ADVANCES IN VIDEO COMPRESSION

Jens-Rainer Ohm

Chair and Institute of Communications Engineering, RWTH Aachen University
Melatener Str. 23, 52074 Aachen, Germany
phone: + (49) 241-80-27671, fax: + (49) 241-80-22196, email: ohm@ient.rwth-aachen.de
web: www.ient.rwth-aachen.de

ABSTRACT

The market for digital video is growing, with compression as a core enabling technology. Standardized methods as motion-compensated DCT, which have reached a level of maturity over decades, are dominating today's products. During the past years, another wave of innovation has occurred in video compression research and development, which leads to the assumption that it is still far from reaching its final bounds. Key factors are further improvements in motion compensation, better understanding of spatio-temporal coherences over shorter and longer distances, and more advanced encoding methods, as e.g. implemented in the new Advanced Video Coding (AVC) standard. While the adoption of successful research trends into standardization and products seems to occur almost seamlessly, new trends are already showing up which may lead to more paradigm shifts in the video compression arena. In general, the interrelationship between transmission networks and compression technology bears many problems yet to be solved. Efficient scalable representation of video becomes more interesting now, providing flexible multi-dimensional resolution adaptation, to support various network and terminal capabilities and better error robustness. This has finally become possible by the advent of open-loop temporal compression methods, denoted as motion-compensated temporal filtering (MCTF), which are presently investigated for standardization. More improvements seem to be possible in the fields of motion compensation and texture representation, but also new application domains are emerging. 3D and multi-view video seems will become more realistically used in applications, due to the availability of appropriate displays. Mobile and sensor networks are raising a future demand for low-complexity encoders. The talk will analyse some of these trends and investigate perspectives and potential developments.

1. INTRODUCTION

The most important series of recommendations and standards for video compression were defined by groups of the ITU (ITU-T Rec. H.261/262/263/264) for application domains of telecommunications, and by the Moving Pictures Experts Group (MPEG) of the ISO/IEC (ISO standards MPEG-1/-2/-4) for applications in computers and consumer electronics, entertainment etc. They are reflecting leading-edge research results of their respective time. There are commonalities between the standards defined by the both groups, and also some work has been done jointly. For example, ITU-T rec. H.262 is identical to the ISO standard MPEG-2, and ITU-T rec. H.264 is identical to part 10 of the ISO standard MPEG-4, 'Advanced Video Coding'.

Proprietary solutions beyond open standards also exist, but mostly use very similar methods as the open standards, which were developed by strong groups of researchers both from industry and academia. The basic principle behind all compression standards, as relevant today in Internet transmission, broadcast and storage of digital video, is the so-called *hybrid coding* approach, a combination of motion-compensated prediction over the time axis with spatial (2D) block transform coding of the prediction error residual, or of newly encoded intraframe information. All necessary components such as motion vectors, switching modes, transform coefficients are typically variable-length coded for maximum compression performance. Advances have been made over time in all those elements and tools of video codecs, their interrelationships becoming better understood. As an example for recent developments, the youngest member in the series of open standards, known as H.264 or MPEG-4 part 10 "Advanced Video Coding" (AVC) is described by more detail in the following section. Other standards, like MPEG-4 part 2 "Visual", are targeting different functionalities, such as object-based coding (arbitrary shape) or scalability of bit streams. Section 3 reflects more recent progress in Scalable Video Coding (SVC) standardization, which allows more flexibility than ever achieved before in post-compression adaptation and tailoring of video streams, supporting specific needs of applications, networks and terminals, while retaining good compression performance. In section 4, some visible and possible trends for future developments in video compression are discussed.

