Optimized Temporal Scalability for H.264 based Codecs and its Applications to Video Conferencing

Hans L. Cycon[*], Detlev Marpe[†], Thomas C. Schmidt[‡], Matthias Wählisch[§], and Martin Winken[†]
[*]HTW Berlin, Wilhelminenhofstr. 75, 12459 Berlin, Germany
[†]Fraunhofer Heinrich-Hertz-Institut Berlin, Image Processing Dept., Einsteinufer 37, 10587 Berlin, Germany
[‡]HAW Hamburg, Dept. Informatik, Berliner Tor 7, 20099 Hamburg, Germany
[§]Freie Universität Berlin, Institut für Informatik, Takustr. 9, 14195 Berlin, Germany
h.cycon@htw-berlin.de, {marpe,martin.winken}@hhi.fraunhofer.de, {t.schmidt,waehlisch}@ieee.org

***Abstract*—In this paper, we describe and analyze a low complexity scalable video codec that extends our H.264-implementation DAVC. We can show that DSVC, our temporally scaled codec, attains an RD performance identical to the non-scaled version at comparable configuration. We achieve this by QP cascading, a method of assigning gradual refining quantization parameters to the declining temporal layers. The different quantization of frames does not lead to visual distinguishable quality fluctuations. This video codec is the core component of a software-based multipoint videoconference system, which works without MCU on a hybrid P2P network structure.**

***Index Terms*—Video coding, temporal scalabilty, SVC, video applications, video conferencing**

I. Introduction

Video applications in the Internet exhibit significant growth in several market segment. Conversational systems of video conferencing and immersive telepresence, video-assisted games, or high-definition IPTV broadcasting are more and more enabled by flexible and powerful nodes connected to the Internet at high speed, but also by recent advances in video coding and processing technologies. In addition, the number of devices capable of displaying moving images at reasonable quality is rapidly growing due to popular consumer devices such as smartphones and game boxes.

Video data in online applications need to be compressed and decoded with high-performance video codecs and high bit rate flexibility. The most efficient codec in sense of rate distortion (RD) performance has been defined in the H.264/AVC video coding standard [1], [2]. To work efficiently in such heterogeneous environments described above, suitable video codecs need to have some extended scalability properties in addition to high (RD) performance. Scalability in this context refers to the removal of parts of the video bit stream in order to adapt it to the varying terminal capabilities or network conditions [3]. To meet these requirements, the scalable successor SVC of H.264/AVC video coding standard was defined in 2007 [4]. SVC enables the transmission and decoding of partial bit streams to provide video streams with temporal, spatial and quality scalability. Full scalability comes at the price of increased complexity and also some bit rate increase for the same fidelity compared to single layer coding [3]. Note that traditional video transmission systems also had some scalability features, but they came along with a significant loss of coding efficiency, as well as a large increase in decoder complexity as compared to the non scalable versions [5], [6], [7].

In this paper, we describe and analyze a scalable extension called DSVC of our H.264-implementation DAVC [8] which permits temporal scalability. We can show that our temporally scaled video codec (DSVC) has the same RD performance as the non-scaled version with comparable configuration. We achieve this by QP cascading, i.e. assigning to the declining temporal layers gradual refining quantization parameters. The different quantization of frames does not lead to visual distinguishable quality fluctuations. Furthermore, DSVC extensions fully preserve the excellent real-time capabilities of DAVC.

The remainder of this paper is organized as follows. In section II we briefly review basics on temporal scalability and introduce our codec extensions. Its experimental analysis and performance results are presented in section III. In the following section IV we discuss our application to video conferencing in heterogeneous networks, and in the final section V, we conclude with a summary and an outlook.

II. Temporal Scalability

Temporal scalability describes cases in which subsets of the bit stream represent the source content with a reduced frame rate (temporal resolution) [3]. A sequence of temporal layers (TLs) consists of the base layer and temporal enhancement layers. A bit stream obtained by a complete sequence of temporal layers starting from base layer to a suitable enhancement layer forms a valid bit stream for the given decoder. Obviously, if the number of enhancement layers is increased then the bit rate and the frame rate of the video stream also increases.

