RECENT ADVANCES IN VIDEO COMPRESSION

Bernd Girod *Khaled Ben Younes* *Reinhard Bernstein* *Peter Eisert* *Niko Färber* *Frank Hartung*
Uwe Horn *Eckehard Steinbach* *Klaus Stuhlmüller* *Thomas Wiegand*

Telecommunications Institute
University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
girod@nt.e-technik.uni-erlangen.de

ABSTRACT

This paper gives an overview over recent advances in video compression and transmission. The new H.263 video compression standard is an important milestone towards wide-spread use of personal video communication. For mobile video applications and video transmission over networks, error robustness and scalability are important issues. Structure-from-motion methods, region-based coding and model-based coding are research topics pointing into the future, towards MPEG-4.

1. INTRODUCTION

Digital video compression is one of the key issues in video coding, enabling efficient interchange and distribution of visual information. New applications in the field of communication, multimedia, and digital television broadcasting [1] require highly efficient, robust and flexible digital video compression and encoding techniques[2]. Multimedia applications range for example from desktop videoconferencing and computer-supported cooperative work to interactive entertainment networks where video-on-demand, video games, and teleshopping are provided. The integration of motion video as an integral part of multimedia environments is technologically one of the most demanding tasks, due to the high data rates and real-time constraints. According to the importance of the problem, there is a significant amount of ongoing research in the area of video compression. The rate required to encode a full motion video signal at VHS quality has come down from 20 Mbit/s around 1980 to well below 1 Mbit/s today. For head-and-shoulders scenes, typical for videoconferencing applications, rates can be substantially lower. In a parallel effort, international standardization organizations are actively developing standards for video compression, which facilitate interoperability among different video systems and motivate the development and production of VLSI systems and devices in addition to software solutions. Since standards reflect the state of the art, they serve as references on developing compatible extensions and novel algorithms. One of the most recent video coding standards, ITU-T H.263 is discussed in section 2. To ensure robust video transmission, standard compatible extensions are proposed in section 3. Potential tools for new standards such as scalable video coding and region based coding are outlined in section 4 and 5. 3D scene and motion estimation for model-based video coding schemes as discussed in section 6 can be supported by structure from motion algorithms, finally described in section 7.

2. THE H.263 VIDEO CODING STANDARD

The H.263 video coding standard [3] is a descendant of the motion-compensated DCT methodology prevalent in several existing standards such as H.261 [4], MPEG-1 [5], and MPEG-2 [6]. Together, their primary applications span the gamut from very low bit-rate video telephony to high quality HDTV with H.263 focusing initially on the very low bit-rate end, i.e. bit-rates below 64 kbps. However, the final specification of H.263 allows picture resolutions ranging from sub-QCIF with 128×96 to 16CIF with 1408×1152 pixels for the luminance component. Therefore, H.263 has emerged to a high compression standard for moving images, not exclusively focusing on very low bit-rates.

The improvements in H.263 compared to H.261 are mainly obtained by improvements to the motion compensation (MC) scheme. At 64 kbps, the performance gain of H.263 in its default mode compared to H.261 is approximately 2 dB in terms of PSNR for the Foreman sequence in QCIF resolution at 12.5 fps, mainly achieved by MC with half-pixel accuracy. The optional capabilities "Advanced prediction mode" which utilizes overlapped MC and "PB-frames mode" that is based on bidirectional MC together produce an additional PSNR gain of 1 dB. The improvements obtained by syntax changes of the bit-stream strongly depend on the bit-rate, whereas the advantage introduced by the "Syntax-based arithmetic coding mode" is rather marginal. A detailed description of the performance of H.263 in comparison to H.261 can be found in [7].

Several strategies to further improve the performance of existing standards have been exploited. Such standard compatible improvements, e.g., see [8] for MPEG and [9] for H.263, are based on the fact that the operational control of the encoder is not subject to recommendation. In [9], by employing rate-distortion optimized mode selection in H.263, a PSNR gain of 0.75 dB over a wide range of bit-rates for the Car Phone sequence compared to conventional mode decision is reported. Although the gains achievable with improved encoder control are significant, the methods remain limited by the underlying block-based compression approach.

