SUBJECTIVE QUALITY ANALYSIS OF BIT RATE EXCHANGE BETWEEN TEMPORAL AND SNR SCALABILITY IN THE MPEG4 SVC EXTENSION

M.A.J. Barzilay, J.R. Taal and R.L. Lagendijk

Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, the Netherlands
M.A.J.Barzilaij@student.tudelft.nl, {J.R.Taal, R.L.Lagendijk}@tudelft.nl

ABSTRACT

It is well known that a compression trade-off exists between the spatial and temporal video quality. Various temporal scalability techniques have been considered for lowering the encoded frame rate such that the freed bit rate can be used to *increase* the PSNR quality of the remaining frames. Temporal scalability in the SVC extension of the MPEG4/AVC codec [1] is realized by the hierarchical B-frame structure or the open loop MCTF approach. We investigate the usefulness of the hierarchical B-frame structure as a tool for scaling the bit rate. We use the patented Video Quality Metric (VQM) to measure the subjective quality, in this way explicitly taking into account temporal artifacts due to frame rate reduction. Our results indicate that there is little to no quality gain by exchanging frame rate for increased spatial quality at a given constant bit rate.

Index Terms— Temporal scalability, hierarchical B-frames, subjective quality, Video Quality Metric (VQM)

1. INTRODUCTION

The human visual system is less sensitive to temporal details and more sensitive to spatial details if a video recording depicts a stationary scene. For video that contains a lot of motion, the opposite is true. It is therefore often postulated that the overall quality of static video sequences benefits relatively more from a spatial quality increase than from an increase in frame rate. If a video encoder supports temporal scalability we can decide to lower the frame rate in exchange for an increase of the bit rate and hence spatial quality of the remaining frames.

Temporal scalability is usually realized by dropping predicted B or P frames. Since these frames can often be efficiently predicted, only relatively little bit rate is saved in this way. Lowering the frame rate by a factor of 2 achieves a PSNR gain of 0.5 to 2 dB for the encoded remaining frames [1],[9][1]. This spatial quality increase comes, however, with a substantial increase in temporal quality degradation. It is therefore worthwhile to investigate whether this spatiotemporal bit rate exchange yields also a better overall subjective quality.

For measuring the perceptual quality of encoded video sequences with reduced frame rate, we need to grade jerky or unnatural motion in addition to the spatial quality of individual frames. The PSNR-measure is useless in expressing temporal quality. In this paper we use the VQM (Video Quality Metric) software to obtain a quality grading of original and reduced frame rate video sequences that is much closer to subjective human judgments [2].

In Section 2, we first briefly discuss the temporal coding structure of MPEG4/SVC that allows for temporal scalability. In Section 3, we summarize the operation of the VQM quality measure used for subjective quality evaluation. Experimental scenarios and simulation results as well as the evaluation of the results are presented in Section 5. A discussion and conclusion of our work is presented in Section 6.

2. TEMPORAL SCALABILITY STRUCTURE

The hierarchical B-frame structure, as used in the Joint Scalable Video Model Reference Software (JSVM), introduces temporal scalability. Frames in the lowest temporal layer are referred to as key frames (which are typically I or P frames). A key frame and all the frames that are temporally located between the key frame and the previous key frame are considered a Group Of Pictures (GOP). Within a GOP, frames are predicted in a dyadic structure as illustrated in Figure 1. This structure is superior to the traditional ("IBBP…") coding structure. Especially for low motion sequences it realizes excellent compression performance [3],[4].

For optimal overall coding efficiency, the quantization step, controlled by the Quantization Parameter (QP), differs per temporal layer. Typically, key frames have the lowest QP values and as a result have the highest PSNR value. For higher temporal layers, the QP value increases, yielding lower PSNRs. The justification for increasing the QP parameter value in higher temporal layers is that

[1] Our experimental results in Figure 2(a) show similar PSNR gains.

quantization noise introduced in a lower temporal layer should not be re-encoded in the higher temporal layers. Further, a higher quality for key frames improves the motion-compensated prediction of all non-key frames, which also improves the overall compression efficiency.

