IS FINE-GRANULAR SCALABLE VIDEO CODING BENEFICIAL 
FOR WIRELESS VIDEO APPLICATIONS?

Thomas Stockhammer
Institute for Communications Engineering
Munich University of Technology
80290 Munich, Germany

ABSTRACT

In this work we compare standard-compliant H.264 video transmission with an efficient fine granular scalable coding method based on H.264, the progressive texture video coding (PTVC) scheme in a wireless conversational application environment. Therefore, we operate on a simple but meaningful wireless channel model, the block-fading AWGN channel. This model is suitable to represent wireless transmission conditions with respect to time-variance and resource limitations. Suitable error control schemes based on punctured high-memory convolutional codes and appropriate decoding methods are applied. Rate-control based on rate-distortion criteria allow selecting optimized source and channel coding rates for both video coding schemes. In addition, the exploitation of network feedback - typically available in conversational services – is employed using NEWPRED techniques. A fair comparison of the systems shows that the standard-compliant non-scalable system outperforms the advanced scalable systems in terms of average decoded PSNR. The benefits of the scalable system are discussed and also presented via experimental results.

1 INTRODUCTION

Applications like video telephony, video conferencing, multimedia streaming, or multimedia messaging to wireless clients will be important features in emerging 2.5G, 3G, and future mobile systems and may be a key factor to their success. The challenges involved in wireless video are manifold, e.g., to specify proper video coding techniques, to design networks appropriately, to apply suitable error protection and transmission schemes, as well as to limit encoding and decoding complexity. Due to the heterogeneity of transmission systems and the multitude of wireless devices having different resource capabilities, it has been discussed recently whether scalable video coding techniques are essential for universal multimedia access schemes [1]. Whereas in MPEG-4 this discussion was accommodated with the standardization of fine-granularity-scalability (FGS) [2], the experts within the Joint Video Team (JVT) did not include scalability, especially FGS, in the final standard of MPEG-4 AVC/H.264 [3] although network friendliness and wireless transmission were major concerns of the standardization project. Due to the likely business models in emerging wireless systems that the end-user’s costs are proportional to the transmitted data volume, compression efficiency is the main target for wireless video and multimedia applications. Although new approaches, e.g. [4], show promising potentials, scalable coding usually performs inferior compared to non-scalable. Therefore, for typical wireless video applications over mainly IP-based networks other means have been considered to compensate the shortcomings of heterogeneous wireless networks. For Multimedia Messaging Services (MMS) download and playback of the video file is usually completely separated and, therefore, reliable link layer techniques including re-transmissions as well TCP/IP on the transport layer can be used to guarantee reliable delivery. Real-time constraints can be ignored completely. More challenges are involved in streaming multimedia and especially video to wireless clients as playback usually starts before the download has been completed. In case of IP-based and/or wireless media streaming, delay jitter occurs due to packet delays and congestions in routers as well as end-to-end or link layer re-transmissions in case of packet losses. This can partly be compensated by a sufficiently large receiver buffer and an initial delay of several seconds. In addition, channel-adaptive streaming technologies based on non-scalable and scalable standard-compliant video coding systems have gained significant interest (see [5] and references therein). These techniques can be grouped into adaptive media play-out, rate-distortion optimized (RDO) packet scheduling, and channel-adaptive packet dependency control. The latter two techniques can successfully be combined. The video is encoded and packetized such that the streaming server can adaptively react to channel and network conditions by, e.g. prioritizing certain packets. Obviously, FGS encoded video can easily be integrated in this frame work. However, in general an efficient version switching with sync-pictures [6] combined with temporal scalability can provide better overall quality at the same bit-rate although the storage requirements at the streaming server might be slightly higher. It has been shown that in multicast and broadcast systems scalable coding is beneficial to serve users with different bit rates, but these applications are not within the primary scope of nowadays mobile systems.

