This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2010 proceedings.

Adaptive Temporal Scalability of H.264-compliant Video Conferencing in Heterogeneous Mobile Environments

Hans L. Cycon*, Valeri George†, Gabriel Hege‡, Detlev Marpe†,
Mark Palkow§, Thomas C. Schmidt‡, Matthias Wählisch¶
*HTW Berlin, †HHI Berlin, ‡HAW Hamburg, §Daviko GmbH, ¶FU Berlin
hcycon@htw-berlin.de, hege@fhtw-berlin.de, marpe@hhi.de, palkow@daviko.com, {t.schmidt, waehlisch}@ieee.org

***Abstract*—In this paper, we present a multipoint video conferencing system that adapts to heterogeneous members including mobiles. The system is built upon a low complexity scalable extension of our H.264 codec DAVC, and a congestion-aware dynamic adaptation layer. 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 gradually refining quantization parameters to the declining temporal layers. We present and analyse a mobile-compliant version of DSVC at reduced complexity that still admits comparable performance. Finally, we report on early work of dynamic layer tuning. Derived of delay variation measures, senders exploit scalable video layering to adapt the video transmission to varying network conditions. Initial results indicate that video performance remains close to optimal.**

I. Introduction

Recent advances in video coding, network and computing technology are enabling more efficient video distribution in the Internet. IP-based video systems today range from video telephony to high-definition telepresence or IPTV broadcasting. This variety of video applications are transmitted via networks with a wide range of capabilities, and displayed by largely heterogeneous receivers. Prior to transmission, video data need to be compressed and decoded with high-performance real-time video codecs that admit high flexibility in bit rate. 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, 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 generate video flows with temporal, spatial, and quality scalability. The fully implemented SVC, however, comes with some increases of complexity and bit rate for the same fidelity as compared to single layer coding.

In this paper we present a real-world video conferencing system built upon a scalable extension of our H.264-implementation DAVC [5]. 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. This codec is also part of our conferencing client for mobile devices, where it operates at reduced complexity with the specific need for dynamic scaling. Network adaptive dynamic scaling in our system is shown to solely work under the control of senders based on generally available delay variation measures. We discuss initial measurement results of this ongoing work on self-adaptive scaling.

The remainder of this paper is organized as follows. In Section 2, we explore the problem space of scalable adaptive conferencing and review related work. Section 3 is dedicated to a detailed analysis of our scalable codec. Network adaptive video scaling is discussed in Section 4. Finally, in Section 5 we conclude with an outlook.

II. Problem Statement and Related Work

A. Heterogeneous Video Conferencing with Mobiles

Video conferencing is a highly conversational application and thus bound to rigid real-time constraints. Any component involved in preparing, transmitting, or displaying the audio-visual streams must be carefully controlled to not exceed performance limits. Such limits are threatened by resource exhaustion of processing capacities at end systems, as well as by network overloads. Conferencing demands are in contrast to those of streaming applications, even when transmitted in real-time, where many performance violations may be compensated by play-out buffers, elastic timing, etc.

In heterogeneous settings where capabilities are unevenly distributed among end-system and network connectivity is under-provisioned or temporally degraded, video conferencing performance at weak end points may easily drop down to an alienating experience for users. Audio-visual flows may stall or even come to a complete halt, whenever frames cannot be delivered in time for play-out. Conventional systems must adopt their resource demands (and thus quality) to comply with the lowest capacities available in a conference, or require a transcoding service, e.g., by an MCU, to reduce individual streams.

Challenges tighten when mobiles join a heterogeneous conference. For handhelds, bandwidth as well as processing and battery capacities commonly remain at least one order of magnitude below those of fixed systems. Using a highly optimized H.264 software codec, a mobile smartphone can

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

reliably and simultaneously encode and decode a QCIF video at 10 to 15 fps [5], resulting in data rates that comply to 3GPP offers. Thus, special treatment must be added to include mobile end systems into a conference of conventional quality at 30 fps of 384 x 288 resolution.

