

THE INFLUENCE OF BIT RATE ALLOCATION TO SCALABILITY LAYERS ON VIDEO QUALITY IN H.264 SVC

Andreas Unterweger
Salzburg University of Applied Sciences
Urstein Süd 1
5412 Puch/Salzburg, Austria
aunterweger.tks2004@fh-salzburg.ac.at

Herbert Thoma
Fraunhofer Institute for Integrated Circuits
Am Wolfsmantel 33
91058 Erlangen, Germany
tma@iis.fraunhofer.de

ABSTRACT

After the finalization of the scalable video coding standard, H.264 SVC faces a wide range of usage areas. In an effort to save bandwidth (and therefore costs) while satisfying the customers' need for high perceptual quality, it is necessary to find an optimal tradeoff between base and enhancement layer bit rate. This paper will analyse the influence of the possibilities H.264 SVC offers to maximize the video quality for a given total bit rate. This will be achieved by varying the number of spatial and quality layers in the SVC stream and the bit rate spent for each layer. The bit rate distributions considered ideal in terms of overall quality will be outlined and compared with the distributions analysed.

Index Terms— H.264 SVC, Video quality, Bit rate allocation, SSIM

1. SCALABLE VIDEO CODING WITH H.264 SVC

H.264 SVC is a video coding standard which allows the creation of bit streams which are temporarily, spatially and qualitywise (SNR) scalable, using a layer concept for each scalability dimension[1].

For the tests carried out, temporal scalability is used because temporal scalability using hierarchical B pictures in general increases coding efficiency compared to classical IBBP coding[1]. The effect of varying frame rates between the spatial layers (resulting in different GOP sizes in the spatial layers) is not taken into consideration. Tests in [1] show that the maximum coding efficiency can be achieved with a GOP size of 4 frames for sequences with fast and complex motion and a GOP size of 8 frames for sequences with high spatial detail. Thus, GOP sizes of 4 and 8 are used for the tests.

All tests are performed using version 9.1 of the reference software of the SVC project, the JSVM (Joint Scalable Video Model[2]) software.

2. VIDEO QUALITY MEASUREMENT

To compare the visual quality of image sequences, video quality measurement methods which consider the human visual perception are required. With SSIM (structural similarity) and VQM (video quality metric) two such

methods are taken into account. Due to the fact that the Pearson's product moment correlation coefficient between subjective quality and the quality measured by the SSIM metric is higher than the coefficient between subjective quality and the quality measured by the Y-PSNR and the VQM metric[3], the most suitable measurement method for comparison purposes in this paper is SSIM.

The SSIM index measures structural distortion by determining the average luminance and contrast of the picture and the linear relationship between the original and the distorted picture as a value between -1 (totally distorted reconstruction) and 1 (exact copy of the original) [4, 5]. When calculating SSIM values of an image sequence, the average SSIM (MSSIM) is used. To measure the MSSIM index, the MSU Video Quality Measurement Tool (compares two sequences by computing quality metrics; http://www.compression.ru/video/quality_measure/video_measurement_tool_en.html) is used due to the fact that the GUI offers the possibility to visualize the SSIM index for each frame which makes the recognition of errors in measurements easier.

3. TEST SETUP

The goal of the tests carried out is to determine the optimal bit rate allocation to spatial and CGS layers in terms of quality. In the following section, the test setup and the sequences chosen for testing are described.

Three sequences with high spatial detail and three sequences with fast motion are accounted for the test in order to cover two important types of sequences. As the importance of HD has been increasing over the last couple of years and will continue to do so in the near future, both SD and HD sequences are used for the tests. The sequences are obtained from the publicly accessible FTP servers of the VQEG (SD) and the TU München (HD) respectively. The following sequences (name (cropped spatial resolution@frame rate)) are used for testing: src5_ref_625 (704-576@50), src6_ref_625 (704-576@50), pedestrian_area (1920-1024@25), src14_ref_525 (704-448@60) src21_ref_525 (704-448@60) and shields_ter (1280-704@50). The first three sequences are characterized by fast motion, the last three are characterized by high spatial detail.

