MULTI-FRAME MOTION-COMPENSATED VIDEO COMPRESSION FOR THE DIGITAL SET-TOP BOX

Bernd Girod and Markus Flierl

Information Systems Laboratory
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
{bgirod,mflierl}@stanford.edu

Invited Paper

ABSTRACT

Motion-compensated prediction based on multiple previous or future frames can enhance the compression efficiency of video coding. Multi-frame prediction can be applied as an extension to P-Pictures, but also to B-Pictures in the form of multi-hypothesis prediction. We review recent advances, several of which have been embraced by the ITU-T Rec. H.263 and the emerging JVT/H.26L standard, jointly developed by ITU-T and ISO/IEC MPEG. Already, JVT/H.26L video outperforms MPEG-2 by a factor greater than 2X. These recent developments could play a significant role in future digital set-top boxes.

1. INTRODUCTION

The average US household consumes more than 1500 hours of broadcast television per year. While most of this contents is still delivered as analog signals today, it is the equivalent of 3,000 Gigabytes (or 3 Terabytes), assuming typical bit-rates with today's MPEG-2 compression [1]. Multiplied by 70 million US households, we arrive at an aggregate bit-rate of $>$200 Exabyte/year (1 ExaByte = 10^{18} Byte). This number is all the more staggering, if we compare it to the annual production of original contents by mankind as a whole (estimated at 1-2 Exabyte in [1]) or the total annual Internet backbone traffic which is also in the order of 1 Exabyte in the US currently, growing by an annual factor of 2X [2].

Digital cable and satellite broadcasting today use the ISO/IEC standard MPEG-2 for video compression, and MPEG-2 decompression is the core digital set-top boxes. Cost-effective solutions are built around 2 integrated circuits, an integrated signal processor for demodulation, video and audio decompression and graphics overlay functions, augmented by several MByte of external random access memory (RAM). Advanced ICs for digital set-top boxes are typically powerful enough to decode several standard-definition television (SDTV) streams simultaneously. Video compression functionality has started to appear in digital set-top boxes for built-in magnetic disk storage of video programming. We expect future set-top boxes to store 1000s of hours of compressed video from a variety of sources, including digital broadcast over cable and satellite, as well as Internet video-on-demand. Moreover, we might also see integrated video conferencing capability. Further, as home networking evolves, digital set-top boxes will become media gateways, serving several audio and video play-out devices over a local area network that might comprise both wired and wireless segments.

Highly efficient video compression is essential for all these new application converging in the digital set-top box. Our introductory numbers game suggests that even a modest improvement of compression efficiency, say by 10%, would reduce the aggregate bit consumption of US households by an amount that is equivalent to 20X the current Internet backbone traffic! Since the standardization of MPEG-2, impressive progress has been made, which is now finding its way into the emerging JVT/H.26L standard [3], that is being developed jointly by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Motion Picture Experts Group (MPEG). While the new standard will be superior in many ways over current standards, one of the most important advances is the use of multi-frame motion-compensated prediction techniques, which, after first becoming part of the H.263 standard, are now being fully embraced.

As multi-frame prediction has important consequences for digital set-top boxes, both in terms of memory and computational requirements, we concentrate on these techniques in this paper. In Section 2, we introduce multi-frame motion-compensated prediction, the extension of current P-Pictures to multiple frames, in Section 3 multi-hypothesis prediction, the extension of current B-Pictures. In Section

4, we show that the emerging JVT/H.26L standard outperforms MPEG-2 by more than 2X.

2. MULTI-FRAME MOTION-COMPENSATED PREDICTION

About 20 years ago, video compression made the leap from intra-frame to inter-frame techniques. While the gains were initially unimpressive, inter-frame compression became more and more sophisticated over time, and significantly lower bit-rates were achieved at the expense of memory and computational requirements that were two orders of magnitude larger. Today, with the continuously dropping cost of semiconductors, we are able to afford another leap by dramatically increasing the memory and computational power in video codecs.

Multi-frame motion-compensated prediction extends the spatial displacement vector utilized in block-based hybrid video coding by a variable frame reference permitting the use of more frames than the previously decoded one for motion-compensated prediction [4]. The multi-frame buffer stores frames at encoder and decoder that are efficient for motion-compensated prediction. The use of multiple frames for motion compensation in most cases provides significantly improved coding gain. The frame reference parameter has to be transmitted as side information requiring additional bit-rate. To control the bit-rate budget, rate-constrained motion estimation is utilized.

