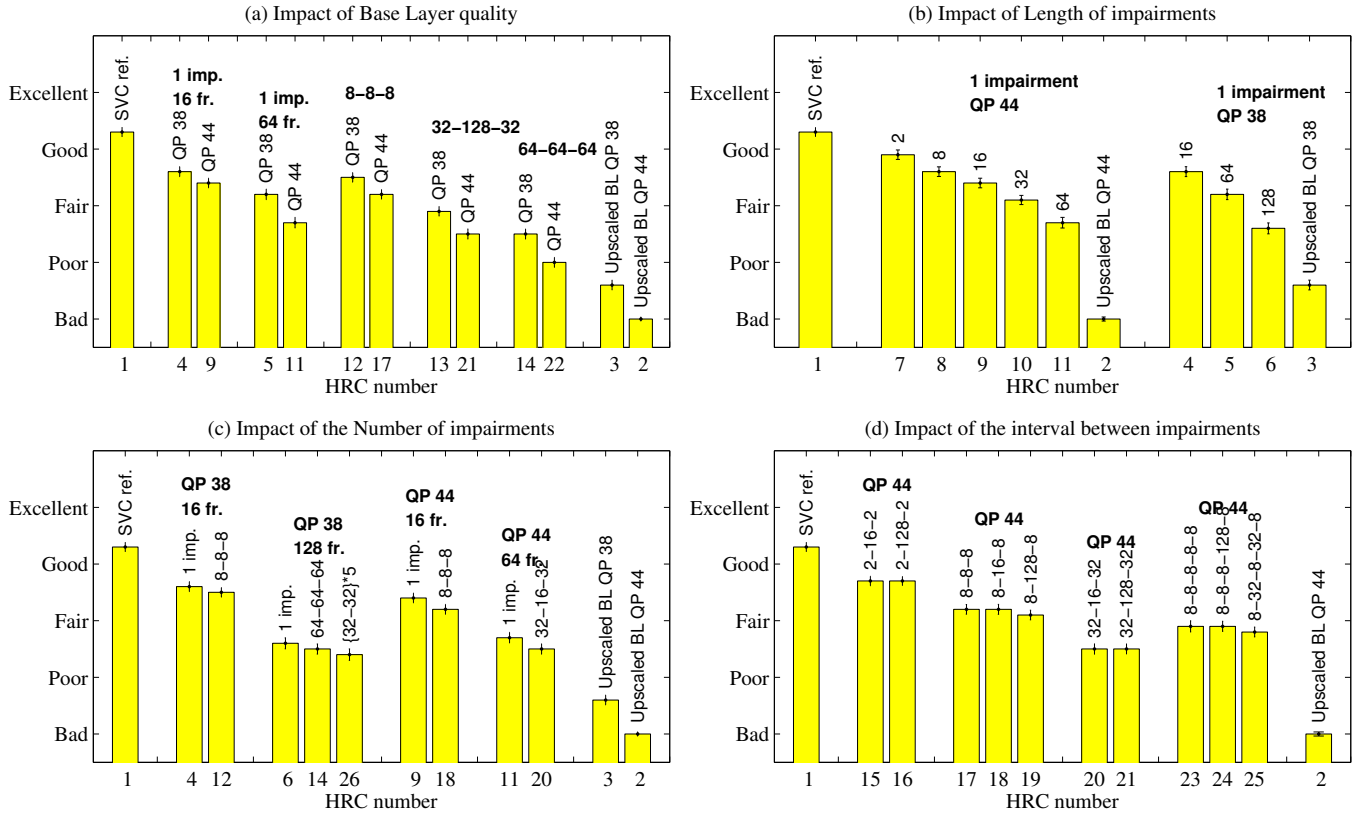

Fig. 2. Comparison of the Mean Opinion Scores for the four evaluated factors in our experiment.