It turned out that bit rates of 1500 kbit/s are suitable for the SD sequences used for this test because the MSSIM

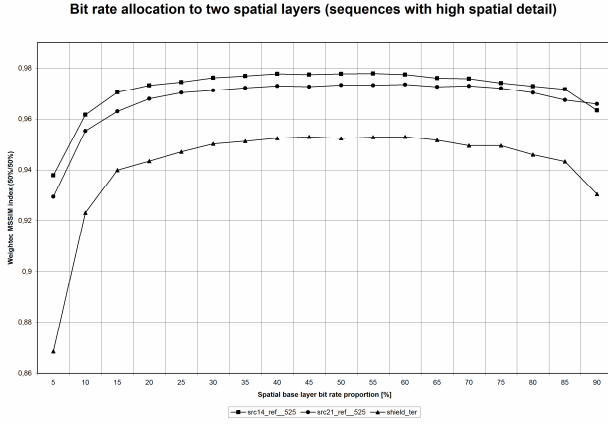


Figure 2: Weighted quality (50%/50%) of two spatial layers (sequences with high spatial detail)

value of every encoded sequence is below 1. Since the sequence shields_ter has 2.222 times more pixels than the SD sequences, the bit rate for this sequence is set 2.222 times higher to 3333 kbit/s. The bit rate for the sequence pedestrian_area is set to 3750 kbit/s, taking the reduced frame rate into consideration.

All sequences are encoded with level 3.1, CABAC, 8x8 transform and n reference frames (where n is the GOP size) using a hierarchical coding structure which presumably corresponds to the scalable high profile. The VQEG sequences are deinterlaced prior encoding using the TDeint AviSynth plugin (http://avisynth.org/warpenterprises/files/tdeint_25_dll_20060410.zip) with its default settings.

After extracting 65 frames of each sequence for encoding, the extracted frames are downsampled dyadically if necessary using the MPEG-4 downsampling filter[2]. After this, the encoder is run multiple times for each bit rate distribution using a fixed Lagrangian parameter which is adjusted in order to meet the layer bit rates specified. The range of the bit rate distributions is explained in the corresponding test section. A step size of 5% per layer with a maximum error of 2% is used in order to ensure that the results are not influenced by the error.

In the evaluation phase, the reconstructed files created for each bit rate distribution during the encoding process are compared with the originals, measuring the MSSIM index of each layer. The results of all layers of one distribution are summed up using weights which consider each layer equally in order to take the quality requirements of every layer into consideration. This leads to the weighted sum for two layers

$$WMSSIM_{2L} = \frac{MSSIM_{L1}}{2} + \frac{MSSIM_{L2}}{2}$$

and three layers

$$WMSSIM_{3L} = \frac{MSSIM_{L1}}{3} + \frac{MSSIM_{L2}}{3} + \frac{MSSIM_{L3}}{3}$$

Local maxima of the WMSSIM (weighted MSSIM) graphs then reveal maxima of quality in the distribution of bit rates for each sequence. Differences below 0.001 between a value and the maximum are treated as being

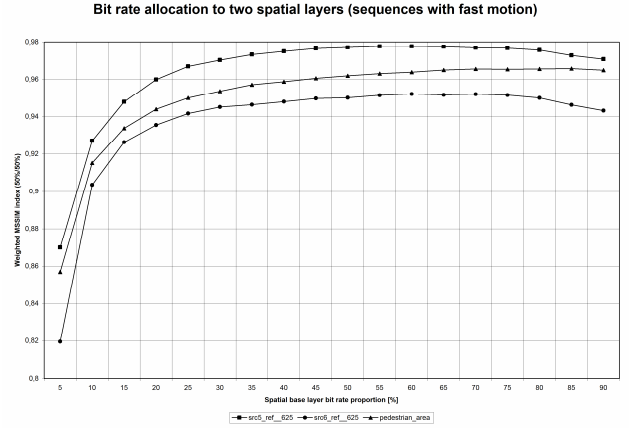


Figure 3: Weighted quality (50%/50%) of two spatial layers (sequences with fast motion)

unnoticeable and therefore zero, causing multiple maxima.