Multi-frame motion-compensated prediction was first proposed as a technique to improve the error-resiliency of compressed video, either by using a randomly varying lag in the early work by Budagavi and Gibson [5], or by adaptive reference picture selection in response to acknowledgments, as incorporated into the Annex N of H.263 [6]. Multi-frame motion-compensated prediction for improved compression performance was first introduced by Wiegand, Zhang, and Girod in 1997 [7], [4]. These techniques became part of Annex U ”Enhanced Reference Picture Selection” of H.263 in 1999 [8] and are now an integral part of the emerging JVT/H.26L standard.

Fig. 1 provides video compression efficiency results for multi-frame prediction with the emerging JVT/H.26L standard. Coding efficiency achieved with prediction from the previous reference frame is compared to multi-frame prediction from five previous reference frames. With a buffer size of 5 frames, bit-rate savings of more than 12% can be observed for the CIF video sequence *Mobile & Calendar*. In the development of JVT/H.26L, simulations are typically compared against a test model with 5 previous reference frames.

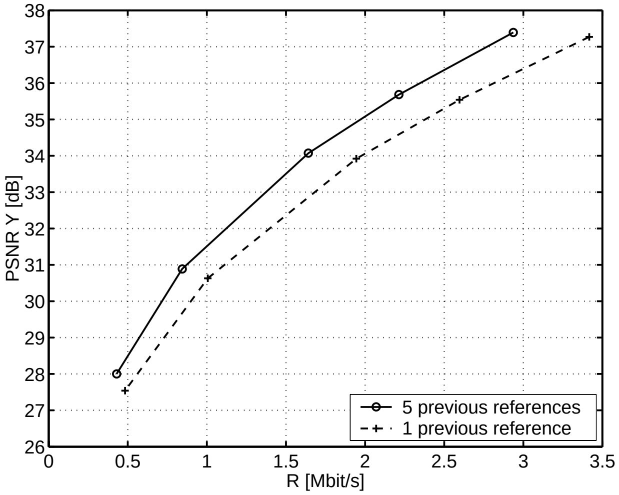

Fig. 1. Luminance PSNR vs. bit-rate for the CIF video sequence *Mobile & Calendar* (30 fps) compressed with JVT/H.26L. The quality of P-Frames with 1 previous reference frame is compared to P-Frames with 5 previous reference frames.

3. MULTI-HYPOTHESIS MOTION-COMPENSATED PREDICTION

B-Pictures are pictures in a motion video sequence that are encoded using both past and future pictures as references. A linear combination of forward and backward prediction signals enables bi-directional prediction (Fig. 2). However, such a superposition is not necessarily limited to forward and backward prediction signals [9, 10]. Multi-hypothesis prediction as proposed in [11] allows a more general form of B-Pictures.

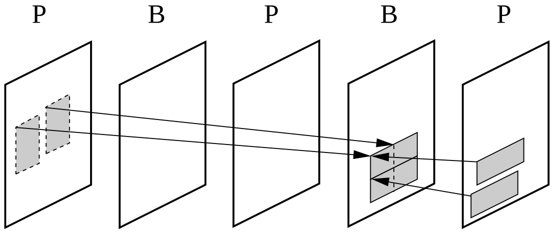

Fig. 2. A bi-directional prediction mode allows a linear combination of one past and one subsequent macroblock prediction signal.

For bi-directional prediction, independently estimated forward and backward prediction signals are practical but the efficiency can be improved by joint estimation. For multi-hypothesis prediction in general, a joint estimation of two hypotheses is necessary [12]. An independent estimate might even deteriorate the performance.

Multi-hypothesis prediction removes the restriction of bi-directional prediction allowing only linear combinations of forward and backward pairs. The additional combinations (forward, forward) and (backward, backward) are obtained by extending a unidirectional picture reference syntax element to a bi-directional picture reference syntax element (Fig. 3). With this picture reference element, a generic prediction signal, which we call hypothesis, can be formed with the syntax fields for reference frame, block size, and motion vector data.

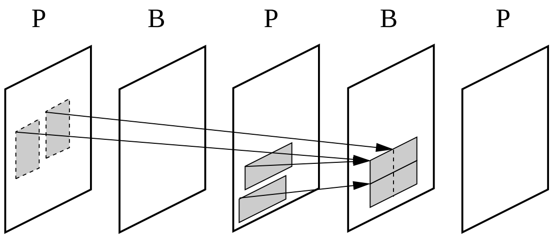

Fig. 3. The multi-hypothesis mode also allows a linear combination of two past macroblock prediction signals.