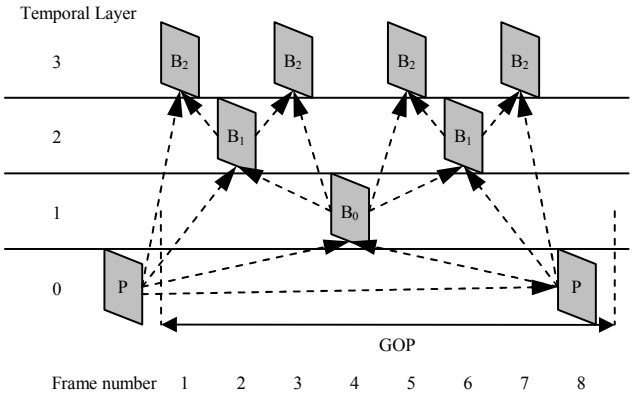


Figure 1: Hierarchical B frame structure (GOP = 8).

The value to increase the QP per hierarchical (temporal) layer is content dependent. Coding experiments with H.264/MPEG4-AVC [3] suggest the following QP settings, which are currently also used in the JSVM:

$$QP(T) = QP(T-1) + 1 \text{ for } T > 1$$
$$QP(1) = QP(0) + 3 \quad (1)$$

Here T enumerates the temporal layer, and $QP(0)$ is the QP setting for the lowest temporal layer, from which all others are derived.

The above-mentioned QP settings result in PSNR differences between key and non-key pictures of up to 4 dB (measured on luminance information) depending on the video content. The large difference in quality is not only due to less texture detail of the (hierarchical) B-frames, but is mostly caused by inaccuracies of the sub-pixel motion prediction. Since these errors are, however, less than ¼ pixel, motion in the reconstructed sequence still appears smooth and the subjective quality is therefore hardly affected [5].

In the context of our research, it is important to realize that due to the larger QP values, B-frames in high temporal layers require only a minimal amount of bit rate. This is particularly true for video sequences with static content. We anticipated, however, that especially for static content frame rate reduction would allow for increasing the subjective quality of the sequence. We therefore may already expect that spatiotemporal bit rate exchange via temporal scalability will not be as efficient as often postulated in literature.

3. VQM SOFTWARE

The VQM software is used to grade the compressed video sequences with reference to the original sequences. We use the double stimulus impairment quality scale, ranging from 1 (very bad, maximum impairment) to 5 (perfect, no impairment). A viewer does not easily notice VQM grade differences below 0.2. The final grade is calculated from a set of computed intermediate parameters, the details of which can be found in [2]. As a side result, the VQM model computes percentages of blurriness, jerky motion, global noise and block distortion confirming our own informal subjective evaluations.

The intermediate parameters for computing the VQM grade are computed on spatiotemporal regions and not individual frames, evaluating both temporal and spatial impairments at the same time. One important parameter is the absolute temporal-information loss, a measure for the amount of jerky motion. Another important parameter is the spatial-information loss. This parameter basically computes the spatial quality compared to the original sequence and is a measure for the amount of blurriness. The low PSNR values of high level B-frames due to errors in motion compensation (barely influencing subjective quality) might affect the VQM grade slightly since comparison with the reference sequence is made. It should therefore be noted that the VQM might slightly under grade higher framerate sequences compared to lower framerates.

4. SCENARIOS AND RESULTS

4.1. Test Setup

The objective of our experiments is to measure the loss or gain in overall (visual) quality when reducing the frame rate while at the same time keeping the encoded bit rate constant. We carry out experiments for different bit rates and sequences. Since we wish to evaluate the effectiveness of a temporal scalable video stream at different bit rates, we obtain the different bit rates by employing three Fine Granular Scalability (FGS) enhancement layers. Spatial scalability as employed in the JSVM is not the objective of our work and therefore not used.

