Quality Assessment of the MPEG-4 Scalable Video CODEC

Florian Niedermeier, Michael Niedermeier, and Harald Kosch

Department of Distributed Information Systems
University of Passau (UoP)
Passau, Germany

niederme@fim.uni-passau.de, niedermm@fim.uni-passau.de,
harald.kosch@uni-passau.de

Abstract. In this paper, the performance of the emerging MPEG-4 SVC CODEC is evaluated. In the first part, a brief introduction on the subject of quality assessment and the development of the MPEG-4 SVC CODEC is given. After that, the used test methodologies are described in detail, followed by an explanation of the actual test scenarios. The main part of this work concentrates on the performance analysis of the MPEG-4 SVC CODEC - both objective and subjective. Please note that this document is only a shortened version of the assessment. Further experimental results can be found in the extended version available at the Computing Research Repository (CoRR).

1 Introduction

As both high visual quality and low bandwidth requirements are key features in the emerging mobile multimedia sector, MPEG and VCEG introduced a new extension to the MPEG-4 AVC standard - scalable video coding (SVC). Its focus lies on supplying different client devices with video streams suited for their needs and capabilities. This is achieved by employing three different scalability modes: Spatial, temporal and SNR scalability. Because these new features are still in development and their impact on visual quality has not often been independently tested, this paper covers this subject. The performance evaluation is done using both objective and subjective methods. Additionally to the evaluations covering the matter of visual quality, test runs are performed to check the encoding speed of the SVC CODEC. The assessment is divided into two parts: The first is a MPEG-4 SVC stand-alone test, which examines the impact of different encoding settings on the CODEC's performance. The second part consists of a competitive comparison of the SVC reference CODEC, x264 (MPEG-4 AVC based) and Xvid (MPEG-4 ASP based), to analyze each CODEC's advantages and disadvantages.

2 Related Work

Some comparisons of subjective and objective assessment methods have been conducted, especially the CS MSU Graphics & Media Lab Video Group ran several evaluations concerning CODEC competitions [3] [2]. The emerging MPEG-4

SVC standard, however, has not been tested in such a manner. Although both objective [14] and subjective tests [5] have already been run separately, an analysis offering both test methodologies was yet outstanding. The MPEG-4 SVC CODEC was also evaluated in an official ISO test [7], which did however not assess a broad range of quality-impacting parameters. Another problem concerning this evaluation is that it only focused on the comparison of MPEG-4 SVC and its direct predecessor MPEG-4 AVC. In this paper, a broader range of quality-affecting settings and scenarios is assessed. Additionally, a comparative synthesis that comprises both subjective and objective test methods is conducted.

3 Used Test Methodologies

To provide comparable results, it is important for both objective and subjective assessments to be run under strictly specified conditions. This means for objective tests that the used metric and the encoding parameters are kept throughout the whole assessment. The subjective evaluations also need to have a fixed testing setup and environment as environmental influences can bias a users' opinion.

3.1 Objective Metrics

PSNR: The PSNR is the currently most widely used metric for quality evaluations of compression techniques. Even though this metric can be calculated for luminance as well as chrominance channels, it is common to just calculate the difference in luminance (Y-PSNR). The correlation of PSNR to subjective quality impression is discussed controversially: The results of the video quality experts group [4] come to the conclusion that PSNR correlation is on par with that of other metrics. In contrast, newer tests like [2] claim that the correlation of PSNR is significantly lower than that of the SSIM metric [13]. Still, PSNR is the standard metric used in most quality assessments and literature. To ensure comparability, this metric will be used in the following tests too.

PSNR adaption for temporally scaled videos: As shown in [6], normal PSNR calculation is not suitable for quality assessment of videos with temporal scalability. The calculated values are too low to accurately reflect perceived quality, so the following adapted quality score based on PSNR was proposed: $QM = PSNR + m^{0.38} (30 - FR)$. QM is the metrics score, FR is the framerate of the processed video. To calculate $PSNR$ in this equation, the frames of the temporally scaled video are repeated to match the frame count of the original sequence. Using this sequence, standard PSNR is calculated. The parameter m is the normalized average magnitude of large motion vectors, which is used to measure motion speed. The exact calculation is given in [6]. The equation was specifically designed for videos with a maximum framerate of 30 Hz. As the source videos used in the following work have different framerates, the following has to be considered: A simple adaption of the equation to fit the new source framerate ($QM = PSNR + m^{0.38} (60 - FR)$) does not lead to reasonable results,

so the impact of temporal decimation is only considered if the framerate drops below 30 Hz. This means that sequences with a framerate of 30 Hz or lower are always compared against those with 30 Hz, so the metric described in [6] can be used without modification.