4.1. Impact of the quality of the base layer

Figure 2(a) compares four couples of HRC in which the influence of the quality of the base layer can be identified. For each couple, the same impairment pattern is applied and only the QP of the base layer varies. We logically observe that the HRC using a QP of 38 for the base layer reaches higher quality than the HRC using a QP of 44. This can be easily understood as the better the quality in the base layer, the lower the difference between the upscaled version and the full-size layer. Moreover, we observe that the difference between HRC using a QP of 38 and 44 increases when the number of frames displayed from the base layer increases. For instance, the difference in MOS between HRC 4 and 9 (which display 16 frames from the base layer) is higher than the difference between HRC 5 and 11 (which display 64 frames from the base layer). A similar difference can be observed between HRC 12-17 and 14-22. However, this observation does not apply for HRC 11-19 and 12-20. In the former, 128 frames are displayed from the base layer while only 64 for the latter (both equally dispatched in 2 impairments). Meanwhile the difference in MOS are similar between the HRC of each couple. This might illustrate the impact of two other factors: the length of one impairment (64 for HRC 11-19 and 32 frames for HRC 12-20), and/or the interval between impairments (64 and 128 frames, respectively).

4.2. Impact of the length of one impairment

Figure 2(b) compares the MOS of a set of HRC with one single impairment. HRC 7 to 11 use a QP of 44 for the base layer, while HRC 4 to 6 use a QP of 38. Both groups show a relatively linear decrease in quality according to the length of one impairment. This shows that the loss in quality is closely related to the length of one impairment. Whereas it is beyond the scope of this work, the results tend to indicate that modelling this relation could be feasible using a simple linear function. One should notice that this result is quite different from what was observed in [11] with frame freeze error-concealment, where the authors identified a highly non-linear relation. Another interesting result is that a high quality base layer makes longer impairments more acceptable for viewers. For instance, HRC 4 and 8 show that an impairment of 16 frames with a QP of 38 gets a quality score equivalent to an impairment of 8 frames with a QP of 44. This statement is verified for longer impairments, such as for HRC 5-10 and HRC 6-11. One can observe here the joint impact of the quality of the base layer and the duration of an impairment.

4.3. Impact of the number of impairments

Figure 2(c) compares the influence of the number of impairments for 5 couples of HRC. Within each couple, the two HRC share the same impairment duration. We globally observe that one single impairment gets a slightly higher score than 2 impairments with the same number of frames displayed from the base layer. On the other hand, no difference is made between two and four impairments leading to 128 frames displayed from the base layer, such as demonstrated by HRC 14 and 26. As we are not able to confirm this result with other configurations, we can only make the assumption that the observers sensitivity to the number of impairments saturates between two and four impairments. Therefore we assume that the number of impairments has an influence on the visual quality, but this influence is quite limited and saturates rapidly. According to the scores of HRC 4 and 12, no significant difference is observed between one single impairment of 16 frames and 2 impairments of 8 frames. Meanwhile, the scores of HRC 9 and 18 which have the same impairment configuration, are not equal. The only difference between HRC 4-12 and 9-18 is the QP used to encode the base layer. This confirms the effect of the quality of the base layer on the visual quality and shows that this factor has a stronger impact than the number of impairments.

4.4. Impact of the interval between impairments

Figure 2(d) shows the impact of the interval between impairments. As one can observe, this factor has a very limited impact on the subjective score. This is especially meaningfull in the case of HRC 17, 18 and 19, which obtain very similar MOS values. It means that two impairments of 8 frames are perceived the same way, whether they are separated by 8, 16 or 128 frames. As a result, no merging effect is observed between two impairments which are very close to each other. Among the four factors we evaluated in our experiment, the interval between impairments has the weakest impact on the visual quality.

A summary of the impact of the four factors tested in our experiment is proposed in the conclusion section.