0-7803-3073-0/96/$5.00

3. STANDARD-COMPATIBLE EXTENSIONS FOR ROBUST VIDEO TRANSMISSION

Many existing networks cannot provide a guaranteed quality of service. This may result from the underlying medium access control, like in IEEE 802.3 based Local Area Networks (Ethernet), or from the limitations of the transmission channel, e.g., in mobile environments where remaining errors can not be avoided during fading periods. Networks with these limitations are characterized as "best effort" networks and require increased robustness for the transmission of video. This is especially true when motion-compensated prediction is utilized, which is essential for high coding efficiency, but also causes spatio-temporal error propagation, i.e., the loss of information in one frame has considerable impact on the quality of the following frames.

Error propagation can be efficiently reduced when a feedback channel between transmitter and receiver is utilized. The system proposed in [10] tolerates errors, but limits their effect by error control techniques in the source codec. Error concealment is employed to hide visible distortion and residual errors are compensated using acknowledgment information from the receiver. Compensation is achieved in a standard-compatible way by coding image regions in INTRA-mode which could not be concealed successfully.

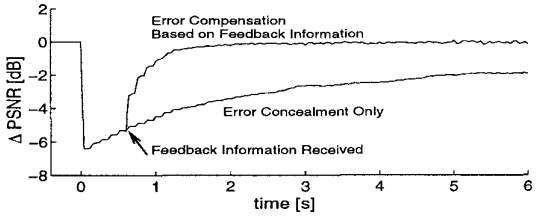

Figure 1. Recovery of PSNR.

Fig. 1 shows averaged simulation results for the application of this approach to H.263 [3]. The loss of picture quality ($\Delta PSNR$) compared to the correctly decoded sequence is shown for an error covering 30% of one frame. A residual loss of approximately 3 dB on average still remains after 3 seconds if the decoder relies on concealment only. Enhanced by error compensation, however, the image quality recovers rapidly as soon as the feedback information arrives at the encoder (a round-trip-delay of 700 ms is assumed in Fig. 1).

4. SCALABLE VIDEO CODING

Scalable video codecs offer substantial new features which address the special demand of today's multimedia communication systems, like computational limited decoding and video transmissions over heterogeneous packet networks or wireless channels [11]. Scalable coders are characterized by their ability to generate bitstreams which can be decoded at a variety of bit rates.

Common standards like MPEG-2 already contain basic mechanisms towards scalability, but only to a limited extent. More promising approaches can be built upon spatio-temporal resolution pyramids, first proposed by Uz and Vetterli in 1991 [12]. Open-loop as well as closed-loop pyramid coders can be used for their efficient encoding and multiscale motion compensation can be easily included [13]. Filters for spatial downsampling and interpolation operations can be kept very simple such that fast and efficient codecs become feasible. Since one is completely free in choosing appropriate filters, morphological filters can be used to further improve subjective image quality [14].

A pyramid coder is a multistage quantization scheme. Therefore, efficient compression requires careful bit allocation to the various quantizers depending on the image to code. Performing an optimal bit allocation for each frame becomes computational infeasible in pyramids with more than two layers. A comparison between open-loop and closed-loop pyramid coders shows that closed-loop coders are better suited for practical applications since they are less sensitive against suboptimal bit allocations such that simpler heuristics can be used [8].

Multiscale motion compensation can be utilized in several ways [15]. Motion vectors can be efficiently computed and encoded by hierarchical motion estimation. All compensation schemes we investigated so far perform very similar under the condition of equal quantization noise in each spatial resolution layer [16].

A simulation model we set up for coding experiments is built upon a four layer closed-loop pyramid coder combined with multiscale motion compensation [15]. For quantization, an E_8-lattice vector quantizer is used [17]. Spatial and temporal resolution of the finest resolution layer correspond to the ITU-R 601-4 standard. In each subsequent layer, spatial and/or temporal resolution is reduced. With a single encoded bitstream, bit rates between 64 kbit/s up to 4 Mbit/s are supported. On a Sun SparcStation5 realtime decoding with a software-only decoder is possible up to 300 kbit/s. Simulations have shown that realtime performance is roughly proportional to bit rate.