Depending on the application, different weights are possible. Figure 1 depicts the quality maxima of a test setup with two spatial layers as a function of possible weights to illustrate their effect on the position of the maxima.

4. BIT RATE ALLOCATION TO SPATIAL LAYERS

This first test analyses the optimal bit rate distribution between two spatial layers. The distribution of bit rates starts at 5% for the spatial base layer and 95% for the spatial enhancement layer respectively. With a step size of 5% (compare section 3), the bit rate of the spatial base layer is increased until 90% are reached.

To visualize the results, the sequences with high spatial detail and the sequences with fast motion are separated and depicted in two diagrams due to their different conditions for coding. Each diagram shows the weighted MSSIM index (as described in section 3) over the bit rate of the spatial base layer. The according bit rate proportion of the spatial enhancement layer is 100% minus base layer bit rate proportion.

Figure 2 depicts the results for the sequences with high spatial detail, figure 3 depicts the results for the sequences with fast motion. The former show common quality maxima between 40%:60% (base:enhancement layer bit rate proportion) and 60%:40%, the latter between 65%:35% and 75%:25%. For sequences with high spatial detail, the weighted quality decreases significantly when choosing base layer bit rate proportions less than 30%.

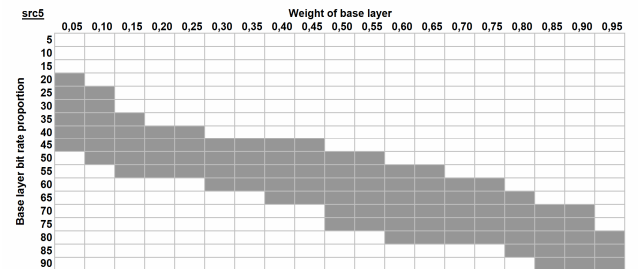


Figure 1: Quality maxima of src5_ref_625 (test setup with two spatial layers) with different weights. Grey areas indicate maxima of the corresponding column ± 0.001 .

Sequence	Maxima (3 layers)	Position of maxima (%:%:%)	Maxima (2 layers)	Position of maxima (%:%)
<i>src5_ref_625</i>	0.97850 0.97920 0.97940 0.97897 0.97861	15:35:50 15:40:45 15:45:40 15:50:35 15:55:30	0.97701 0.97760 0.97757 0.97739 0.97687 0.97667	50:50 55:45 60:40 65:35 70:30 75:25
<i>src6_ref_625</i>	0.95551 0.95594 0.95519 0.95518	15:45:40 15:55:30 15:60:25 15:65:20	0.95145 0.95208 0.95162 0.95203 0.95156	55:45 60:40 65:35 70:30 75:25
<i>pedestrian_area</i>	0.96552 0.96617 0.96591 0.96602 0.96601	15:55:30 15:60:25 15:65:20 15:70:15 15:75:10	0.96492 0.96550 0.96535 0.95575 0.96570 0.96488	65:35 70:30 75:25 80:20 85:15 90:10
<i>src14_ref_525</i>	0.98004 0.98044 0.98044 0.98051 0.97972	15:30:55 15:35:50 15:40:45 15:45:40 15:50:35	0.97696 0.97777 0.97751 0.97783 0.97795 0.97754	35:65 40:60 45:55 50:50 55:45 60:40 70:30
<i>src21_ref_525</i>	0.97867 (see text)	15:45:40 (see text)	0.97305 0.97344 0.97388 0.97363 0.97308	40:60 50:50 55:45 60:40 70:30
<i>shields_ter</i>	0.95897 0.95861	15:40:45 15:45:40	0.95248 0.95306 0.95243 0.95293 0.95303	40:60 45:55 50:50 55:45 60:40

Table 1: Maxima of weighted quality of three spatial layers, compared with the maxima of two spatial layers. Sequences with fast motion are in italics; all values are rounded to 5 decimals

The same applies for sequences with fast motion for base layer bit rate proportions less than 30%.

In order to extend these conclusions, this test analyses the optimal bit rate distribution between three spatial layers. The distribution of bit rates starts at 5% for the spatial base layer, 5% for the first spatial enhancement layer and 90% for the second spatial enhancement layer respectively. With a step size of 5% (compare section 3), the bit rate of the first spatial enhancement layer is increased until the bit rate proportion of the second spatial enhancement layer equals the step size.