B. Scalable Coding

Scalable coding of video flows is a promising approach of adapting data rates to capacities in heterogeneous and mobile environments [6]. In general, a video bit stream is called scalable when parts of it can be removed in a way that the resulting sub-stream forms another valid bit sequence for some target decoder. The sub-stream represents the source content with a reconstruction quality that is less than that of the complete original data. Bit streams that do not provide this property are referred to as single-layer. The usual modes of scalability are temporal, spatial, and quality scalability. Current standard H.264/AVC-decoders do not support spatial scalability, as this requires support of dedicated extension as defined in the Scalable Video Codec (SVC) amendment [4], while temporal and quality scaling can be achieved in compatibility to the widely deployed H.264/AVC [1] standard players.

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 consists of the base layer and temporal enhancement layers. Any bit stream obtained by a complete sequence of temporal layers starting from base layer to a suitable enhancement layer forms a valid input for the given decoder. Obviously, if the number of enhancement layers is enlarged, the bit rate and the frame rate of the video stream also increases.

Reference pictures form the basis for uni- or bidirectional predictions at the enhancement layer pictures. Conceptually, H.264/AVC allows for coding of picture sequences with arbitrary temporal dependencies. Following our real-time objective in conferencing, we consider only hierarchical prediction structures, here, where reference pictures are always temporally preceding the enhancement layer pictures (see Fig. 1). This implies to adhere to unidirectional predictions, which cause zero structural delay. In general, such low-delay structures decrease coding efficiency. For a general discussion we refer to [3].

C. Adaptive Video Distribution

With the ability to scale video communication, a conferencing application may adapt streams to individual participants without the burden of transcoding. Layers can be selected statically according to initial media negotiations, or dynamically in reaction to network and runtime conditions. Dynamic adaptation faces the problems of reliably detecting network conditions, enabling a sender reaction in time so that media data neither accumulate in buffers or drop, nor under-utilize the available transmission resources. Finally, adaptation rates and algorithms need to remain stable and resistant against oscillating states even when network capacities fluctuate.

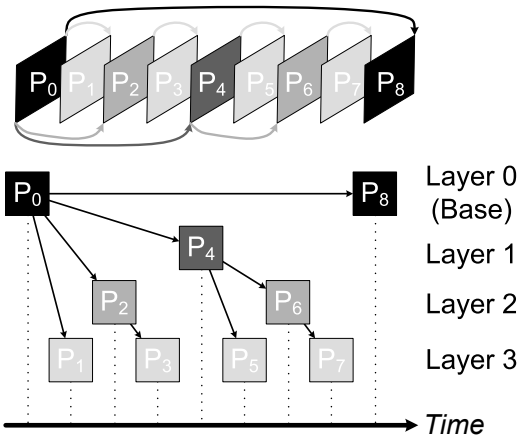


Fig. 1. Unidirectional dyadic hierarchical prediction structure with 4 temporal layers