Reference pictures form the basis for uni- or bidirectional prediction of the enhancement layer pictures. In principle H.264/AVC allows the coding of picture sequences with arbitrary temporal dependencies. We consider here only simple hierarchical prediction structures, where reference pictures are always temporal preceding the enhancement layer pictures as shown in Figure 1. Following a hierarchical precedence scheme implies that we only use unidirectional predictions in

This work is supported by the German Ministry of Research and Education within the projects Moviecast and HAMcast (see http://www.realmv6.org).

978-1-4244-6673-3/10/$26.00 ©2010 IEEE

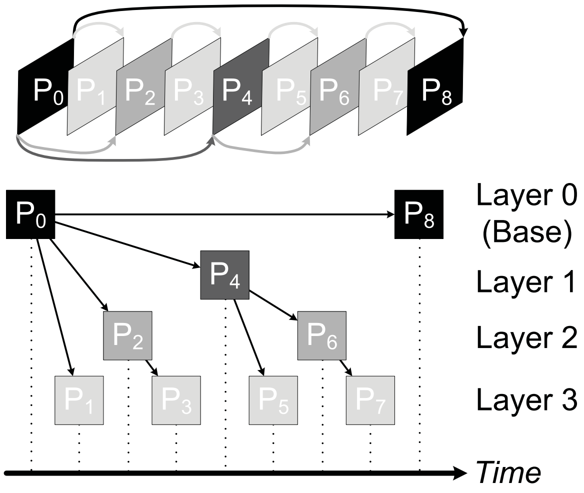
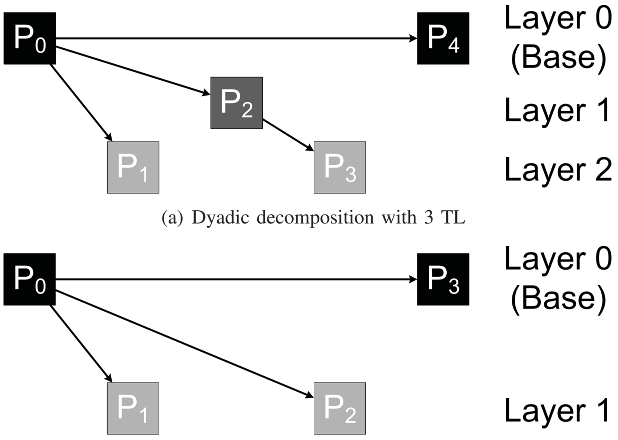


Fig. 1. Unidirectional dyadic hierarchical prediction structure with 4 temporal layers

our implementation. This provides zero structural delay. In general, such low-delay structures decrease coding efficiency. For a general discussion we refer to [3].

a) Implementation of Temporal Scalability: The proposed DSVC codec provides up to two temporal enhancement layers (TLs). For the configuration of three layers, we implemented several decomposition strategies, which implicate different bit rates caused by each layer. To achieve a higher weight at the base layer, we use a dyadic prediction decomposition. The ratio of the frame rate regarding the base layer, the 1st, and 2nd enhancement layer is 1/4, 1/4, and 1/2 respectively. The base layer covers approx. 63 % and the 1st enhancement layer approx. 17 % of the overall bit rate. To allow for a lower weight of the base layer, we implemented two variants by a combination of both a dyadic (cf., Figure 2(a)) and non-dyadic (cf., Figure 2(b)) layer decomposition. The first variant changes the frame rate by a factor of 1/3 at the transition from the second to the first enhancement layer, and by a factor of 1/2 at the transition from the first enhancement to the base layer. In this case, the base layer accounts only for 50 % of the overlay bit rate, and the first enhancement layer attains 15%. The second variant realizes a more balanced bit rate between the first and second enhancement layer.