The multi-hypothesis mode includes the bi-directional prediction mode when the first hypothesis originates from a past reference picture and the second from a future reference picture. The bi-directional mode limits the set of possible reference picture pairs. Not surprisingly a larger set of reference picture pairs improves the coding efficiency of B-Pictures.

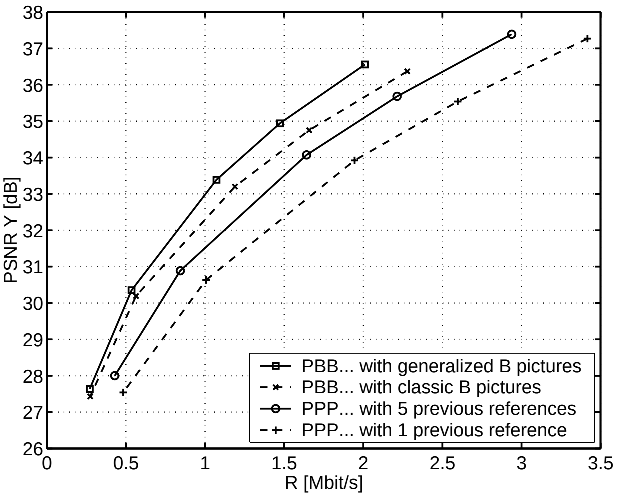

Fig. 4. Luminance PSNR vs. bit-rate for the CIF video sequence *Mobile & Calendar* (30 fps) compressed with JVT/H.26L. A P-Frame is followed by 2 generalized B-Frames. Generalized B-Frames with multi-hypothesis prediction and several reference frames outperform classic P-Frames with bi-directional prediction from the most previous and subsequent P-Frames.

Fig. 4 compares the video quality of generalized B-Pictures to classic B-Pictures for the CIF video sequence *Mobile & Calendar*. The generalized B-Pictures utilize multi-hypothesis prediction with up to two hypotheses, 5 previous and 3 subsequent reference frames. The classic B-Pictures are based on bi-directional prediction from the previous and the subsequent P-Picture. For this experiment, two B-Pictures follow a P-Picture. When comparing this to display-order P-Picture encoding, it turns out that out-of-display-order encoding is still more efficient, even when P-Picture encoding utilizes multi-frame motion-compensated prediction.

The concept of generalized B-Pictures separates picture reference selection and linear combination of prediction signals. For example, generalized B-Pictures with forward-only prediction may be utilized like P-Pictures with the advantage of linearly combined prediction signals without extra coding delay.

4. COMPARISON BETWEEN JVT/H.26L AND MPEG-2

With MPEG-2 being the predominant video compression standard for digital set-top boxes currently, it is interesting to compare its compression efficiency to that of the emerging JVT/H.26L standard. Such a comparison has been carried out by ITU-T VCEG, and example results are shown in Fig. 5.

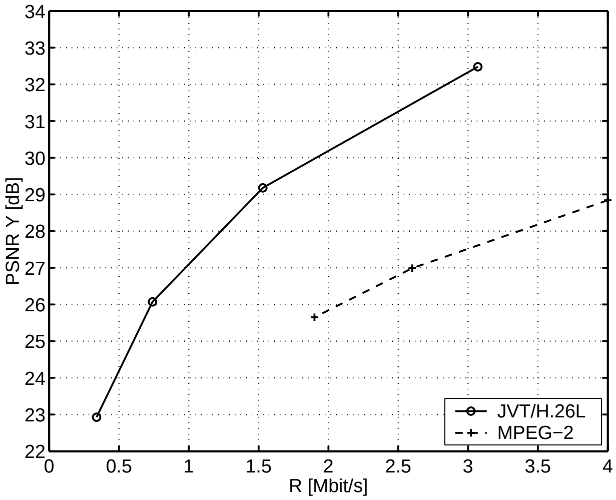

Fig. 5. Luminance PSNR vs. bit-rate for the HHR interlaced video sequence *Mobile & Calendar* (352×480) compressed with JVT/H.26L and MPEG-2. Results are reported in [13] and [14].

Unlike the experiments previously reported in this paper, an interlaced video sequence (*Mobile & Calendar*) was compressed. With the exception of DVD movies, nearly all entertainment video for television sets is still in interlaced format, and MPEG-2 has been extensively optimized for interlace. At the time that the results reproduced in Fig. 5 were obtained, JVT/H.26L was not yet optimized for interlace, but it still achieved a bit-rate 60% lower than that for MPEG-2 at the same PSNR [13]. The results for JVT/H.26L were obtained by field coding [14]. Further improvement will be achieved by adaptive frame/field selection on the picture and macroblock level. Also, note that these results do not yet include generalized B-Pictures.