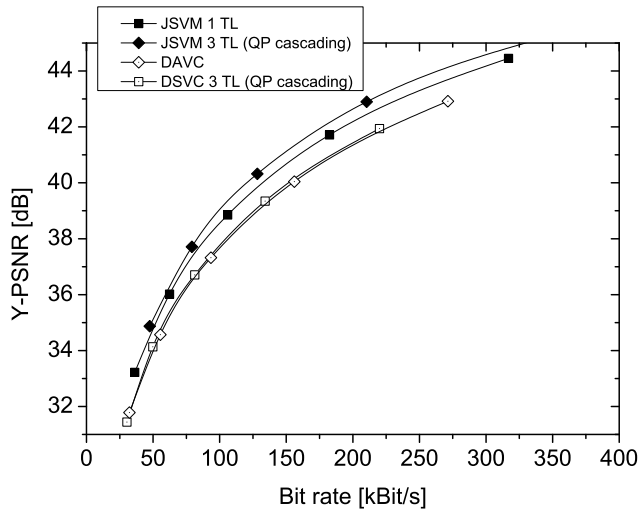
Previous work on adaptive layering has mainly focused on streaming scenarios. PALS [7] introduces a receiver-driven approach to a layered peer-to-peer video concast. Receivers monitor sender performance and layer-wise attempt to maximize throughput from multiple sources. Another method of selecting peers from heterogeneous neighbors according to their support of quality layers is presented in [8]. In a progressive quality adaption, the authors dynamically select temporal layers according to a continuously measured network throughput at receivers. Baccichet et al. [9] distribute SVC streams via multiple overlay multicast trees with layer-awareness, while adopting to heterogeneous uplink capacities and network conditions. Kofler et al. [10] introduce an RTSP/RTP proxy at WiFi routers to facilitate scalable video distribution to mobiles, while an authentication scheme for SVC videos that enables verification of all possible substreams is presented in [11]. In an early study on multipoint video conferencing, Eleftheriadis et al. [12] compared the performance of a traditional MCU with a corresponding server system based on the SVC reference implementation and could identify a significant reduction in delay and complexity for the SVC. Vidyo [13] provides stream adaptation by application-layer routers from an infrastructure perspective. To the best of our knowledge, there is no scalable adaptive solution for infrastructureless, peer-to-peer video conferencing systems.

III. SVC Implementation

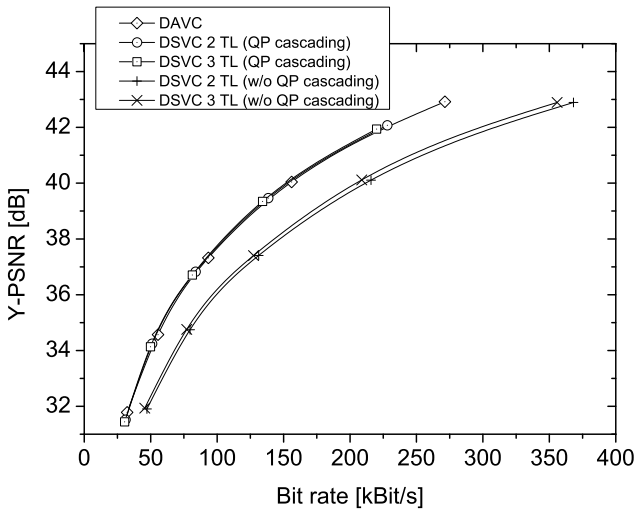
A. DSVC Codec Capabilities

Temporal Scalability. The DSVC codec provides up to three temporal layers (TL). Depending on the frame partitioning, different bit rates per layer are attained. For the configuration of two enhancement layers, we implemented several decomposition strategies. To allow for a reduced weight of the base layer, we introduced a combination of both a dyadic and non-dyadic layer decomposition. This reduces the data rate at the base layer down to 50 %. A strict dyadic prediction decomposition leads to a higher weight at the base layer, approx. 63 % of the overall bit rate in this case.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE



(a) Comparing reference codec JSVM with DAVC/DSVC



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

Fig. 2. Encoding quality for the test sequence "G4" in 384x288 resolution at at frame rate of 30 Hz.

	Bit rate			Frame rate		
Layer	0	1	2	0	1	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

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

Quantization in Layers. The coding efficiency for hierarchical prediction structures is 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*.

We have chosen the following strategy for QP cascading (cf., [15]): 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 $QP_T = QP_0 + 3 + T$.

Real-time Complexity. Real-time compliance of the codec can be measured by evaluating the maximum number of frames that can be encoded per second on 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 [5]. 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.

B. Evaluation of Encoding Quality

In this section, we analyze the encoding quality of our SVC codec. For reproducibility, we use as input data the HHI video test sequence "G4" 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 codec.

1) The Full DSVC 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 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. We disable the intra refresh option.

DSVC is compared with the SVC reference software Joint Scalable Video Model (JSVM) version 9.16 [16]. 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.