2. ADVANCED VIDEO CODING

The demand for ever increasing compression performance has initiated the creation of a new part of the MPEG-4 standard, ISO/IEC 14496-10: *'Coding of Audiovisual Objects Part 10: Advanced Video Coding'*, which is identical by text with ITU-T Rec. H.264. The development of AVC was performed by the *Joint Video Team* (JVT), which consists of members of MPEG and of the ITU-T Video Coding Experts Group. To achieve highest possible compression performance, any kind of backward or forward compatibility with earlier standards was given up. Nevertheless, the basic approach is again a hybrid video codec (MC prediction + 2D transform), as in the predecessors. The most relevant new tools and elements are as follows:

- Motion compensation using variable block sizes (denoted as sub-macroblock partitions) of 4x4, 4x8, 8x4, 8x8, 8x16, 16x8 or 16x16; motion vectors encoded by hierarchical prediction starting at the 16x16 macroblock level;
- Motion compensation is performed by quarter-sample accuracy, using high-quality interpolation filters;
- Usage of an integer transform of block size 4x4, optionally switchable to 8x8. This is not a DCT, but could be interpreted as an integer approximation thereof. The transform is orthogonal, with appropriate normalization to be observed during quantization. For the entire building block of transform and quantization, implementation by 16-bit integer arithmetic precision is possible both in encoding and decoding. To compensate for the negative impact of possibly small block sizes on the compression performance, extensive usage of spatial inter-block prediction is made, and a sophisticated deblocking filter is applied in the prediction loop.
- Intraframe coding is performed by first predicting the entire block from the boundary pixels of adjacent blocks. Prediction is possible for 4x4 and 16x16 blocks, where for 16x16 blocks only horizontal, vertical and DC prediction (each pixel predicted by same mean value of boundary pixels) is allowed. In 4x4 block prediction, directional-adaptive prediction is used, where eight different prediction directions are selectable. The integer transform is finally applied to the residual resulting from intra prediction.
- An adaptive de-blocking filter is applied within the prediction loop. The optimum selection of this filter highly depends on the amount of quantization error fed back from the previous frame. The adaptation process of the filter is non-linear, with lowpass strength of the filter steered firstly by the quantization parameter (step size). Further parameters considered in the filter selection are the difference between motion vectors at the respective block edges, the coding mode used (e.g. stronger filtering is made for intra mode), the presence of coded coefficients and the differences between reconstruction values across the block boundaries. The filter kernel itself represents a linear 1D filter with directional orientation perpendicular to the block boundary, impulse response lengths are between 3 and 5 taps depending on the filter strength.
- Multiple reference picture prediction allows defining references for prediction of a macroblock from any of up to F previously decoded pictures; the number F itself depends on the maximum amount of frame memory available in a decoder. Typically, values around F=5 are used, but for some cases F=15 could also be realized.
- The bi-directional prediction approach (denoted as *bi-predictive* or *B*-type slices) is generalized compared to previous standards. This in particular allows to define structures of prediction from *two previous* or *two subsequent* pictures, provided that a causal processing order is observed. Furthermore, follow-up prediction of *B*-type slices from other *B*-type slices is possible, which e.g. allows implementation of a *B*-frame based temporal-differences pyramid. Different *weighting factors* can be used for the reference frames in the prediction (formerly, this was restricted to averaging by weighting factors 0.5 each). In combination with multi-reference prediction, this allows a high degree of flexibility, but also could incur less regular memory accesses and increased complexity of the encoder and decoder.
- Two different entropy coding mechanisms are defined, one of which is *Context-adaptive VLC* (CAVLC), the other *Context-adaptive Binary Arithmetic Coding* (CABAC). Both are universally applicable to all elements of the code syntax, based on a systematic construction of variable-length code tables. By proper definition of the contexts, it is possible to exploit non-linear dependencies between the different elements to be encoded. As CABAC is a coding method for binary signals, a binarization of multi-level values such as transform coefficients or motion vectors must be performed before it can be applied; four different basic context models are defined, where the usage depends on the specific values to be encoded [1].
- A *Network Abstraction Layer* (NAL) is defined for simple interfacing of the video stream with different network transport mechanisms, e.g. for access unit definition, error control etc.

The key improvements are indeed made in the area of motion compensation. The sophisticated loop filter allows a significant increase of subjective quality at low and very low data rates, and also compensates the potential blocking artifacts produced by the block transform. State-of-the-art context-based entropy coding drives compression to the limits. On the other hand, the high degrees of freedom in mode selection, reference-frame selection, motion block-size selection, context initialization etc. will only provide significant improvement of compression performance when appropriate optimization decisions, in particular based on rate-distortion criteria, are made by the encoder. Such methods have been included in the reference encoder software (Joint Model, JM) used by the JVT during the development of the standard.