(a) Dyadic decomposition with 3 TL

(b) Non-dyadic decomposition with 2 TL

Fig. 2. A dyadic and non-dyadic of temporal layers (TL)

Layer	Bit rate 0	Bit rate 1	Bit rate 2	Frame rate 0	Frame rate 1	Frame rate 2
3 TL	50 %	15 %	35 %	1/6	1/6	2/3
	52 %	27 %	21 %	1/6	1/3	1/2
	63 %	17 %	20 %	1/4	1/4	1/2
2 TL	70 %	30 %	—	1/3	2/3	—

TABLE I
CONFIGURATIONS OF DSVC – RELATIVE BIT RATES AND CORRESPONDING FRAME RATES AT DIFFERENT LAYERING

In the case of one enhancement layer, we implemented a non-dyadic composition with two non-referenced P–frames at the topmost time level. The frame rate of the base layer compared to the enhancement layer is 1/3. The base layer contains approx. 70 % of the overall bit rate. For details of the temporal decomposition and resulting bit and frame rates, we refer to Table I.

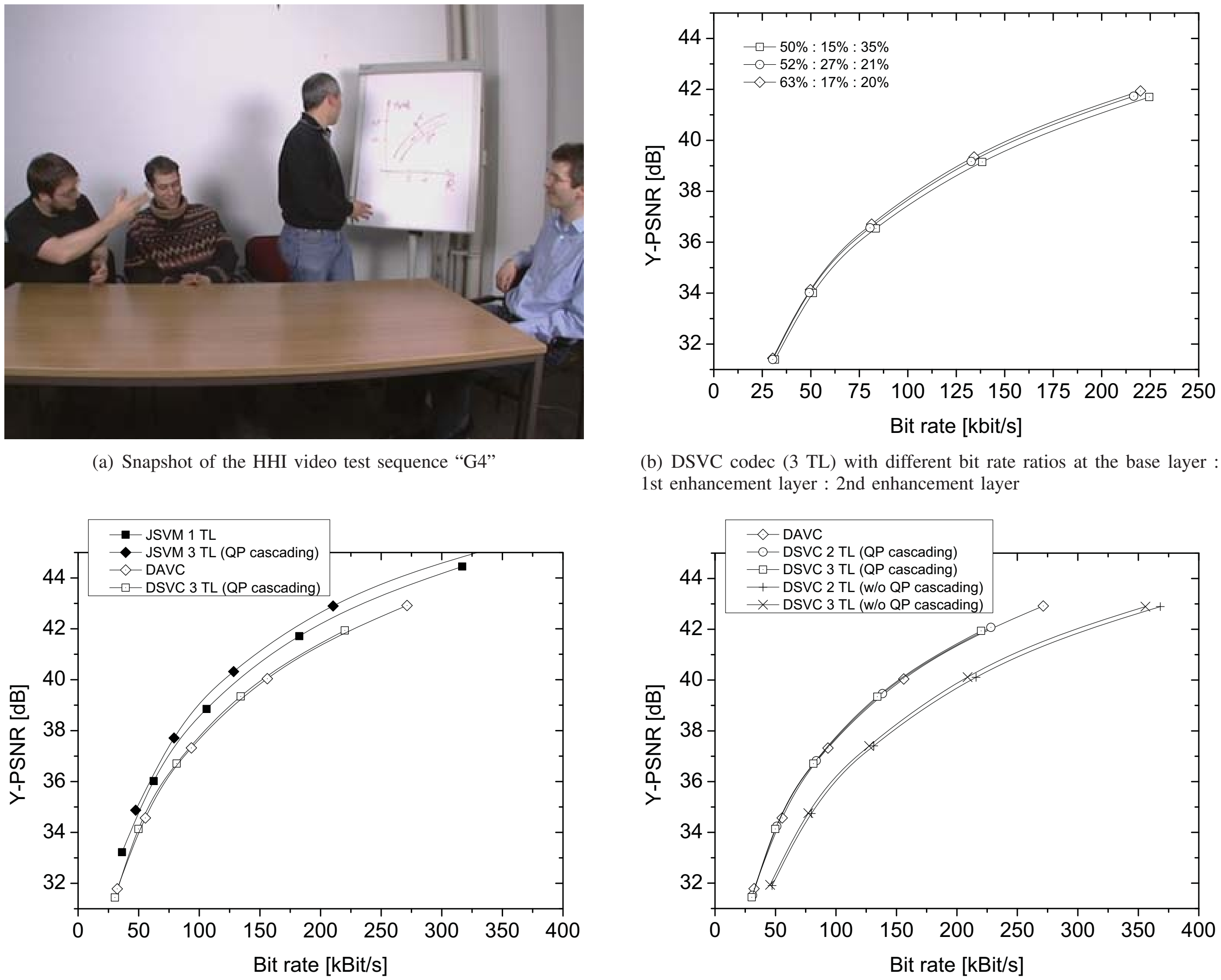
b) Quantization in Layers: The coding efficiency for hierarchical prediction structures are based on the amount of quantization per temporal layer This will be configured by the quantization parameter QP. Frames of the temporal base layer should be coded with highest fidelity, since they are used as references for all temporal enhancement layers. Consequently, a larger quantization parameter should be chosen for subsequent temporal layers as the quality of these frames influences fewer pictures [3]. A gradual quantization depending on the layer is called QP cascading. For the 3 TL we have chosen the following strategy for QP cascading (cf. [9]): Based on a given quantization parameter QP_0 for pictures of the temporal base layer, the quantization parameter QP_T for pictures of a given temporal layer with an identifier $T > 0$ is determined by $QP_T = QP_0 + 3 + T$. This algorithm has proofed empirically to give good RD results [3], [9]. In case of 2 TL we chose a simple QP cascading of the form $QP_1 = QP_0 + 5$.