Comparison with Reference Encoder JSVM. The coding efficiency of our real-time SVC implementation is compared to

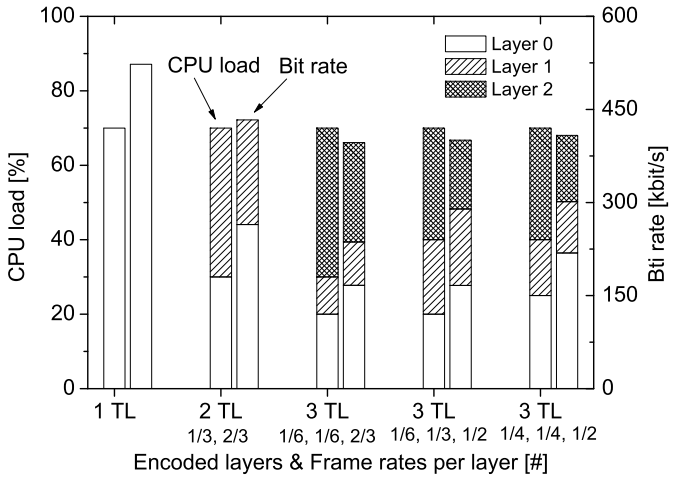


Fig. 3. CPU and bandwidth consumption at mobile device while receiving different layers of G4 384x288@15 Hz

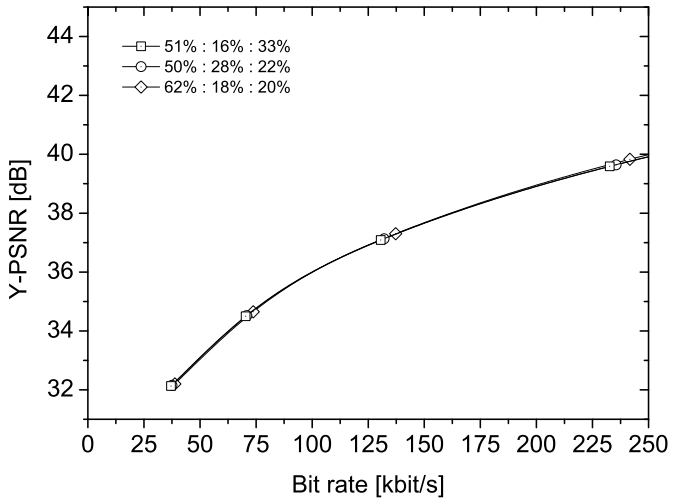


Fig. 4. Encoding quality for the Mobile-DSVC (3 TL) and G4 384x288@30 Hz at different bit ratios between layers

the reference encoder JSVM for one and three temporal layers in Figure 2(a). The DSVC and DAVC encoder achieve a RD performance similar to the reference encoder, with decreases only about 1 – 1.5 dB. It should be recalled, though, that JSVM is far from operating in real-time.

Effects of Layering & QP Cascading. Figure 2(b) shows the RD performance of the DSVC encoder for a varying number of available temporal layers and different QP options (i.e., a layer-dependent and layer-independent quantization of frames). A configuration with cascaded quantization outperforms an equal quantization of frames: (1) The overall RD performance reduces for multi-layered streams without QP cascading. (2) Applying gradual quantization per temporal layer yields an RD performance equal to the single layer stream. (3) Without QP cascading, the coding quality becomes dependent on the number of layers in use. Thus we can observe that signaling overheads due to temporal layering are fully compensated by QP cascading in our DSVC implementation.

2) DSVC on Mobiles: The generic DSVC codec has been ported and tuned to the Windows Mobile and iPhone OS platform. This included adaptation and optimization of the ANSI compliant C version to the wireless MMX instruction set for the mobile systems with target-specific code.

In order to enable real-time encoding performance, even when appropriately reducing the resolution of the input video to a 240 x 144 pixel format, the DSVC codec has been restricted to perform motion estimation only for integer-pel displacements. At this coder configuration, our test system Samsung Omnia II i8000 with 800 MHz SC36410 ARM11-Kernel could reliably encode 8 fps while simultaneously decoding 15 fps without CPU exhaustion or packet drop. Corresponding values for the iPhone were slightly lower at about 6 : 12. It should be noted, though, that in the absence of an open API support for video capturing, we had to dedicate noticeable resources of the iPhone to video image extraction of display buffers.