Preparations for this test showed that a base layer bit rate proportion above 15% is not reasonable because the base layer MSSIM is then above 0.995 which does not allow a meaningful conclusion because a MSSIM of 1 indicates a quality loss of zero and therefore requires low base layer QPs which are not suitable in this setup.

Table 1 summarises the quality maxima of all sequences for two and three spatial layers. The results are separated into the two sequence types described in the tests with two spatial layers.

As can be seen in table 1, sequences with fast motion

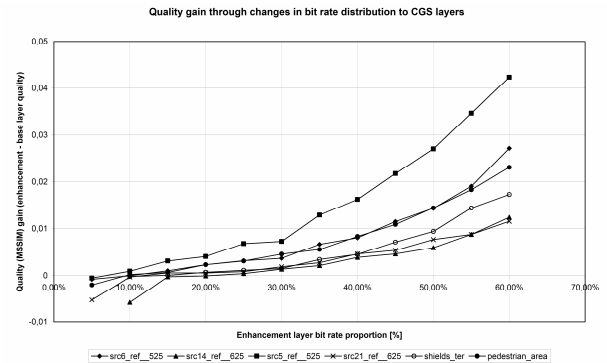


Figure 4: The effect of varying bit rate distribution to 2 CGS layers on quality (all sequences)

show a common maximum at 15%:55%:30% (base:first enhancement:second enhancement layer bit rate proportion). The maxima of sequences with high spatial detail cannot be generalised because bit rate mismatches of more than 2% appeared during the tests of *src21_ref_525* at the bit rate proportions where *src14_ref_525* and *shields_ter* show maxima. Increasing the step size to 10% (omitting every second test result) allows a conclusion because all mismatches are below 5%. This leads to a maximum at 15%:45%:40% for sequences with high spatial detail and a step size of 10%.

5. BIT RATE ALLOCATION TO CGS LAYERS

Besides spatial scalability, H.264 SVC supports two forms of SNR scalability: CGS (Coarse Grain Scalability) and FGS (Fine Grain Scalability). CGS can be seen as a special form of spatial scalability with layers of equal size but different quality[2]. This test extends the optimal bit rate distributions for spatial configurations to CGS configurations with 2 layers.

Due to the fact that CGS coding involves layers of equal size, the overall bit rate spent is increased to 2000 kbit/s for the SD sequences and multiplied accordingly for the HD sequences, as described in section 3.

Figure 4 shows the gain in quality when allocating more bit rate to the enhancement layer than to the base layer. It is clearly visible that sequences with fast motion profit from a higher enhancement layer bit rate proportion to a greater extent than sequences with high spatial detail – the faster the motion the higher the gain.

Enhancement layer bit rate proportions greater than 60% are not taken into consideration because the bit rate of the according base layer cannot be met within the mismatch range. Negative quality differences indicate an enhancement layer quality lower than the base layer quality. This can occur when the enhancement layer bit rate proportion is below 10%.

The greatest difference between base and enhancement layer quality can be measured when allocating 60% of the overall bit rate to the enhancement layer. In this configuration, the base layers of all sequences show absolute MSSIM values greater than 0.91.

In order to take advantage from the ability of CGS streams to provide various levels of different quality,

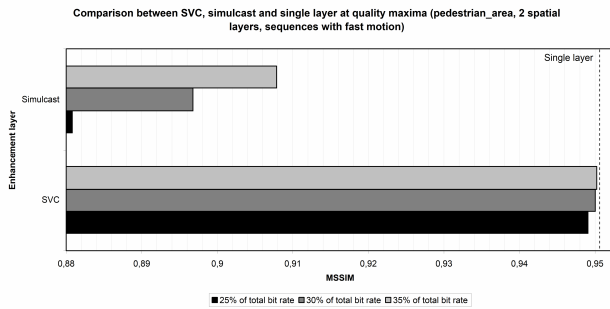


Figure 5: Comparison between SVC, single layer and simulcast at the quality maxima (enhancement layer quality only)

distributions which cause greater differences between two consecutive layers are preferred. This leads to the conclusion that it is reasonable to allocate 40% of the bit rate to the CGS base layer and 60% to the CGS enhancement layer. This distribution also causes quality maxima for sequences with high spatial detail in the tests with 2 spatial layers.