III. EXPERIMENTAL RESULTS

In this section, we analyze the encoding quality of our DSVC codec.

For reproducibility, we use as input data the HHI video test sequence “G4” (cf., snapshot Figure 3(a)) in 384 x 288 resolution at a frame rate of 30 Hz. Experiments have been conducted for other test sequences, which achieve similar results. When configured with a single temporal layer, the DSVC corresponds to our DAVC [8] codec.

We evaluate the quality of our SVC codec by measuring the peak signal-to-noise ratio (PSNR) depending on different bit rates. This quantifies the pure encoding quality, i.e., the distortion of the compressed stream in contrast to the original



(a) Snapshot of the HHI video test sequence "G4"

(b) DSVC codec (3 TL) with different bit rate ratios at the base layer : 1st enhancement layer : 2nd enhancement layer

(c) Comparing reference codec JSVM with DAVC/DSVC

(d) Coding efficiency of DSVC for different temporal layers and QP cascading configurations

Fig. 3. Encoding quality for the test sequence "G4" in 384x288 resolution at at frame rate of 30 Hz.

data without including network disturbances or layer adaptation. The rate distortion (RD) is analyzed for different layer configurations, which reflect the number of temporal layers (TL) used for encoding, effects of QP cascading, and variable bit rates per layer.

DSVC is compared with the SVC reference software Joint Scalable Video Model (JSVM) version 9.16 [10]. It is worth noting that the JSVM encoder is designed for complete RD characteristics, but has no real-time abilities in contrast to the DSVC.

We compare the coding efficiency for the 2 TL and 3 TL case with and without QP cascading versus the single layer DAVC version. The results for the first case show for 3 layer DSVC with QP cascading similar/better RD performance than the single layer DAVC (see Fig. 3(b)) . Compared with the reference implementation JSVM_9-16 we loose only 1-1.5 dB PSNR in average as shown in Figure 3(d). Note that there is a considerable loss of RD performance if higher layering is used without QP cascading (see Figure 3(d)). In summary, it can be observed that signaling overheads due to temporal layering are fully compensated by QP cascading in our DSVC implementation.

Table II shows similar results for different resolutions and additional sequences, i.e., there is no loss of RD performance in 3TL case if we use QP cascading in contrast to uniformly assigned QP parameters for the frame levels (no QP cascading).