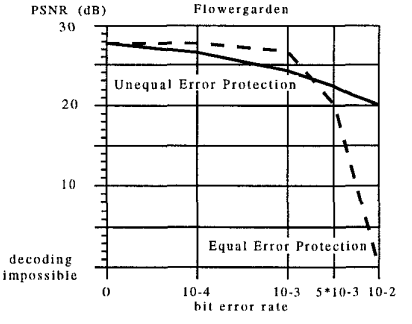

Figure 2. Scalable video coder combined with equal and unequal error protection

By adding an unequal forward error correction scheme to a scalable source coder, robustness against data loss can be noticeably increased. Fig. 2 shows simulation results for five different bit error rates on a binary symmetric channel for the Flowergarden test sequence. Unequal error protection is compared against equal error protection where all layers are protected at the same level. In both cases the increase in bit rate due to error protection is 26%. From Fig. 2 it can be seen that scalability combined with unequal error protection leads to a gracefully degrading coding scheme.

Scalability can also be utilized for transmission over heterogeneous packet networks since it offers an easy way to reduce the bit rate of transmitted video data in case of congestion [18]. By giving each packet a priority the network layer itself is able to reduce the bitrate without informing the coder or knowing the content of a packet.

5. REGION-BASED MOTION COMPENSATION

In current standards [3, 4, 5, 6], motion compensation and encoding of the prediction error is block-based. However, the emerging MPEG-4 standard is supposed to provide contents based functionalities in addition to efficient video coding [19]. To fulfill this task a source model is needed that adapts better to the scene contents than blocks. In so-called *object-based image coding techniques* [20] the partition of images into a fixed block structure is replaced by arbitrarily shaped regions that are related either to texture/color [21] or to motion [22]. The shape of the regions has to be transmitted as additional side information. The efficiency of such a region-based coding scheme depends very much on properly adjusting the amount of bits used for region coding.

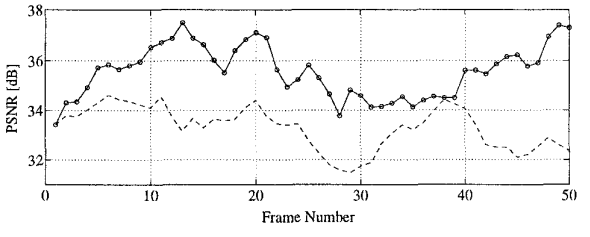

Figure 3. PSNR for the decoded QCIF sequence "Miss America" at 14 kbit/s, 8.33 Hz. '- -': block-based motion compensation, 'oo': region-based motion compensation

A solution for the optimum trade-off by applying rate-distortion theory has been presented recently for region-based motion compensation [22]. The regions are optimized with respect to a Lagrangian cost function by variation of the region contours. The resulting optimal contours do not necessarily coincide with the actual contours of the objects in the scene. However, for the optimized regions the improvement in distortion and the region's rate are well balanced in a rate-distortion sense. The improvement that has been achieved in [22] with a coding scheme using optimized regions is shown in Fig. 3. It can be seen that motion compensation with rate-distortion optimized regions is about 2 dB better than block-based motion compensation. In both cases a block-based DCT-coder is used for coding the prediction error image.

6. MODEL-BASED VIDEO CODING

A very promising technique for achieving very low data rates in video coding is model-based coding. It has been reported [23] that rates of less than 1000 bit/s are possible without reducing the resolution of the image. In a model-based system the coder analyzes the scene using three dimensional models of the objects. At the decoder the scene is synthesized with respect to the same models using parameters received from the coder. An important application for model-based coding is video telephony. The structure of such a system is shown in Fig. 4.