3.2 SAMVIQ

The Subjective Assessment Methodology for Video Quality (SAMVIQ) is an invention of the EBU (European Broadcasting Union), which started in 2001 and finished in 2004. It is incorporated in ITU-R BT.700 by now [10]. SAMVIQ was developed because most other subjective test methodologies (for example DSIS, DSCQS, SSCQE and SDSCE) are specialized in rating videos shown on TV screens, and not on home computer or even mobile devices. At the beginning of the test process, the subject watches the reference sequence. After that the expert has to watch and rate all impaired sequences, which are randomly ordered and made anonymous to the expert by labeling them alphabetically. If required, every sequence may be repeated as often as the tester likes. It is also possible to change the rating of a sequence anytime. The reference is also hidden among the impaired sequences and is therefore rated as well. For voting, a linear, continuous scale with a range of 0 to 100 points is used, where a higher value represents better image quality and a lower one worse quality respectively [10] [8].

4 Test Setup

Selection of experts: A total of 21 persons of all age and working classes are included in the test. None of the experts was previously trained as a subjective tester or had a job associated with visual quality testing. However, before a person is approved as an expert in the evaluation, two aptitude tests are run: A visual acuity and a color blindness test. The visual acuity of every viewer is tested using the Freiburg Visual Acuity, Contrast & Vernier Test (FrACT). The process is thoroughly described in [1]. An acuity minimum of 1.0 is necessary to take part in the following quality evaluation. Vision aids like glasses are permitted in the test. The color perception is also an important factor when assessing graphical material. Persons with a visual impairment of the color perception are excluded from the test [9]. This test is executed using the standard Ishihara test charts. After these tests, one person had to be excluded, leaving 20 test subjects for the subjective assessment.

Subjective test environment: The testing environment is set up as follows: To prevent any unwanted display-related influences, the same device (a Samsung R40-T5500 Cinoso notebook, further technical details are shown in table 2) is used for every test session and expert. The black level and contrast of the display are adjusted using a PLUGE (Picture Line-Up Generation Equipment) pattern [12]. During the playback of the sequences the test room's background lighting is provided by a faint, artificial light source. The viewing distance is set concerning

the rules of Preferred Viewing Distance (PVD) for an 15.4" LCD device. The display is aligned both horizontally and vertically to provide a viewing angle of $\leq 20^\circ$ to the expert.

Encoder settings: Three CODECs are assessed in the comparison: Xvid 1.1.3 (MPEG-4 ASP), x264 core 59 r808bm ff5059a (MPEG-4 AVC) and the new MPEG-4 SVC reference encoder 9.12.2. All encoder parameters are kept at default settings except for the settings listed in table 1.

Table 1. x264 and SVC encoder settings

x264		SVC	
Encoding Type	Single pass - bitrate-based (ABR)	GOPSize	4
Max consecutive	2	SearchMode	4
Threads	4	BaseLayerMode	2

The 'GOPSize' parameter is changed to a value of 4 to enable the usage of B frames. Encoding a video sequence without B frames would result in a significant drop in compression efficiency. The fast search algorithm is used, so 'SearchMode' is adjusted to 4. The parameter 'BaseLayerMode' is altered as the default setting is invalid.

5 Conducted Evaluations

The assessment is split in two separate evaluations: Firstly, the MPEG-4 SVC CODEC is tested in a stand-alone test, to document the impact of different encoder settings on the resulting quality and assess the CODEC's features. Secondly, the characteristics of the MPEG-4 SVC CODEC are compared to those of x264 and Xvid in a comparison test.

5.1 MPEG-4 SVC Stand-Alone Test

Quantization parameter test: During this test, the impact of the quantization parameter (QP) on the video quality is evaluated. The higher the QP value, the stronger is the quantization of the sequence and the lower is the resulting quality. The QP can either be a constant integer or - using rate control - automatically dynamically adjusted to match a selected bitrate. For the evaluation, the 'Foreman' (CIF, 30 Hz), 'Crew' (4CIF, 60 Hz) and 'Pedestrian Area' (720p, 25 Hz) sequences are each encoded with a single layer and constant QPs of 0, 10, 20, 30, 40 and 50. These sequences are used as they provide a wide range of different motion and spatial details. All other encoder settings are left at standard values.