The combination of all these different methods has led to a significant increase of compression performance compared to previous standard solutions. Reduction of the bit rate at same quality level by up to 50%, as compared to H.263 or MPEG–4 Simple Profile, and up to 30% as compared to MPEG–4 Advanced Simple Profile have been reported [2]. On the other hand, the complexity of encoders must significantly be increased as well to achieve such performance. Regarding decoder complexity, the integer transform and usage of systematic VLC designs clearly reduces the complexity as compared to previous (DCT based) solutions, while memory accesses are also becoming more irregular due to usage of smaller motion block sizes and possibility of multi-frame access. The loop filter and the arithmetic decoder also add complexity.

3. SCALABLE VIDEO CODING

Video and motion pictures are frequently transmitted over variable-bandwidth channels, both in wireless and cable networks. They have to be stored on media of different capacity, ranging e.g. from memory sticks to next-generation DVD. They have to be displayed on a variety of devices, including a range from small mobile terminals up to high-resolution projection systems. *Scalable video coding* (SVC) schemes are intended to encode the signal once at highest rate and resolution, but enable decoding from partial streams in any situation, depending on the specific rate and resolution required by a certain application. This enables a simple and flexible solution for transmission over heterogeneous networks, additionally providing adaptability for bandwidth variations and error conditions. Both multicast and unicast streaming applications are enabled, with minimal processing at server/network and low-complexity decoding. It further allows simple adaptation for a variety of storage devices and terminals. For highest flexibility, scalability providing a fine granularity at the bitstream level, and universality in terms of different combinations, are desirable. The most important scalability dimensions that have to be provided are various spatial, temporal and quality-level resolutions; the latter is often referred to as SNR scalability.

For video coding, a lack of efficiency had been observed in the past (also in standard implementations such as MPEG-2 and MPEG-4) when combining scalable coding with motion-compensated prediction and block transform encoding. A theoretical investigation shows that this is mainly caused by the recursive structure of the prediction loop, which causes a drift problem whenever incomplete information is decoded [3][4]. As a consequence, wider acceptance of scalable video coding in the market of its prospective applications had never occurred in the past.

Within the last 5 years, a breakthrough was made in developing open-loop concepts, designated as motion-compensated temporal filtering (MCTF). Originally based on wavelet concepts, any infinite recursions in encoding and decoding are abandoned. This has finally enabled implementation of highly efficient scalable video codecs. As a consequence, MCTF schemes can provide flexible spatial, temporal, SNR and complexity scalability with fine granularity over a large range of bit rates, while maintaining a very good compression performance. The inherent prioritization of data in this framework leads to added robustness and considerably improved error concealment properties as a by-product.

Early implementations of MCTF did not provide the capability for perfect reconstruction under constraints of arbitrary motion vector fields [5]. A very efficient implementation of pairs of bi-orthogonal wavelet filters is made by the *lifting structure* [6]. The first step of the lifting filter is a decomposition of the signal into its even- and odd-indexed polyphase components. Then, the two basic operations are *prediction steps* $P(z)$ and *update steps* $U(z)$. The prediction and update filters are primitive kernels, typically of filter lengths 2-3. As the lifting method is inherently reversible and guarantees perfect reconstruction, the application of this approach for MCTF in video compression was straightforward, and first proposed in [7],[8] and [9]. Here, the polyphase components are even and odd indexed frames "A" and "B". An illustration for the case of the 5/3 bi-orthogonal filter is given in Fig. 1. The result of the prediction step is a highpass frame "H", which is generated as motion-compensated prediction difference of a frame A and its two (averaged) neighbours B. The complementary update step shall usually revert the motion compensation and leads to a lowpass filtered (weighted average) representation "L" of a frame B, including information from its two previous and two subsequent neighbours.