This argumentation shows that for wireless MMS and streaming applications FGS coded does in general not provide significant benefits. This leaves the third major application category, low-delay real-time conversational services. The time-variant behavior of wireless channels is especially challenging for conversational services as neither deep interleaving on the link layer nor any kind of re-transmissions can be performed to avoid delay and delay jitter. Fast power control is only partly integrated in wireless systems and cannot always guarantee a stationary behavior of the channel. In the following we focus on this scenario as it potentially provides an application of FGS coded video with appropriate error correction schemes. We present a conventional solution based on H.264 in combination with equal error protection and compare it with a system employing a FGS coding method based on H.264, the progressive texture video coding (PTVC) scheme [7]. The PTVC is combined with an adapted error control system as already presented in [8]. For both systems appropriate network feedback techniques and rate allocation methods are applied. We show experimental results and compare both systems. Conclusions from these results will be drawn and, the title question can at least partly be answered.
2 CONVERSATIONAL VIDEO OVER LOW BIT-RATE WIRELESS CHANNEL

2.1 Application Constraints and Wireless Channel Characteristics

Wireless channels are usually constrained in transmit power and available transmission bandwidth. In addition, a highly time-variant behavior in terms of receive power due to short-term fading effects and interference is experienced. An appropriate modeling and simulation of mobile channels and systems on its own is a very challenging task. Additionally, to exploit and compare coding schemes in the area of source and channel coding is even more complex. Therefore, it is reasonable to define simple, but meaningful models for wireless channels to evaluate the transmission of video and multimedia data. In [9] a block-fading additive white Gaussian noise (BF-AWGN) channel with perfect channel state information at the receiver is introduced as an appropriate model for many mobile channels. The characteristics of time- and frequency-hopping mobile systems are appropriately modeled by a block-fading channel. Additionally, if we assume random activity of users over time a multi-user system can be modeled as block-fading. In the remainder of this work we assume that the transmitter has knowledge of the channel statistics, i.e., the distribution of the channel gain, and the average signal-to-noise ratio SNR. The propagation channel is assumed slowly time-varying and frequency-flat for each time slot. In particular, the channel gain is assumed to be constant on the entire slot and iid Rayleigh.

If we apply interleaving and channel coding over \( S \) radio slots, the variance of the channel can be reduced. In our case we use high-memory (memory 96) punctured convolutional codes with sequential decoding with rates from 1/7 to 1 and puncturing period 32. For details on the channel coding and interleaving we refer to [8]. The effects of channel coding and interleaving over different number of radio slots is shown in Figure 1: For average SNR 4 dB, only one radio slot \( S=1 \) and channel coding rate 0.5 the outage probability is 0.3, whereas for increased number of radio slots \( S=4 \) and \( S=16 \) the variance of the channel can be reduced and the outage probability decreases. For TDMA systems such as GSM interleaving usually spreads one channel coded block over only a few radio slots to limit the delay. For the remainder we apply the commonly used interleaver spread \( S=4 \).

According to Figure 1, the outage probability for \( S=4 \) of this typical wireless channel varies significantly with the applied channel coding rate. We assume a fixed bandwidth and binary modulation such that the maximum bit-rate with channel coding rate \( \lambda \) is 64 kbit/s. The maximum bit-rate results by occupying 40 radio slots per second, each with 1600 binary channel symbols. For low channel coding rate the outage probability is low, whereas for increasing channel coding rate the outage probability increases also. However, lower channel coding rate also results in lower effective bit-rate for the video and vice versa. Therefore, residual channel coding block errors and video coding rate has to be traded. As conversational video applications require low-delay encoding and transmission to maintain real-time impression, re-transmissions of lost radio blocks are usually not possible. However, in general a fast and low bit-rate feedback channel is available in bi-directional conversational applications. In addition we require a constant total bit-rate for each frame to fulfill the stringent delay constraints of the application.