5. DISCUSSION AND CONCLUSION

In this paper, we presented the results of large-scale subjective experiment to evaluate the impact of network behavior and coding parameters on the visual quality using SVC-based error-concealment. We showed that the quality of the base layer to be upscaled is quite important and that a high-quality base layer allows for longer impairments to be accepted by the viewers. The duration of an impairment seems to have a linear impact on the visual quality, which is in disagreement with previous work on frame-freeze error-concealment. This goes in favour of SVC-based error-concealment as it allows for a progressive loss in quality when the duration of the impairment increases, whereas the frame-freeze faces a fast drop in quality for short impairments. The two first evaluated factors showed a stronger impact than the number of impairments and the interval between impairments. Indeed, we demonstrated that the influence of the number of impairments is only significant between one and two impairments, while the interval between impairments does not seem to have any significant influence.

In our future work, we will use the results from this experiment to build a model of the percieved quality for SVC encoded videos in a network impairment context. A thorough analysis of the variability of the results among video contents shall also be performed, in order to evaluate the influence of spatial and temporal activity on the performance of SVC-based error-concealment.

6. REFERENCES

[1] I. E. Richardson, *H.264 and MPEG-4 Video Compression*, John Wiley and Sons, 2003.

[2] Y. Wang and Q.-F. Zhu, "Error Control and Concealment for Video Comm.: A Review," in *Proc. of the IEEE*, 1998.

[3] H. Ha, C. Yim, Y. Y. Kim, "Packet loss resilience using unequal forward error correction assignment for video transmission over comm. networks," in *Computer Comm. 30*, 2007.

[4] J. Reichel, H. Schwarz and M. Wien, "Joint Scalable Video Model JSVM-11, doc. JVT-X202," Tech. Rep., Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 2007.

[5] R. Zhang, S. L. Regunathan and K. Rose, "Switched Error Concealment And Robust Coding Decisions In Scalable Video Coding," in *Proc. of ICIP*, 2000.

[6] Q. Ma, F. Wu, M.-T. Sun, "Error Concealment For Spatially Scalable Video Coding Using Hallucination," in *Proc. of ICSAS*, 2009.

[7] D. Hands and K. Brunnström, *Multimedia Group Test Plan Draft Ver. 1.21*, Video Qual. Experts Gp (VQEG), 2008.

[8] ISO/IEC JTC1/SC29/WG11 MPEG2007/N9189, "SVC Verification Test Plan, Ver. 1," Tech. Rep., Joint Video Team, 2007.

[9] T. Liu, Y. Wang, J. M. Boyce, H. Yang, Z. Wu, "A Novel Video Quality Metric for Low Bit-Rate Video Considering Both Coding and Packet-Loss Artifacts," in *IEEE Journal Of Selected Topics In Signal Proc.*, 2009.

[10] U. Reiter and J. Korhonen, "Comparing Apples And Oranges: Subjective Quality Assessment Of Streamed Video With Different Types Of Distortion," in *Proc. of QoMEX*, 2009.

[11] R. Pastrana-Vidal, J.-C. Gicquel, C. Colomes and H. Cherifi, "Frame Dropping Effects On User Quality Perception," in *WIAMIS*, 2004.

[12] Y. Pitrey, M. Barkowsky, P. Le Callet, R. Pepion, "Evaluation of MPEG4-SVC for QoE protection in the context of transmission errors," in *Proc. of SPIE Optical Imaging*, 2010.

[13] Joint Video Team, "JSVM Ref. Softw. Ver. 9.18," http://ip.hhi.de/imagecom_G1/savce/downloads/, 2009.

[14] N. Staelens, N. Vercammen, Y. Dhondt, B. Vermeulen, P. Lambert, R. Van de Walle, P. Demeester, "VIQID: A No-Reference Bit Stream-Based Visual Quality Impairment Detector," in *Proc. of QoMEx*, 2010.

[15] Question ITU-R 211/11, "ITU-R BT.500-10 Methodology for the subjective assessment of the quality of television pictures," *ITU-R BT.500-10*, 1974.

[16] ITU-T Study Group 12, "ITU-T P.910 Subjective video qual. assessment methods for multimedia appl.," *ITU-T P.910*, 1997.