We did observe in all measurements that in spite of the relatively large jumps in the quantization parameters (QP) and PSNR qualities for the different frames in a sequence PSNR, the reconstructed video appears temporally smooth and does not show visually distinguishable fluctuations in quality. Similar effects have been already reported in different configurations [3].

Sequence	Resolution@frame rate	PSNR [dB] 1 TL	PSNR [dB] 3 TL	PSNR [dB] 3 TL no QP Cascading
G4	768x576@30Hz	38,2	38,2	36,9
KO	768x576@30Hz	38,1	38,1	36,8
SM	768x576@30Hz	42,6	42,8	41,7
TC	768x576@30Hz	41	41	39,7
TW	768x576@30Hz	42	42	41

TABLE II
ADDITIONAL RD MEASUREMENTS AT BIT RATE OF 400 KBIT/S AND 30 HZ FRAME RATE (INTERPOLATED)

The real-time compliance of the DSVC codec can be measured by evaluating the maximum number of frames that can be encoded per time unit on a target hardware. The single layer version of the DSVC codec achieves up to 284 frames per second on a standard desktop PC. It slightly outperforms comparable H.264 codecs [8]. Compared to a single layer encoded stream, the hierarchical prediction structure only changes the pointer to the reference frame. Parameters for the motion prediction (in particular the search range) remain unchanged. Temporal scalability thus does not introduce additional overhead and does not decrease the run time performance.

IV. THE VIDEO CONFERENCING SYSTEM

Our target application is a software-based video conferencing solution running on ordinary PC hardware or mobile devices which requires low complexity scalable video codecs. The proposed DSVC technology is implemented as the core component of the video conference system daViKo. It is built upon DAVC described in [8], which is a fast constraint base line H.264/AVC video codec, optimized for real time encoding and decoding. The codec along with the H.264/AVC design also includes context-adaptive mechanisms to recover quickly from video packet loss. The Internet conferencing tool works without MCU server on a hybrid P2P network structure [11], and seamlessly complies with different Internet Protocol versions, as well as the conference management signaling of SIP [12] and H.323 legacy MCUs [13]. It is designed for heterogeneous network conditions and components, where the scalability of the codec enables dynamic adaption the data stream to the available capacity at network and receiving side. The adaption of the data stream is controlled by an adaption layer. Parameters like packet loss, inter arrival jitter or round trip time generated by receivers as sender feedback, e.g., as foreseen by the RTCP protocol are used to give suitable information on the network status. Details on this will be given in a forthcoming report.

This highly portable conferencing software is professionally available for desktop computers running MS-Windows or Linux, on handhelds equipped with the Windows Mobile operating system, and on the Apple iPhone.

V. CONCLUSIONS AND OUTLOOK

Video communication in real-world heterogeneous environments needs significant scaling abilities to flexibly adapt to various network and host conditions. In this paper we presented a corresponding conferencing application built upon a fast and efficient temporally scalable video codec called DSVC. We could show that that our temporally scaled video codec attains the same RD performance as its non-scaled version with comparable configuration. This promising result has been achieved by QP cascading, an adaptive quality scaling between layers. The different quantization of frames does not lead to visual distinguishable quality fluctuations. Further work will proceed into two different directions. At first, we will extend experimental analysis and optimizations of the network adaptation to maximize video performance. Second we will extend the scalability of our codec in particular by including spatial scaling options as offered by the SVC standard.