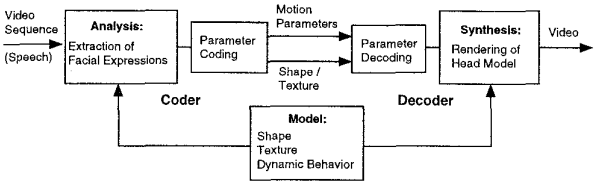

Figure 4. Concept of model-based coding

From a head-and-shoulders scene, the coder extracts both global and local motion caused by changes of facial expressions. The set of 3D motion parameters is then coded and transmitted. These few parameters are sufficient to reconstruct the whole image provided that the model of the head is known at the decoder. Model failures can be taken into consideration by transmitting shape and texture updates to the decoder. The 3D model specifying shape and color of the head is often modeled by a triangular wire frame with mapped texture [20].

The deformation of the wire frame model can be controlled with a mesh of B-spline control points [24], whereby the facial expressions are separated by *action units* based on the facial action coding system (FACS). The FACS system developed by Ekman and Friesen only represents the instantaneous state of the expression. Recent approaches extend this system with respect to the dynamic behavior [25]. To control the motion of the model, 3D motion parameters have to be extracted from the 2D video sequence. This can be done by tracking special features in the face or by using the optical flow field of successive frames. In both cases the knowledge of the scene and the 3D model improve these techniques leading to more accurate and efficient solutions.

7. STRUCTURE FROM MOTION

Model-based video communication systems attempt to extract information about the 3-dimensional structure of the scene to be transmitted. Assuming rigid objects, the relative motion between the camera and the objects permits to extract information about the scene depth. This is exploited in *structure from motion* algorithms recovering simultaneously 3D motion parameters and relative depth values. In [26] we present a new technique for 3D scene and motion estimation that is based on the observation that image point correspondences for a given 3D motion are constrained to lie on a straight line in the image, the epipolar line [23]. Precomputed displaced frame difference surfaces undergo an *Epipolar Transform* where the best match along an epipolar line is stored as a function of epipolar line parameters. A 5-dimensional motion parameter space is searched in a coarse-to-fine manner and a particular parameter set is evaluated by accumulation of the best match along the epipolar lines for predefined measurement windows.

A natural test scene consists of two frames of the sequence *Flowergarden*. Fig. 5 shows the estimated depth map. White image regions have small depth values, black blocks represent scene parts far away.

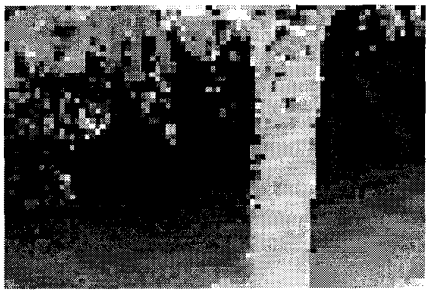

Figure 5. Estimated depth map

8. CONCLUSIONS

In this paper we have presented recent advances in video compression and transmission. H.263 and its extensions for mobile applications are important steps into a wide-spread use of personal video communication. For typical bit rates, H.263 performs up to 3 dB better than H.261. Using a feedback-channel can efficiently reduce the impact of transmission errors. Scalable video coding remains an important issue for video transmission over wirebound and wireless networks, especially when combined with unequal error protection at the different resolution layers which leads to graceful degradation with increasing bit error rate. Furthermore, it has been shown that the use of irregular regions for motion compensation rather than fixed blocks can yield gains of more than 2 dB, even considering the additional information to be transmitted. Structure-from-motion methods and model-based video coding are active research topics but promise significant improvements compared to todays standards, and will be considered in MPEG-4.

REFERENCES

[1] Special issue on advances in image and video compression. *Proceedings of IEEE*, Feb. 1995.

[2] Special issue on digital television - part 1: Enabling technologies. *Proceedings of IEEE*, June 1995.

[3] ITU-T Recommendation H.263. Video coding for low bitrate communication. (Draft), December 1995.

[4] ITU-T Recommendation H.261. Video codec for audiovisual services at $p \times 64$ kbits. December 1990, March 1993 (revised).