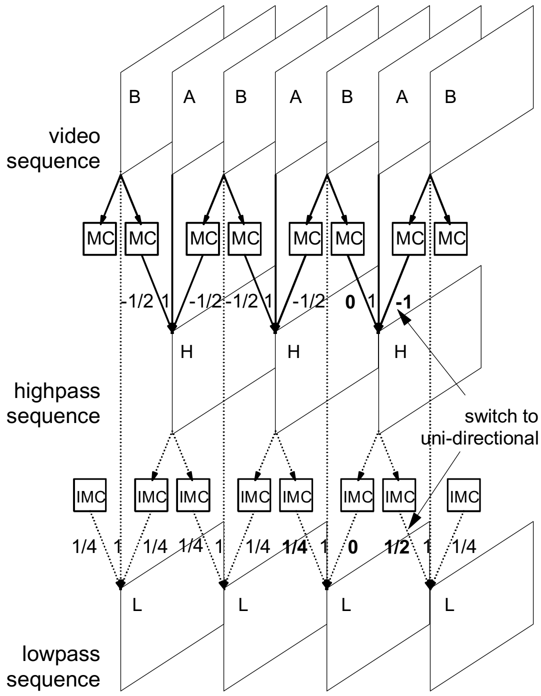

Fig. 1. MCTF lifting scheme for decomposition of a video sequence into lowpass (L) and highpass (H) frames

The lifting process needs to be adapted in cases where motion estimation and compensation are likely to fail, e.g. in case of fast inhomogeneous motion or occlusions. This can e.g be implemented by switching to uni-directional prediction and update, by omitting the update step, or by omitting the prediction step (intraframe coding modes). Typically, in a wavelet pyramid, the MCTF decomposition is iteratively performed over several hierarchy levels of L frames. In fact, the B-slice pyramid as specified in the AVC standard can be interpreted as a simple variant thereof, omitting the update step.

The MCTF process including prediction and update step establishes an orthogonal or bi-orthogonal transform with beneficial effect on equal distribution of coding errors among the different frames, including a de-noising effect in case of discarded highpass frame information. Due to the non-recursive (open-loop) structure, scalability can easily be implemented

- temporally, by discarding H frames
- quality-wise, by reducing the quality of H and L frames without drift or leakage effects

– spatially, by running a pyramid (wavelet-based or layered differences) over several spatial resolutions.

The layered concept is presently implemented in the Working Draft of the most recent SVC standardization activity of ISO-MPEG's and ITU-VCEG's Joint Video Team (JVT), which is planned to become an amendment to the AVC standard. The most important properties are:

– SVC supports scalable bit streams with typically up to 4 spatial resolutions, up to 4 different frame rates, and fine granularity of quality scalability at the various spatio-temporal resolutions;
– SVC will become an extension of AVC with capability to implement an AVC-compatible base layer;
– Motion-compensated temporal filtering (MCTF) with adaptive prediction and update steps is implemented for efficient open-loop compression;
– Encoder and decoder are run in a layered structure with "bottom-up" prediction from lower layers; this is simply implemented by allowing additional coding modes that switch between intra-layer and inter-layer prediction. For a more simple implementation, only one motion compensation loop is used;
– The fine granularity of scalability (FGS) functionality is implemented as an extension of CABAC, where additional contexts are defined between different quality layers. The basis of binarization is quantization of residual signals, where each subsequent quantizer provides typically half of the quantizer step size.
– All bitstream scalability is supported by the high-level syntax, defining new types of NAL units. This shall allow extremely simple truncation of packetized streams, which could be implemented by evaluating only one byte in the NAL unit header.

Results as achieved by the current status of development already indicate that the compression performance of the SVC extension is quite competitive when compared to non-scalable AVC coding. The steepness of rate-distortion plots over a wide range of rates and all resolutions appears quite "typical", indicating that the new codec does not dramatically penalize either high or low rate points (as previous scalable solutions did). In average, even though additional syntax overhead is necessary to allow scalable stream access, the rate increase as compared to non-scalable coding is not higher than 10%. Therefore, the new SVC scheme offers a clear compression advantage as compared to simulcast or parallel storage of non-scalable streams, as would otherwise be necessary to support different rates and resolutions. For certain sequences, in particular with smooth and uniform-motion video, SVC is even compressing better than non-scalable AVC at some rate points, most probably due to the usage of the open-loop MCTF tool.