2.2 Conventional System

A conventional video transmission system based on H.264 is shown in Figure 2. Video encoding is based on a sequential encoding of frames. In most existing video coding standards including H.264, within each frame video encoding is typically based on sequential encoding of macroblocks (MBs). Although slices could be formed, the strategy of one frame in one packet is usually beneficial [10] to exploit the full intra-frame correlations within one frame. Each generated frame is channel encoded and transmitted over the wireless channel. The applied channel coding maps the wireless channel into a perfect packet erasure channel, i.e., frames are either lost or perfectly decoded. With that, we can define the frame loss or channel behavior \( C \) as a binary sequence of length \( n \) indicating whether the frame is lost (indicated by \( 0 \)) or correctly received (indicated by \( 1 \)). As already mentioned, in addition to the forward link it possible that a low bit-rate reliable back-channel from the decoder to the encoder is available which allows reporting a \( d \)-frame delayed version the observed channel behavior \( c_{n-d} \) at the decoder to the encoder. The decoder processes the received sequence of packets. Whereas correctly received packets are decoded as usual for the lost packet an error concealment algorithm has to be invoked. In our case we simply copy the previous frame. In Inter mode, i.e., when motion-compensated prediction (MCP) is utilized, the loss of information in one frame has a considerable impact on the quality of the following frames, if the concealed image content is referenced for MCP. Because errors remain visible for a longer period of time, the resulting artifacts are particularly annoying. Therefore, due to the motion compensation process and the resulting error propagation the reconstructed image depends not only on the lost packets for the current frame but in general on the entire channel loss sequence.
From this system perspective an error-resilient video coding standard suitable for conversational wireless services has to provide features to combat two problems, while focusing on prime goal of high compression efficiency. On the one hand it is necessary to avoid errors completely and to minimize the visual effect of errors within one frame. On the other hand, as errors cannot be avoided, the well-known problem of spatio-temporal error propagation in hybrid video coding has to be limited. In [10], standard compliant error resilience techniques for packet lossy channels have been investigated. For the expected frame error rates 2-10% and with the low-delay feedback channel available, the best performance is provided by a system which employs multiple reference frames and feedback information to predict from acknowledged frames only [11]. The system outperforms any kind of channel-adaptive intra updates, e.g. [12], as well as any kind of intra-frame packetization schemes in combination with error concealment schemes.

3 SCALEABLE WIRELESS VIDEO TRANSMISSION SYSTEM

The problem of the non-scalable system comes for the “all-or-nothing” transmission strategy. Either the entire frame can be decoded or everything is lost. The probability of the loss obviously depends on the applied channel coding rate. According to Figure 1, if we reduce the channel coding rate below 0.3 we can almost completely avoid any transmission frame losses. However, with this strategy we have overprotected the source for most channel realizations and therefore limit the average bit-rate for the video transmission. It has been shown for still image transmission that the application of scalable or progressively coded source in combination with unequal error protection can enhance the system significantly [13], [14], [15], [16]. In [16], we have presented a channel and complexity scalable transmission system which outperforms previous approaches especially for wireless fading channels. We have extended this system according to Figure 3 to the transmission of low-delay conversational video in [8]. We will present briefly components of this advanced system, for details we refer to [8].

Figure 3 Scalable video coding system: PTVC, source and channel rate allocation, unequal error protection, far end error decoder, and progressive texture video decoding.

The conventional non-scalable video coder is replaced by the PTVC which allows for progressive coding of the entire texture information. The PTVC is based on the H.26L test model TML4.0 which has been modified to perform motion estimation and compensation only, i.e. coding of transform coefficients has been disabled in the H.26L codec. Instead, we use an embedded bit-plane coding similar to JPEG-2000 to represent I-frames and the displaced frame difference (DFD) resulting from motion compensation. The resulting video bit-stream consists of the combination of the H.26L control and motion component and the progressive texture bit-stream. For details, we refer to [7].

The reconstruction quality of efficient progressive source coding schemes such as the PTVC is severely affected by residual errors. Due to the properties of progressive coding any data after the first decoding error cannot be interpreted any more. In fact, even if the data following the decoding error is correct, usually the reconstruction quality decreases due to effects like synchronization loss, etc. Thus, it is vital in progressive coding to only deliver error-free data to the source decoder, i.e., to detect and localize the first decoding error. To achieve this goal in [8] a forward error correction scheme has been proposed. This system includes a high-memory systematic convolutional code and a puncturing unit such that a regressive redundancy profile can be generated, i.e., unequal error protection over the progressive data frame is applied. Decoding is performed with the Far End Error Decoder (FEED) algorithm. FEED is a decoding algorithm which allows inherent error detection and error localization and aims to delay the first error in the decoded frame as far as possible. The algorithm provides a reliability vector which contains the reliabilities for each decoded sub-path. The shortening unit outputs a reliable sub-path from the beginning of the frame according to some specified reliability. This reliable sub-path of length \( e_{in} \) is used to decode the current video frame in combination with the motion compensation process. In addition, a more reliable shorter sub-path is generated of length \( e_{out} \) which is used to generate a new reference frame. In addition, the length of this highly reliable sub-path is sent back to the encoder via a feedback channel. The usually d-frame delayed version \( e_{out} \) is used to reconstruct the identical reference frame at the decoder such that drift between encoder and decoder MCP process can be avoided. For more details on the system we refer to [8].