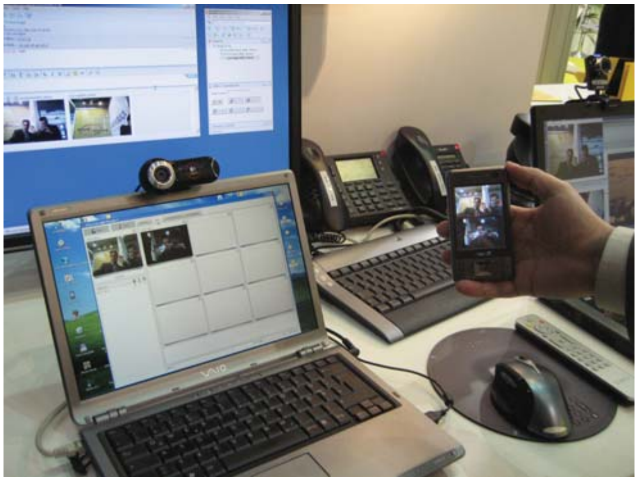


Fig. 4. The heterogeneous daViKo video conferencing system

ACKNOWLEDGMENTS

We would like to thank the following collaborators for fruitful discussions and implementation work (in alphabetical order): Valerie George, Gabriel Hege, and Mark Palkow.

REFERENCES

[1] "Advanced Video Coding for Generic Audiovisual Services," ITU-T, Tech. Rep. Recommendation H.264 & ISO/IEC 14496-10 AVC, v3, 2005.

[2] J. Ostermann, J. Bormans, P. List, D. Marpe, N. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video Coding with H.264/AVC: Tools, Performance and Complexity," *IEEE Circuits and Systems Magazine*, vol. 4, no. 1, pp. 7–28, April 2004.

[3] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, September 2007.

[4] “Advanced Video Coding for Generic Audiovisual Services, Annex G,” ITU-T, Tech. Rep. Recommendation H.264 & ISO/IEC 14496-10 AVC/ Amd.3 Scalable Video Coding, November 2007.

[5] “Generic Coding of Moving Pictures and Associated Audio Information – Part 2: Videos,” ITU-T, Tech. Rep. ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2 Video), ITU-T and ISO/IEC JTC 1, 1994.

[6] “Video Coding for Low Bit Rate communication,” ITU-T, Tech. Rep. ITU-T Rec. H.263, ITU-T, Version 1, 2, 3, 1995 (1), 1998 (2), 2000 (3).

[7] “Coding of audio-visual objects – Part 2: Visual,” ISO/IEC, Tech. Rep. ISO/IEC 14492-2 (MPEG-4 Visual), ISO/IEC JTC 1, Version 1, 2, 3, 1999 (1), 2000 (2), 2004 (3).

[8] H. L. Cycon, T. C. Schmidt, G. Hege, M. Wählisch, and M. Palkow, “An optimized H.264-based Video Conferencing Software for Mobile Devices,” in *ISCE2008 – The 12th IEEE International Symposium on Consumer Electronics*, A. Navarro, Ed., IEEE. Piscataway, NJ, USA: IEEE Press, April 2008, pp. 1–4. [Online]. Available: http://dx.doi.org/10.1109/ISCE.2008.4559439

[9] J. Reichel, H. Schwarz, and M. Wien (Eds.), “Joint Scalable Video Model 11 (JSVM 11),” ISO/IEC MPEG & ITU-T VCEG, Doc. JVT-X202, July 2007.

[10] “SVC Reference Software,” http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm, 2010.

[11] T. C. Schmidt, G. Hege, M. Wählisch, H. L. Cycon, M. Palkow, and D. Marpe, “Distributed SIP Conference Management with Autonomously Authenticated Sources and its Application to an H.264 Videoconferencing Software for Mobiles,” *Multimedia Tools and Applications*, 2010, to appear. [Online]. Available: http://dx.doi.org/10.1007/s11042-010-0500-8

[12] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, “SIP: Session Initiation Protocol,” IETF, RFC 3261, June 2002.

[13] H. L. Cycon, G. Hege, D. Marpe, M. Palkow, T. C. Schmidt, and M. Wählisch, “Connecting the Worlds: Multipoint Videoconferencing Integrating H.323 and IPv4, SIP and IPv6 with Autonomous Sender Authentication,” in *ISCE2009 — 13th International Symposium on Consumer Electronics*. Piscataway, NJ, USA: IEEE Press, May 2009, pp. 890–893.