Figure 2 shows PSNR vs. rate plots, comparing the performance of JVT's Joint Scalable Video Model (JSVM-0) software against the non-scalable (AVC-based) Joint Model (JM 9.4) software. More improvements are still to be expected, and up-to-date results will be presented at the conference.

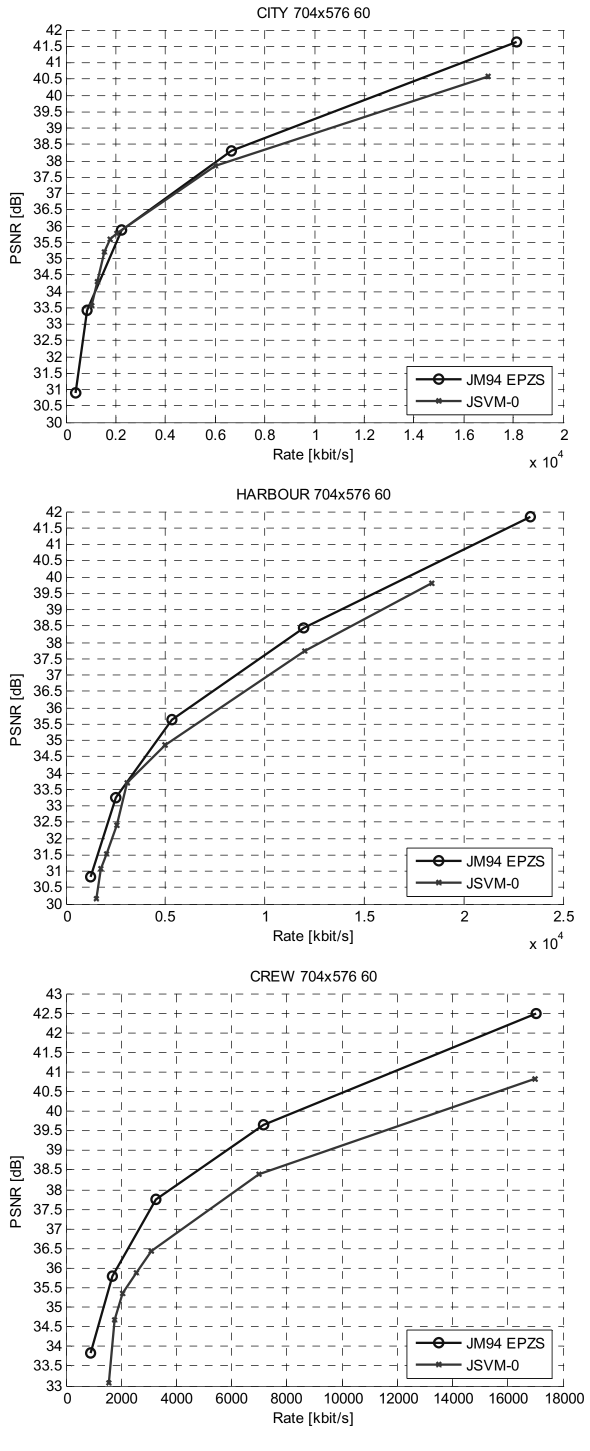

Fig. 2. PSNR vs. rate plots gained from the software implementations of JSVM-0 (scalable) and JM 9.4 (non-scalable), spatial resolution 4CIF, fully down-scalable into CIF and QCIF.

The new standard amendment is planned to be finalized by mid of 2006. Its capability to be run backwards compatible with AVC streams (base layer decoding capability), as well as the re-use of many existing AVC elements are expected to support fast acceptance in future video compression devices.

4. TRENDS FOR THE FUTURE

As described in the previous sections, video coding has shown a continuous and significant progress for improved compression in the past, which enabled a data rate reduction by approximately a factor of 2-3 per decade. The price that had to be paid was increased encoder and decoder complexity, which was however affordable considering the progress in microelectronics. It is not unlikely that this trend for better compression will continue, because video represents a highly complex multi-dimensional class of signals with many properties not yet fully understood. As an example, only recently capabilities of longer-term redundancies in video signals were started to be exploited by multi-reference prediction and MCTF. This is of particular importance, as it opens the door for better usage of spatio-temporal masking effects, fundamentally differentiating video coding from intraframe or still image coding. Further progress in compression efficiency seems feasible e.g. by improved motion compensation models, improved texture coding or synthesis of motion and texture appearance at the receiver end, also taking into account the distinction between visible and invisible artefacts instead of pixel-wise fidelity [10].