We used 240 frames of CIF-format sequences at 30 fps. The sequences contain different amount of motion and texture information namely: *foreman* (medium texture and medium irregular motion), *mobile* (high texture and low natural motion), *football* (medium texture and high irregular motion), and *bridge* (medium texture and barely any motion). The bit rates selected for 7.5, 15 and 30 fps are 100, 200 and 500 kbit/s. For the static video *bridge* we also evaluated 50 kbit/s.

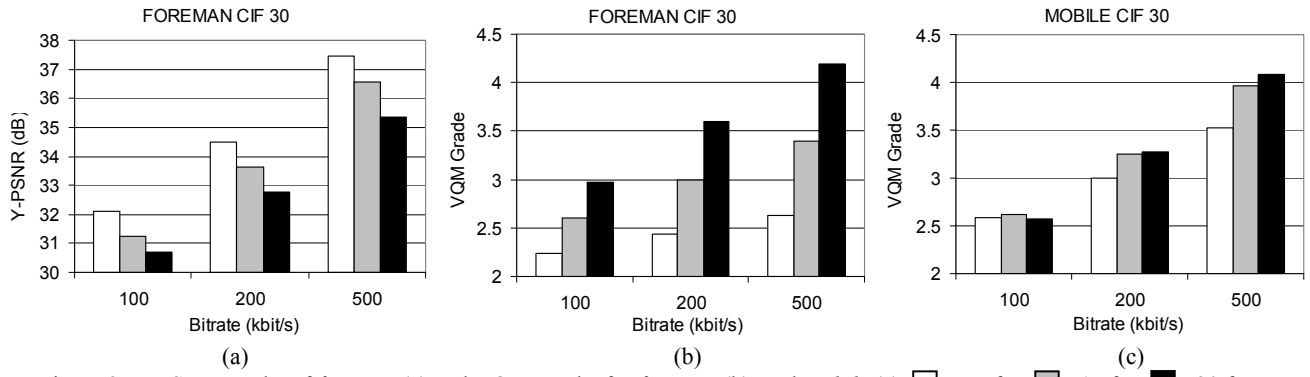


Figure 2: Y-PSNR results of *foreman* (a) and VQM results for *foreman* (b), and *mobile* (c). □ =7.5 fps, ▒ =15 fps, ■ =30 fps.

In order to make fair comparisons, the decoded 7.5 fps and 15 fps sequences were temporally up-converted to 30 fps. We do this in two different ways. The first up-conversion is by simple frame repetition, which is a method most display devices will use. The second up-conversion utilizes an advanced temporal interpolating method to obtain the original 30 fps. Details of this method are outside the scope of this paper [6]; we consider the results to be an upper bound to the quality that can be achieved by advanced temporal post processing.

4.2. Optimization of the JSVM encoder settings

The effectiveness of temporal scalability to reduce bit rate is dependent on the amount of data needed for the higher layer B-frames. There are various encoder settings that influence the quality of these B-frames. The mode decision parameters (QPMD) influence the coding mode chosen for the macro-blocks. Every temporal layer has an individual QPMD parameter to be set. For more information about these parameters the reader is referred to [7]. The layer dependency of the QP parameters is given by Eq. (1), which influences the optimal setting of QPMD. The relation between QP and QPMD is chosen as described in [8]. The QP parameter is configured such that the encoded bit rates cover the desired range of 100 to 500kbit/s. Selecting a specific $QP(0)$ indirectly defines the smallest rate of the lowest FGS layer (base layer) and the largest rate of the highest FGS layer since each FGS layer corresponds to a QP-decrease of 6 [4].

The coding efficiency at high bit rates is affected by using a low base layer bit rate. This is inherent to the SVC and inevitable since we want to encode the sequence for relatively low bit rates and a wide range. Our comparisons are, however, still fair since sequences with the same bit rate are compared.