To quantify resource consumption of layered video processing in a standardized, reproducible experiment, we sent the G4 test sequence (384x288@15 Hz) to the Samsung mobile using different layers. Results obtained for the layer configurations in Table I are displayed in Figure 3. The measurements revealed a large scaling factor of 3.5 for processing consumptions, while sustained bandwidth scales down to about 40 %.

Figure 4 presents RD diagrams for the scaled-down mobile codec and the G4 test sequence in full 384 x 288 resolution. Frame rates have been chosen to distribute among layers as displayed in Table I for the full codec. Algorithmic downscaling, however, leads to a slight modification in bandwidth ratios as visible in the legend. Results show a moderate loss of 0 – 2 dB in rate-distortion performance relative to our full DSVC encoder. Thus, our mobile-based scalable video encoder produces acceptable video quality when conforming to the tight resource constraints of the mobility regime.

IV. Adaptation Layer

Heterogeneous, e.g., mobile clients can signal their system capacities within initial SDP negotiations, while network conditions may change at runtime. The objective of the adaptation layer is to achieve a dynamic scaling of the video transmission appropriate to network resources at the sender without explicitly involving receivers. Network congestions or overloads should be quickly detected and immediately answered by reduction in layers. After a reduced network load has been observed, the activation of layers should act tardy to avoid oscillating transmission rates and to maximize the user experience of continuous, uninterrupted play out.

The temporal scaling in our conferencing system is adjusted according to the available bandwidth between sender and receivers. Bandwidth estimation follows the general observation that the delay continuously increases when links start to become congested [17]. Identifying the transformation from an almost constant delay to a continuous growth can approximate the available bandwidth.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

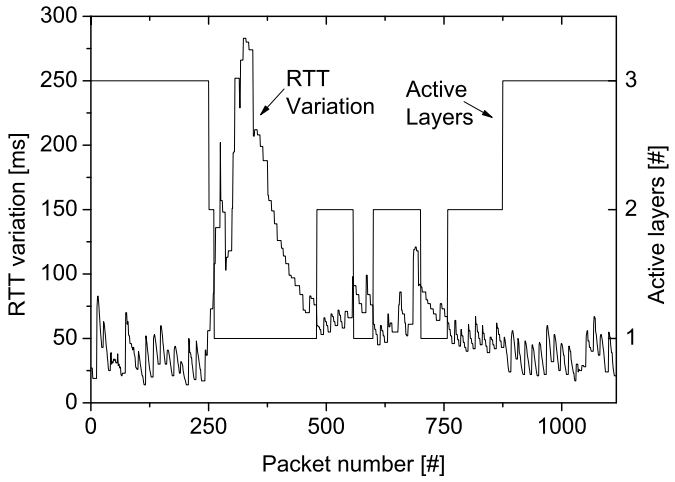


Fig. 5. Video layer adaptation stirred by changing jitter values

The layer adaptation of DSVC is inspired by SLoPS [18], which analyzes the one-way delay measured by sending a predefined number of equal-sized packets periodically. With respect to a minimal deployment support and lightweight mobility regimes, we make the following changes. To allow for an autonomous bandwidth estimation by senders without additional measurement traffic in the restricted wireless regimes, the DSVC layer adaption is based on the inter-arrival jitter. The inter-arrival jitter is commonly available, e.g., by RTCP. Its changes can likewise serve as an indicator for accumulating queuing delays at routers and access gateways. DSVC thus does not introduce additional packet overhead or requires functional updates at the receiver.

In detail, we observe the inter-arrival jitter packet wise. Whenever the jitter increases by at least a threshold during a short sequence of packets (e.g., 10), the adaptation reduces the video layers. Layers are added again to the transmission after a long sequence of packets (e.g., 50) is observed that continuously decreases delay variation. Parameters in this ongoing work are subject to current optimizations. To evaluate our adaption scheme, we set up a test network. The source transmits the "G4" test sequence to a receiver, while the available bandwidth between source and receiver is manually varied. First results are displayed in Figure 5, which show a smooth layer adaptation to jitter conditions in the test network.