While improved compression efficiency seems to remain the main driving factor of further advances in video compression, it could be accompanied by supporting additional functionality for specific applications, as shown above for the Scalable Video Coding development. Demand for new elements in video coding could arise in the future by the following factors:

- Development of multiview applications and displays: When a scene is captured by multiple cameras, e.g. for the purpose of user-controlled navigation through a video scene, or for 3D display with view adaptation, the video data rate may in the worst case linearly increase by the number of cameras that are used for simultaneous capturing. It will therefore be necessary to further compress the overall rate by proper exploitation of coherences between the different available views (inter-view redundancy) instead of independently compressing the different views.
- Beyond the scalability solution, heterogeneous and error-prone networks could even better be served by closer integration of source and channel coding, which should be similarly adaptable depending on network error conditions as SVC is adapting for rate conditions. It is still to be proven whether such approach can overcome unequal error protection strategies.
- Envisaged sensor networks, where myriads of video cameras (possibly mobile) would produce up-stream content that needs to be evaluated by stationary devices. In this case, the broadcast paradigm that has influenced video coding standardization so far (requirement for many low-cost decoders) would be reverted. Coding concepts which support such solutions are known as *distributed source coding* [11]. According to the theoretic bounds of Wyner-Ziv coding, compression performance should in principle not suffer when sufficient estimation complexity is shifted to the decoder side, and appropriate side information is available at the receiver end. However, the present state of the art in distributed video coding still seems far out from achieving such goal. Nevertheless, understanding of video decoders as more universal "source state estimators", and supplementing encoders by more normative and reproducible behaviour might lead to other interesting insights into further compression advances.

All these developments are ruled by the fact that microelectronics as well as software technology are still fast developing. This has already led to a situation where the present generation of video coding standards allows much more flexibility in supporting different modes and configurations than it was the case in the early 90ies. It also leaves much more headroom for improvements, by proper understanding and integration of different tools (elements) as flexibly needed by specific applications.

REFERENCES

[1] D. Marpe, H. Schwarz, T. Wiegand: "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 13 (2003), pp. 620-636

[2] T. Wiegand, G. J. Sullivan, G. Bjontegaard, A. Luthra: "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circ. Syst. Video Tech.*, vol. 13 (2003), pp. 560-576

[3] J.-R. Ohm: "Advances in Scalable Video Coding," *Proc. IEEE*, vol. 93 (2005), pp. 42- 56

[4] J.-R. Ohm: "Multimedia Communication Technology," Berlin, New York: Springer, 2004

[5] J.-R. Ohm: "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Process.*, vol. 3 (1994), pp. 559-571

[6] W. Sweldens: "The lifting scheme : A new philosophy in biorthogonal wavelet constructions," in *Proc. SPIE*, vol. 2569 (1995), pp. 68-79

[7] B. Pesquet-Popescu and V. Bottreau: "Three-dimensional lifting schemes for motion-compensated video compression," in *Proc. IEEE ICASSP* (2001), pp. 1793-1796

[8] L. Luo, J. Li, S. Li, Z. Zhuang, and Y.-Q. Zhang: "Motion compensated lifting wavelet and its application in video coding," in *Proc. IEEE Intl. Conf. on Multimedia and Expo* (ICME 2001), Tokyo, Japan, Aug. 2001

[9] A. Secker and D. Taubman: "Motion-compensated highly-scalable video compression using an adaptive 3D wavelet transform based on lifting", in *Proc. IEEE ICIP* (2001), pp. 1029-1032

[10] A. Dumitras and B. G. Haskell: "An Encoder-Decoder Texture Replacement Method with Application to Content-based Movie Coding," IEEE Trans. on CSVT, Vol. 14 (2004), pp. 825-840

[11] B. Girod, A. M. Aaron, S. Rane, D. Rebollo-Monedero: "Distributed video coding," *Proc. IEEE*, vol. 93 (2005), pp. 71-83