5. CONCLUSIONS

Multi-frame motion-compensated prediction improves compression efficiency and is rapidly becoming part of the capabilities of modern video compression standards. As always, when new techniques require more memory and computation, there will be predictable objections from hardware manufacturers, but in the end, Moore's Law wins and performance takes priority over complexity. As Gary Sullivan, the leader of the JVT/H.26L standardization effort has formulated it: *"What once seemed like a strange and wasteful idea of requiring storage and searching of extra old pictures is becoming accepted practice – indeed it is the previous practice of throwing away the old decoded picture that has started to seem wasteful."* [15]

We expect multi-frame techniques to first appear in digital set-top boxes for video applications over the Internet and for applications that benefit from efficient compression, but do not have to adhere to MPEG-2, e.g., built-in magnetic disk storage or home networking. Once set-top boxes are programmable with the new format, it might be tempting to make better use of the bit-rate by using JVT/H.26L for digital broadcasting instead of MPEG-2.

6. REFERENCES

[1] P. Lyman and Hal R. Varian, "How much information," 2000, Retrieved from http://www.sims.berkeley.edu/how-much-info on April 26, 2002.

[2] K. G. Coffman and A. M. Odlyzko, "Internet growth: Is there a "Moore's Law" for data traffic," 2001, Retrieved from http://www.research.att.com/ ˜amo/doc/internet.moore.pdf on April 26, 2002.

[3] ITU-T Video Coding Experts Group and ISO/IEC Moving Picture Experts Group, *Working Draft Number 2, Revision 7*, Apr. 2002, ftp:// ftp.imtc-files.org/ c:/ inetpub/ ftpsites/ imtc/ jvt-experts/ draft_standard/ jwd2r7.zip.

[4] T. Wiegand, X. Zhang, and B. Girod, "Long-Term Memory Motion-Compensated Prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, Feb. 1999.

[5] M. Budagavi and J. Gibson, "Multiframe block motion compensated video coding for wireless channels," in *Thirtieth Asilomar Conference on Signals, Systems and Computers*, Nov. 1996, vol. 2, pp. 953–957.

[6] ITU-T, *Recommendation H.263 (Video Coding for Low Bitrate Communication) Annex N*, 1997.

[7] T. Wiegand, X. Zhang, and B. Girod, "Block-Based Hybrid Video Coding Using Motion-Compensated Long-Term Memory Prediction," in *Proceedings of the Picture Coding Symposium*, Berlin, Germany, Sept. 1997, pp. 153–158.

[8] ITU-T, *Recommendation H.263 (Video Coding for Low Bitrate Communication) Annex U*, 1999.

[9] M. Flierl, T. Wiegand, and B. Girod, "Rate-Constrained Multi-Hypothesis Motion-Compensated Prediction for Video Coding," in *Proceedings of the IEEE International Conference on Image Processing*, Vancouver, Canada, Sept. 2000, vol. III, pp. 150–153.

[10] M. Flierl, T. Wiegand, and B. Girod, "Multihypothesis pictures for H.26L," in *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, Oct. 2001.

[11] M. Flierl and B. Girod, "Multihypothesis Predition for B frames," Document VCEG-N40, ITU-T Video Coding Experts Group, Sept. 2001, http:// standards.pictel.com/ ftp/ video-site/ 0109_San/ VCEG-N40.doc.

[12] M. Flierl and B. Girod, "Multihypothesis Motion Estimation for Video Coding," in *Proceedings of the Data Compression Conference*, Snowbird, Utah, Mar. 2001, pp. 341–350.

[13] P. Borgwardt, *Handling Interlaced Video in H.26L*, ITU-T Video Coding Experts Group, Sept. 2001, http:// standards.pictel.com/ ftp/ video-site/ 0109_San/ VCEG-N57.doc.

[14] M. Gallant, L. Winger, and G. Côté, *Interlaced Field Coding Core Experiment*, ITU-T Video Coding Experts Group, Dec. 2001, http:// standards.pictel.com/ ftp/ video-site/ 0112_Pat/ VCEG-O40.doc.

[15] T. Wiegand and B. Girod, *Multi-Frame Motion-Compensated Prediction for Video Transmission*, Kluwer Academic Publishers, 2001.