V. Conclusions & Outlook

Video conferencing in real-world heterogeneous environments needs significant scaling abilities to flexibly adapt to various conditions. Even though the principle methods and tools are around, its realization in a ready-to-use system are still hard to achieve. In this paper, we presented a fast and efficient temporally scalable video codec implementation for mobile-based video conferencing, as well as an adaptation layer that dynamically selects appropriate frame rates. In an extensive analysis we could demonstrate the strength of our video solution, but also identify further needs for optimization.

Future work will proceed in 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.

Acknowledgements

Porting of our mobile video client to the Apple iPhone was provided by Fabian Jäger. This work has been supported by the German Ministry of Research and Education within the projects Moviecast and HAMcast (see http://www.realmv6.org).

References

[1] "Advanced Video Coding for Generic Audiovisual Services," ITU-T, Tech. Rep. Recomm. 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] H. L. Cycon, T. C. Schmidt, G. Hege, M. Wählisch, D. Marpe, and M. Palkow, "Peer-to-Peer Videoconferencing with H.264 Software Codec for Mobiles," in *WoWMoM08 – WS on Mobile Video Delivery (MoViD)*, IEEE Press, June 2008, pp. 1–6.

[6] T. Schierl, T. Stockhammer, and T. Wiegand, "Mobile Video Transmission Using Scalable Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1204–1217, Sept. 2007.

[7] R. Rejaie and A. Ortega, "PALS: Peer-to-Peer Adaptive Layered Streaming," in *Proc. 13th Intern. WS on Network and Operating Systems Support for Dig. Audio and Video*. ACM, 2003, pp. 153–161.

[8] O. Abboud, K. Pussep, A. Kovacevic, and R. Steinmetz, "Quality Adaptive Peer-to-Peer Streaming Using Scalable Video Coding," in *Proc. 12th IFIP/IEEE MMMS 2009*, ser. LNCS, no. 5842. Springer Verlag, 2009, pp. 41–54.

[9] P. Baccichet, T. Schierl, T. Wiegand, and B. Girod, "Low-Delay Peer-to-Peer Streaming using Scalable Video Coding," in *Proc. Intern. Packet Video Workshop (PV2007)*. IEEE, 2007, pp. 173–181.

[10] I. Kofler, M. Prangl, R. Kuschnig, and H. Hellwagner, "An H.264/SVC-based adaptation proxy on a WiFi router," in *Proc. 18th Intern. Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'08)*. ACM, 2008, pp. 63–68.

[11] K. Mokhtarian and M. Hefeeda, "End–to–End Secure Delivery of Scalable Video Streams," in *Proc. 19th Intern. Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'09)*. New York, USA: ACM, 2009, pp. 79–84.

[12] A. Eleftheriadis, R. Civanlar, and O. Shapiro, "Multipoint videoconferencing with scalable video coding," *J. Zhejiang Univ. SCIENCE A*, vol. 7, no. 5, pp. 696–705, 2006.

[13] Vidyo Inc., "Vidyo homepage," 2009, http://www.vidyo.com.

[14] M. Palkow, "The daViKo homepage," 2009, http://www.daviko.com.

[15] 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.

[16] "SVC Reference Software," http://ip.hhi.de/imagecom_G1/savce/ downloads/SVC-Reference-Software.htm, 2010.

[17] R. Prasad, C. Dovrolis, M. Murray, and K. C. Claffy, "Bandwidth Estimation: Metrics, Measurement Techniques, and Tools," *IEEE Network*, vol. 17, no. 6, pp. 27–35, November–December 2003.

[18] M. Jain and C. Dovrolis, "End–to–End Available Bandwdith: Measurement Methodology, Dynamics, and Relation with TCP Thoughput," in *Proc. ACM SIGCOMM'02*. New York, 2000, pp. 295–308.