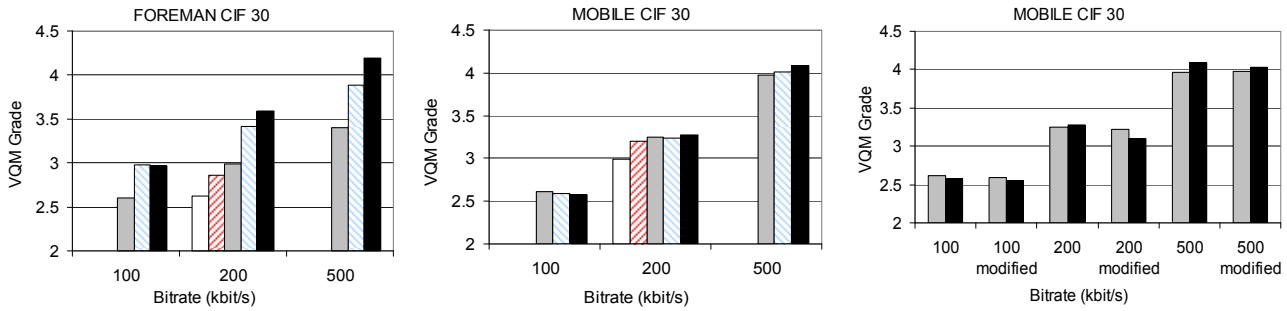
We used closed loop (thus no MCTF) encoding of the GOPs, closing the loop at both base layer and the highest FGS layer. Other encoder parameters are identical to the settings applied in the Palma test conditions included in the reference software [1].

Finally, the Quality Level Assigner tool of the JSVM is used in order to optimize the rate distortion curve of the extracted sub-streams. This is especially useful for our experimental setup since this tool is used for a scalable bit stream containing progressive refinement NAL (Network Abstraction Layer) units. A quality layer identifier is assigned to each NAL unit that can be used during the extraction of a sub-stream with desired bit rate [1].

4.3. Results

Figure 2(a) shows the Y-PSNR values of *foreman* sequence for the three encoded bit rates, and the three frame rates. Luminance PSNR gain is observed for the remaining frames of the lower frame rate sequences. Results of our VQM evaluation using frame repetition on the 7.5 and 15 fps results are shown in Figure 2(b) and (c). The *football* sequence shows results similar to *foreman* and the *bridge* sequence gave results similar to *mobile* [10]. The result in Figure 2 and all our other experiments show that exchanging bit rate freed by lowering the frame rate for an increase in spatial quality of the remaining frames, does not lead to a higher overall subjective quality for any sequence, except for the static *bridge* sequence and the 100 kbit/s version of *mobile*. However, even for *bridge* the VQM grade for lower frame rates is only marginally better, in most cases below the 0.2 grade-difference threshold. Informal subjective evaluations confirm the VQM grading results.

Figure 3(a) and (b) show the VQM result for *foreman* and *mobile* using advanced temporal frame interpolating on the 7.5 and 15 fps coding results. When there is a lot of jerky motion due to low frame rate, the sequences are graded significantly higher than when using simple frame repetition. However, interpolation of low/regular motion video does not improve the subjective quality much compared to simple frame repetition. Although for fast/irregular motion video the interpolated sequences are graded higher than frame repetition, they are not graded higher than the 30 fps sequences. Apparently, not all spatiotemporal details of the dropped frames can be recovered, not even with advanced post processing techniques.



(a) (b) (c)

Figure 3: Frame repetition and advanced interpolation of *foreman* (a) and *mobile* (b). In (c), VQM Grade for mobile with similar QP setting for all temporal layers. □ =7.5 fps frame repetition, ▨ =7.5 fps interpolated, ■ =15 fps frame repetition, ▨ =15 fps interpolated, ■ =30 fps.