CGS / MGS test: In the coarse grain scalability (CGS) / medium grain scalability (MGS) test, the impact of MGS on the video quality is assessed in

comparison with CGS coding. To do so, the three sequences 'Foreman', 'Crew' and 'Pedestrian Area' are encoded with two layers. In CGS mode, only these two layers - using SNR scalability - can be extracted, while the sequence encoded with MGS additionally offers 4×4 MGS vectors to dynamically adjust to changing bandwidth needs. Except for the two layers, the standard encoding settings are employed. During the test, three different bitrates are compared.

Best extraction path test: The different video streams of a SVC bitstream are arranged in a spatio-temporal cube. The best extraction path test is conducted to determine which of the video streams is perceived as the optimal one for a given bitrate in terms of visual quality. To achieve this, the unimpaired original 4CIF sequences are encoded in three spatial (QCIF, CIF, 4CIF) and four temporal (7.5 Hz, 15 Hz, 30 Hz, 60 Hz) resolutions. The QP of each layer is adjusted to match the target filesize of 1000 KB. The outcome of the best extraction path test shows which of the three kinds of impairments (spatial, temporal or SNR) has the biggest impact on perceived quality and, as a result, if there is an extraction path which can generally be recommended.

5.2 Comparison of MPEG-4 SVC to MPEG-4 AVC/ASP

Quality comparison test: During the quality comparison test, nine test sequences are encoded with the three evaluated CODECs Xvid, x264 and SVC. The CIF sequences are encoded with 200 kbps, the 4CIF and HD sequences with 1000 kbps. In the subjective assessment, the experts are then asked to evaluate the sequences: In each test, the subject is first shown the uncompressed reference sequence. After that, the three impaired versions of the same sequence compressed with the three evaluated CODECs are compared to the original. During the objective evaluation, the three impaired sequences of each sequence are compared to the original.

Encoding speed test: In the encoding speed test, the time of each CODEC to encode a given sequence is measured. For this evaluation the standard encoder settings are employed. For the encoding process, three sequences ('Foreman', 'Crew' and 'Pedestrian Area') with different resolutions and a duration of 10 seconds each are used. The sequences are looped 3 times before the encoding process to reduce measuring inaccuracies.

6 Results

6.1 MPEG-4 SVC Stand-alone Test

First, the results from different tests regarding the SVC options are compared. It has to be mentioned that some tests could only be performed using objective metrics as the differences in quality are too small to be evaluated subjectively.

Quantization parameter test: After normalizing the both PSNR and ITU-R quality mark, the objective and subjective quality scores differ significantly. While the objective score degrades almost linearly with the rising QP value, the subjective score shows very little quality impairment up to a QP value of 30, but then quickly falls to a relative score of about 25% at QP 40. Apparently a certain amount of loss in high frequency information does not impair perceived quality much, but of course this loss is already picked up by the PSNR calculation.

CGS / MGS test: The CGS / MGS test showed similar results in both objective and subjective evaluation. At bitrates between the two SNR layers, MGS encoding can lead to a significant increase in quality. As the objective tests showed, the quality level assigner tool can be used to achieve an almost linear PSNR increase with a low number of MGS vectors.

Best extraction path test: While the results of the objective best extraction path assessment showed the best PSNR values for sequences encoded in 4CIF resolution and 30 / 60 Hz, in subjective testing, in contrast, especially the bitstream using the highest possible spatial and temporal level is rated very poor. This finding matches with the ones previously mentioned in the quantization parameter test, where the subjective quality ratings suddenly drops between QP 30 and 40, whereas the objective scores scaled almost linearly throughout the whole QP range. In the following figures, the numbers from 1 to 12 indicate the visual quality of each selectable bitstream, where 1 is the best and 12 the worst rating.

Apart from that, it is additionally visible that QCIF resolution, as well as all streams encoded with 7.5 Hz framerate received very low scores in both test runs. As a result, the selection of the lowest spatial and/or temporal resolution should be avoided if possible.

Fig. 1. Objective and subjective quality marks for different framerates and resolutions

6.2 Comparison of MPEG-4 SVC to MPEG-4 AVC/ASP

Quality comparison test: When looking at the quality comparison test, basically similar results could be observed in both subjective and objective testing.