[5] ISO/IEC 11172-2. Information technology – coding of moving picture and associated audio for digital storage media at up to about 1.5 mbit/s: Part 2: Video. August 1993.

[6] ITU-T Recommendation H.262—ISO/IEC 13818-2. Information technology – generic coding of moving picture and associated audio information: Part 2: Video. (Draft), March 1994.

[7] B. Girod, E. Steinbach, and N. Färber. Comparison of the H.263 and H.261 video compression standards. In *Proc. SPIE*, volume CR60, Standards and Common Interfaces for Video Information Systems, October 1995.

[8] K. Ramchandran, A. Ortega, and M. Vetterli. Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders. *IEEE Trans. on Signal Processing*, 3(5):533–545, Sep. 1994.

[9] T. Wiegand, M. Lightstone, T.G. Campbell, D. Mukherjee, and S.K. Mitra. Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard. Submitted for publication to IEEE Trans. on Circuits and Systems for Video Technology.

[10] N. Färber, E. Steinbach, and B. Girod. Robust H.263 compatible video transmission over wireless channels. In *Proc. PCS*, Mar. 1996.

[11] B. Girod. Scalable video for multimedia systems. *Computers & Graphics*, 17(3):269–276, 1993.

[12] M.K. Uz, M. Vetterli, and D.J. LeGall. Interpolative multiresolution coding of advanced television with compatible subchannels. *IEEE Trans. on Circuits and Systems for Video Technology*, 1(1):86–99, Mar. 1991.

[13] M. Vetterli and K.M. Uz. Multiresolution coding techniques for digital television: A review. *Multidimensional Systems and Signal Processing*, 3:161–187, 1992.

[14] R. Bernstein and U. Horn. Linear and morphological pyramids for scalable image coding - A comparison. submitted to: 1996 International Conference on Image Processing, 1996.

[15] U. Horn, B. Girod, and B. Belzer. Scalable video coding with multiscale motion compensation and unequal error protection. In *Proc. International Symposium on Multimedia Communications and Video Coding*, New York, NY, Oct. 1995.

[16] U. Horn and B. Girod. Performance analysis of multiscale motion compensation techniques in pyramid coders. submitted to: 1996 International Conference on Image Processing, 1996.

[17] B. Girod, F. Hartung, and U. Horn. Subband image coding. In Ali N. Akansu and Mark J.T. Smith, editors, *Subband and Wavelet Transforms*, chapter 7. Kluwer Academic Publishers, Norwell, MA, 1995.

[18] U. Horn, B. Girod, and B. Belzer. Scalable video coding for multimedia applications and robust transmission over wireless channels. 7th International Workshop on Packet Video, Mar. 1996.

[19] R. Schäfer and T. Sikora. Digital video coding standards and their role in video communications. *Proc. IEEE*, 83(6):907–924, Jun. 1995.

[20] H.G. Musmann, M. Hötter, and J. Ostermann. Object-oriented analysis-synthesis coding of moving images. *Signal Processing: Image Communication*, 1(2):117–138, Oct. 1989.

[21] P. Salembier, L. Torres, F. Meyer, and C. Gu. Region-based video coding using mathematical morphology. *Proc. IEEE*, 83(6):843–857, Jun. 1995.

[22] K. W. Stuhlmüller, A. Salai, and B. Girod. Rate-constrained contour-representation for region-based motion compensation. In *Proc. Symp. on Visual Comm. and Image Proc.* SPIE, Mar. 1996.

[23] B. Girod. Image sequence coding using 3D scene models. *SPIE Symposium on Visual Communications and Image Processing 94*, Sept. 1994.

[24] M. Hoch, G. Fleischmann, and B. Girod. Modeling and animation of facial expressions based on B-splines. *The Visual Computer*, pages 87–95, Nov. 1994.

[25] I. A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 1994.

[26] E. Steinbach and B. Girod. Estimation of rigid body motion and scene structure from image sequences using a novel epipolar transform. *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, May 1996.