4.4. Modified settings for the QP of frames

There is a big difference in PSNR of individual frames due to the layer-dependent QP parameter. Frames in higher temporal layers therefore contain relatively less information. We have modified the reference software to give all temporal layers the same QP setting. This on one hand leads to a smoother PSNR over the frames, but on the other hand also to less average PSNR quality at the same rate. However, the effect on subjective quality when using constant-QP settings was not yet clear.

For mobile we have compared the standard encoder using layer-dependent QP-setting with the "modified" encoder using similar QP-settings for all temporal layers. We did that at 100, 200 and 500kbit/s decoding bit rates at 15 and 30fps. The results are shown in Figure 3(c). We see that results labeled as "modified" have an overall lower VQM quality. But we also see that most of this reduction is for 30fps sequences, while the sequences for which the 15fps substream is extracted have roughly kept their quality.

In conclusion, having constant-QP settings does not favor temporal scalability over SNR scalability, especially when we consider that subjective perceptual quality is not much affected by layer-dependent QP settings.

5. DISCUSSION AND CONCLUSION

The usefulness of temporal scalability as a tool for increasing subjective video quality by temporal to spatial bitrate exchange is questionable. High temporal layer B-frames can, in general, be encoded very efficiently. These frames have a lower quantization parameter and PSNR value than their reference frames. Encoding all temporal layers with identical QP setting introduces relative higher quality gain for lower frame rates but decreases the overall coding efficiency. Advanced frame interpolation improves the overall VQM quality of 15fps and 7.5fps sequences, but is never much higher than the 30fps sequence.

Certain devices such as mobiles might demand a lower video frame rate since there is limited computing power on the decoder side. For these applications temporal scalability is useful, however, frame rate selection is limited to factors of 2. Furthermore, FGS offers already bitrate scalability of a factor around 10, which is sufficient in many cases.

6. ACKNOWLEDGEMENTS

The authors would like to thank Rene Klein Gunnewiek from Philips Research for his contribution of interpolating a selection of the sequences.

REFERENCES

[1] ITU-T and ISO/IEC (JVT), "Joint Scalable Video Model 6.8.2," October 13, 2006.

[2] S. Wolf and M. Pinson, "VQM Software and Measurement Techniques," *National Telecommunications and Information Administration (NTIA'02) Report 02-392*, June 2002.

[3] T. Wiegand and J-R. Ohm, "Scalable Video Coding – Standardization and beyond," *IEEE International Conference on Image Processing (ICIP'06)*, Atlanta, GA, USA, October 2006, Tutorial.

[4] H. Schwarz, D. Marpe, "Overview of the Scalable H.264/MPEG4-AVC Extension," *IEEE International Conference on Image Processing (ICIP'06)*, Atlanta, GA, USA, October 2006, Invited Paper.

[5] H. Schwarz, D. Marpe and T. Wiegand, "Hierarchical B pictures," Poznan, Poland, Doc. JVT-P014, July 2005.

[6] E. Bellers, J. van Gurp, J. Janssen, R. Braspenning and R. Wittebrood, "Solving Occlusion in Frame-Rate Up-Conversion," *Proc. The International Conference on Consumer Electronics (ICCE'07)*, January 2007.

[7] S-H Kim and YS Ho, "Optimum Quantization Parameters for Mode Decision in Scalable Extension of H.264/AVC Video Codec," *in PCM*, Vol. 1, pp. 179-190, 2005.

[8] M. Wien and H. Schwarz, "Testing Conditions for Coding Efficiency and JSVM Performance Evaluation," *16th Meeting*, Poznan, Poland, Doc. JVT-P205, July 2005.

[9] W.J. Han, "CE6 response of Samsung Electronics: in-depth comparison of closed-loop and open-loop MCTF structures," *16th Meeting*, Poznan, Poland, Doc. JVT-P084, July 2005.

[10] http://ict.ewi.tudelft.nl/~jacco/icip07/results.xls