Fig. 2. Objective and subjective quality results of the quality comparison test

The overall visual quality of the three tested CODECs in the evaluated scenarios leads to the following ranking:

Under the described test conditions, the Xvid CODEC scores the lowest, which is most likely due to its MPEG-4 ASP base. The visual quality of x264 and SVC are nearly on par, which is expected as AVC is the direct predecessor of SVC. During the quality comparison, a particular flaw in the SVC CODEC became apparent: The rate control. Even though the requested bitrate is delivered in most cases quite accurately, the resulting quality can be unstable under certain conditions. While the maximum fluctuation amplitude of x264 is about 5 dB, the SVC CODEC reaches about 10 dB. Another significant flaw in SVC rate control is that in certain short sequences, the CODEC tends to distribute too much bitrate at the beginning of the sequence. This is followed by an excessive increase of quantization at the end of the file, leading to a significant quality decrease. It is however noteworthy that this behavior did not occur in every sequence.

Encoding speed test: The encoding time is measured on two different test systems to evaluate the impact of different CPU speeds and capabilities on SVC encoding. The details of both test systems are listed in table 2.

The following tables show the detailed results for both test systems. Both the absolute times and the relative speedup with System 2 as reference are given.

Table 2. Hardware configurations of the test systems

	System 1	System 2
OS	Microsoft Windows Vista Business 64-Bit, Version: 6.0.6001 SP1	Microsoft Windows Vista Business 32-Bit, Version: 6.0.6001 SP1
CPU	Intel® Core™ 2 Quad Q9450 4×2.66 GHz	Intel® Core™ 2 Duo T5500 2×1.66 GHz @ 1.00 GHz
RAM	4096 MB DDR2 800	2048 MB DDR2 667
HDD	Samsung Spinpoint T166, 320 GB, 7200 RPM, 16 MB Cache	Hitachi Travelstar 5K100, 100 GB, 5400 RPM, 8MB Cache

Table 3. Average encoding time on different computer systems in seconds

	CIF			4CIF			HD		
	Xvid	x264	SVC	Xvid	x264	SVC	Xvid	x264	SVC
System 1	1.1	0.9	387.7	12.1	7.8	2778.5	15.3	7.7	3902.1
System 2	4.1	7.2	947.8	42.1	60.8	8155.4	41.6	60.1	9575.8

As table 3 shows, there are significant differences in speedup between the different CODECs. SVC just seems to profit from the higher core clock of system 1, as the speed scales linearly with the core clock ($\frac{1.00GHz}{2.66GHz} = 0.376$). Xvid speedup is slightly higher, maybe due to optimizations for the new SSE instruction sets implemented in the quadcore processors. The biggest speed gain can be observed using the x264 CODEC. This is because x264 is the only CODEC that supported multithreaded encoding at the time of testing, so the quadcore processor could be used to its full potential.

7 Current SVC Flaws

7.1 Improvement of Existing Features

While the new MPEG-4 SVC CODEC adds many useful features to its predecessor MPEG-4 AVC, some flaws could still be observed during the subjective as well as the objective evaluations. These are described in the next section.

More reasonable default configuration: Some parameters of the SVC configuration files are by default not reasonably adjusted. The most important is the value of 'BaseLayerMode', whose default value is '3', which is not even a defined setting. Although being allowed and defined, the value of '1' for the setting 'GOPSize' is also not reasonable, as it heavily cripples the amount of temporal scalability possible. Hence, a change of the default parameter to a value of '8' or '16' is purposed. Because the encoding speed of SVC is currently low, the default parameter '0' (= 'BlockSearch') of 'SearchMode' is also not considered to be reasonable and should be switched to '4' (= 'FastSearch').

Improve encoding speed: The previous test have shown that the current MPEG-4 SVC version has a much lower encoding speed than the other tested CODECs. Firstly, it needs to be mentioned again that this is to be expected, as SVC is still in development status, but two main reasons can be identified and are explained in the following.

Multithreading: The benefit of multithreading support becomes more and more visible in modern computer systems, because multicore configurations are already commonly found in private environments today. If a similar encoding speed gain as in x264 when using multithreading is proclaimed, the encoding speed would approximately be accelerated linearly with the number of available logical CPUs.