It has to be noted that the quality difference between the two CGS layers is only between 0.01 and 0.02 for sequences with high spatial detail. Therefore CGS is not reasonable for this type of sequences because of the low quality gain when switching from the base to the enhancement layer.

CGS configurations with more than two layers allow no further conclusions because of high bit rate mismatches.

6. COMPARISON WITH SINGLE LAYER AND SIMULCAST CONFIGURATIONS

When analysing scalable video configurations, comparisons with the corresponding single layer and simulcast configurations are necessary in order to assure that the gain of flexibility through scalability does not result in a loss of coding efficiency. Therefore, the distributions considered as optimal in the conclusions drawn in section 4 and 5 are being compared with their single layer and simulcast equivalents using the JSVM software.

Figure 5 shows this comparison for one sequence with fast motion in a configuration with two spatial layers at the maxima described in section 4. This sequence is representative for all other configurations with two and three spatial layers and two CGS layers.

Simulcast (coding the sequence as two independent AVC streams with the bit rate proportions of the base and the enhancement layer respectively) performs weak compared with SVC because SVC takes advantage of the redundant information in the base layer, using inter-layer prediction mechanisms.

The difference between single layer coding (coding the sequence as AVC stream with 100% bit rate proportion) and the corresponding SVC configurations is insignificantly small (0.0001) when allocating 35% of the overall bit rate to the enhancement layer. When allocating 25% or 30% to the enhancement layer, the difference is

interpreted as zero as described in section 3.

All configurations with two and three spatial layers reveal differences below 0.001 between single layer coding and the corresponding SVC enhancement layer configuration. The differences are 5 times higher when using CGS. In general, sequences with fast motion show greater differences between simulcast and SVC configurations than sequences with high spatial detail.

7. CONCLUSION

The tests carried out analysed the distribution of a fixed total bit rate to a varying number of spatial and CGS layers. The results demand a separation into sequences with high spatial detail and sequences with fast motion and reveal clear maxima which represent optimal bit rate distributions for the layers configured. Whereas the differences between SVC and simulcast configurations are regarded as being significant at the optimal ratio, the differences between SVC and single layer coding are regarded as being insignificant.

The paper did not consider the following aspects which can be subject of future research: first, a different set of weights used for the summarisation of the weighted quality for spatial configurations can be developed in order to reflect a specific application or to integrate the perceptual quality achievable on different devices with different decoding capabilities regarding spatial layers. Moreover, the tests can be rerun distributing the bit rate to more than three spatial layers. Furthermore, the influence of a varying number of temporal layers inside the spatial or CGS layers can be analysed. Finally, the tests can be rerun with an encoder which supports rate control in order to meet the target bit rates in a more precise way, allowing more exact measurements.

Altogether, the results gained through the tests carried out can be seen as a first approach to optimize the perceptual quality of H.264 SVC bit streams based on the bit rate proportion allocated to each layer.

8. REFERENCES

- [1] Schwarz, H., D. Marpe, and T. Wiegand: Overview of the Scalable Extension of the H.264/MPEG-4 AVC Video Coding Standard. In IEEE International Conference on Image Processing (ICIP 2006), Atlanta, GA, USA, October 2006.
- [2] Joint Video Team and ITU-T Video Coding Experts Group: JSVM Software Manual, June 2007.
- [3] Roitzsch, M.: Slice-Level Trading of Quality and Performance in Decoding H.264 Video. Diploma thesis, Technische Universität Dresden, June 2006.
- [4] Dosselmann, R. and Yang X.D.: Existing and Emerging Image Quality Metrics. Canadian Conference on Electrical and Computer Engineering, 2005, pages 1906–1913, May 2005.
- [5] Wang, Y.: Survey of Objective Video Quality Measurements. Tech report, Worcester Polytechnic Institute, June 2006.