4 OPTIMIZED RATE ALLOCATION

The performance of both systems significantly depends on the selected source and channel coding rates. Therefore, we introduce rate allocation schemes which target to minimize the expected end-to-end distortion for each frame to be encoded. The total number of channel symbols for each frame is constrained to \( N \) to meet the delay constraints. For the non-scalable system we search a quantization parameter \( q \) and a channel coding rate for each frame such that the expected end-to-end distortion is minimized:

\[
\min_{r} \left[ D_0 + D_q \right],
\]

where \( D_0 \) and \( D_q \) denote the concealment distortion and the coding distortion, respectively, when encoding with quantizer \( q \). Additionally, \( k \) denotes the source rate when encoding with quantizer \( q \) and \( p_{out}(r) \) the outage probability for channel coding rate \( r \). In case of the scalable coder in combination with the unequal error protection not only one channel coding rate has to be derived but one for all layer \( L \). In addition, the number of layer can be optimized to control the source coding rate. The details for rate allocation as well as a complexity reduced algorithm based on dynamic programming have been presented in [16]. To estimate the channel code performance in both cases information theoretic bounds based on the cut-off rate have been used which predict the performance of the applied codes quite well. For details on this we refer to [8]. The algorithms can be formulated to minimize the expected mean square error (MSE-RDO) or to maximize the expected PSNR (PSNR-RDO). Influence of the choice of the metric have been discussed in [13] and will be shown in the experimental results in the following section.

5 EXPERIMENTAL RESULTS

The following simulations have been carried out with the QCIF test sequence Foreman (30 Hz, 300 frames) at a constant frame rate of 10 Hz. To obtain long test sequences the original sequence
is looped. The number of frames coded for one experiment more than 6000. In the experiment we excluded the influence of the feedback delay and assumed that instantaneous feedback is available. Results with one or several frames delay on the feedback showed the same trends except for decreasing overall performance and lead to the same conclusions.

In Figure 4 the cumulative distribution of the decoded luminance PSNR for each frame is shown for the conventional H.264 based system with EEP and the PTVC system with UEP and FEED.

The average PSNR for the H.264 system is slightly better (about 0.2 dB for PSNR-RDO, about 0.5 dB for MSE-RDO) than the PTVC system with UEP. The gains mainly result from the better average encoding performance which can be observed in Figure 4 from the high probability of relatively good decoded PSNR, e.g. in more than 70% the decoded PSNR is above 28 dB for PSNR-RDO and H.264 with EEP. However, for the conventional system the probability of a bad frame with low PSNR is generally higher as it is more likely that the entire frame is lost because of the EEP error protection. For example, the probability that the decoded PSNR is below 23 dB is only about 2% for the PTVC UEP case, whereas for H.264 the probability is about 9% in case of MSE-RDO and about 7% in case of PSNR-RDO. The loss of an entire frame obviously results in a discontinuous representation of the sequence which might be annoying if for example fluent motion is a main criterion in the display quality. However, in general the advanced system with fine granular scalability and UEP does not provide a significant advantage compared to a conventional system objectively and from subjective observations. A generalization to any kind of sequences and channels is obviously not possible. However, it is not expected that the trend of the experimental results differ significantly for common video sequences and common wireless channels.

6 CONCLUSIONS

The argumentation in the introduction showed that uni-directional video applications such as MMS or streaming usually do not require a scalable codec as with common state-of-the-art non-scalable video codecs in combination with efficient version switching and network functionalities such as re-transmissions, etc., are sufficient. For conversational applications over wireless channels we have compared non-scalable and scalable video coding under similar application and channel constraints. For both systems optimized rate allocation and network feedback has been applied. From the experimental results it was observed that based on the average PSNR the non-scalable system outperforms the FGS system. However, with the FGS system fluent motion can be maintained and the probability of decoding bad frames is reduced significantly. Therefore, FGS is beneficial only under very restrictive requirements for the application and usually the increased system complexity does justify the replacement of conventional non-scalable coders with FGS coding.

REFERENCES