Motion estimation: To further decrease the encoding time needs, it would be essential to optimize the performance of the motion estimation algorithms. As already noted in [11], the currently employed motion estimation technique achieves the best quality possible. However, the computation complexity is very high, which obstructs it from practical use. [11] also proposes a fast mode decision algorithm for inter-frame coding as a solution, which achieves an average encoding time reduction of 53%.

Enhanced, stable rate control mechanism: As shown before, the SVC rate control feature still has minor flaws. Because the exact reasons for these behaviors could not be precisely pinpointed in the tests, no concrete proposal for improvement can be given here. Still, improvements in this area are regarded as necessary.

7.2 Additional Useful Features

In the next section, additional features, that are not implemented in the current SVC release, but would be useful, are described.

Variable, content-dependent framerate: As scalable video technology is especially advantageous in streaming media environments, a useful new technique would be content-aware variable framerate. The basic idea of variable content-dependent framerate is that a reduced temporal level does not impair scenes with no or very low movement, which was already proven by [6]. There could be two main positive results when reducing the framerate: Either the file size of the video sequence could be reduced, or - if the size remains constant - the SNR quality would benefit respectively.

2-Pass encoding mode: 2-pass encoding strategies have been implemented in most modern CODECs, for example Xvid or x264 which have been examined earlier. Implementing this feature into SVC would primarily benefit its suitability for archiving storage. Of course, the poor rate control of SVC would also benefit from the bitrate distribution algorithms in 2-pass mode. In spite of this fact, it is essential that single pass rate control of SVC is improved, as 2-pass encoding mode is not suited for realtime encoding.

7.3 Conclusion

The extensive tests conducted in this work show that the new scalable video coding extensions provide significant improvement in terms of adaptability of the video stream. Using the scalability features of SVC, both high quality and low bandwidth versions of a video stream can be delivered, while at the same time saving bitrate compared to the storage of separate videos. However, there are also several features that still need improvement. First and foremost, the encoding speed of the SVC reference encoder is far too slow. Two methods to speed up the encoding are already proposed before. Additionally, several optimizations and other new useful features are proposed in the previous section. Concluding, SVC is a promising new extension to the MPEG CODEC family.

References

1. Bach, M.: Freiburg Visual Acuity, Contrast & Vernier Test ('FrACT') (2002), <http://www.michaelbach.de/fract/index.html>
2. CS MSU Graphics & Media Lab Video Group. MOS Codecs Comparison (January 2006)
3. CS MSU Graphics & Media Lab Video Group. Video MPEG-4 AVC/H.264 Codecs Comparison (December 2007)
4. Rohaly, A.M., et al.: Video quality experts group: current results and future directions. In: Ngan, K.N., Sikora, T., Sun, M.-T. (eds.) Visual Communications and Image Processing 2000. Proceedings of SPIE, vol. 4067, pp. 742–753. SPIE (2000)
5. Barzilai, M.A.J., et al.: Subjective quality analysis of bit rate exchange between temporal and SNR scalability in the MPEG4 SVC extension. In: International Conference on Image Processing, pp. II: 285–288 (2007)
6. Feghali, R., Wang, D., Speranza, F., Vincent, A.: Quality metric for video sequences with temporal scalability. In: International Conference on Image Processing, pp. III: 137–140 (2005)
7. I.O. for Standardisation. Svc verification test report. iso/iec jtc 1/sc 29/wg 11 n9577 (2007)
8. Institut für Rundfunktechnik. ITU-R BT.500 Recommendation and SAMVIQ, ITU-R BT.700 (2005)
9. Rabin, J.: (Visual Function Laboratory Ophthalmology Branch / USAF School of Aerospace Medicine). Color vision fundamentals (1998)
10. Kozamernik, F., Steinman, V., Sunna, P., Wyckens, E.: SAMVIQ - A New EBU Methodology for Video Quality Evaluations in Multimedia, Amsterdam (2004)
11. Li, H., Li, Z.G., Wen, C.: Fast mode decision algorithm for inter-frame coding in fully scalable video coding. IEEE Trans. Circuits and Systems for Video Technology 16(7), 889–895 (2006)
12. W. Media. Pluge Test Pattern, <http://www.mediacollege.com/video/test-patterns/pluge.html>
13. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Processing 13(4), 600–612 (2004)
14. Wien, M., Schwarz, H., Oelbaum, T.: Performance analysis of SVC. IEEE Trans. Circuits and Systems for Video Technology 17(9), 1194–1203 (2007)