WYNER-ZIV VIDEO CODING WITH HASH-BASED MOTION COMPENSATION AT THE RECEIVER

Anne Aaron, Shantanu Rane and Bernd Girod

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

ABSTRACT

In current interframe video compression systems, the encoder performs predictive coding to exploit the similarities of successive frames. The Wyner-Ziv Theorem on source coding with side information available only at the decoder suggests that an asymmetric video codec, where individual frames are encoded separately, but decoded conditionally (given temporally adjacent frames) could achieve similar efficiency. In previous work we propose a Wyner-Ziv coding scheme for motion video that uses intraframe encoding, but interframe decoding. In this paper we improve on our Wyner-Ziv video codec by sending hash codewords of the current frame to aid the decoder in accurately estimating the motion. This allows us to implement a low-delay system where only the previous reconstructed frame is used to generate the side information of a current frame. Simulation results show significant gains above conventional DCT-based intraframe coding. The Wyner-Ziv video codec with hash-based motion compensation at the receiver enables low-complexity encoding while achieving high compression efficiency.

1. INTRODUCTION

Implementations of current video compression standards, such as the ISO MPEG schemes or the ITU-T recommendations H.263 and H.264 require much more computation for the encoder than for the decoder; typically the encoder is 5 to 10 times more complex than the decoder. This asymmetry in complexity is well-suited for broadcasting or for streaming video-on-demand systems where video is compressed once and decoded many times. However, some applications may require the dual system, i.e., low-complexity encoders, possibly at the expense of high-complexity decoders. This is certainly the case for wireless mobile terminals with a built-in camera that possess the capability to either store compressed video or send it to the fixed part of the network. Examples of such systems include wireless video sensors for surveillance, wireless PC cameras, mobile camera phones, and future networked camcorders. For these applications, compression must be implemented at the camera where memory and computation are scarce.

To achieve low-complexity encoding, we propose an asymmetric video compression scheme where individual frames are encoded independently (*intraframe encoding*) but decoded conditionally (*interframe decoding*). Two results from information theory suggest that an intraframe encoder - interframe decoder system can approach the efficiency of an interframe encoder-decoder system. Consider two statistically dependent discrete signals, X and Y, which are compressed using two independent encoders but are decoded by a joint decoder. The Slepian-Wolf Theorem on distributed source coding states that even if the encoders are independent, the achievable rate region for probability of decoding error to approach zero is $R_X \geq H(X|Y)$, $R_Y \geq H(Y|X)$ and $R_x + R_y \geq H(X,Y)$ [1]. The counterpart of this theorem for lossy source coding is Wyner and Ziv's work on source coding with side information [2]. Let X and Y be statistically dependent Gaussian random processes, and let Y be known as side information for encoding X. Wyner and Ziv showed that the conditional Rate-Mean Squared Error Distortion function for X is the same whether the side information Y is available only at the decoder, or both at the encoder and the decoder. We refer to lossless distributed source coding as Slepian-Wolf coding and lossy source coding with side information at the decoder as Wyner-Ziv coding.

We call our proposed intraframe encoder-interframe decoder system a *Wyner-Ziv video codec*. A Wyner-Ziv video encoder has great cost advantage, since it compresses each video frame independently, requiring only intraframe processing. The corresponding decoder, in the fixed part of the network, exploits the statistical dependence between frames, by much more complex interframe processing. A similar video compression system, using distributed source coding principles, was proposed independently by Puri and Ramchandran [3, 4, 5]. Sehgal et al. also propose Wyner-Ziv coding for a state-free causal video encoder [6].

We first described the pixel-domain version of our system in [7] where Wyner-Ziv coding is applied to the even frames of a video sequence and the odd frames are known as side information at the decoder. In [8], we present a more general framework. The *key frames* of the video sequence are compressed using a conventional intraframe codec. The remaining frames, the *Wyner-Ziv frames*, are intraframe encoded using a Wyner-Ziv encoder. To decode a Wyner-Ziv frame, previously decoded frames (both key frames and Wyner-Ziv frames) are used to generate the side information which is an estimate of the Wyner-Ziv frame to be decoded. In [9], we extend the Wyner-Ziv video codec to a transform-domain codec. The spatial transform enables

This work is supported in part by the National Science Foundation Grant No. CCR-0310376 and a C.V. Starr Southeast Asian Fellowship.

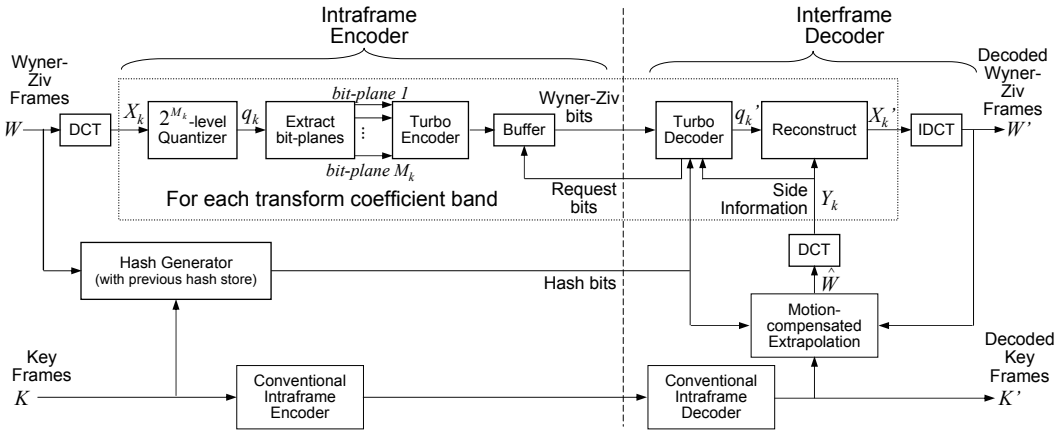

Fig. 1. Wyner-Ziv video codec with hash-based motion compensation at the decoder

the codec to exploit the statistical dependencies within a frame, thus achieving better rate-distortion performance.

To achieve high compression efficiency in a Wyner-Ziv video codec, motion has to be estimated at the decoder. In previous work, we have relied on previously decoded frames to either interpolate or extrapolate the motion without considering the current frame. Conventional motion-compensated coding, however, benefits from extracting the best motion information by directly comparing the current frame with one or more reference frames. The analogous approach for Wyner-Ziv video coding requires *joint decoding and motion estimation*, using the Wyner-Ziv bits, and possibly additional helper information from the encoder. The CRC bits in the system proposed by Puri and Ramchandran [3, 4, 5] is an example of helper information for joint decoding and motion estimation. At the decoder, the CRC of a block is used to choose from many decoded versions of the block, with each version corresponding to a different motion vector.

In this work we propose to send robust *hash codewords* from the encoder, in addition to the Wyner-Ziv bits, to aid the decoder in estimating the motion and generate the side information. These hash bits serve to relay the motion information to the decoder without actually estimating the motion at the encoder. Since the hash bits can enable more accurate motion compensation using only one previous frame, we can implement a low-delay system where only the previous reconstructed frame is used to generate the side information of a current Wyner-Ziv frame. This is analogous to the I-P-P structure used in conventional interframe video coding. In [8], the number of Wyner-Ziv frames in between key frames was limited to a few frames and the side information was generated using interpolation, thus, requiring out-of-order decoding. The improved system allows longer group of pictures (GOP) as well as sequential decoding.

In Section 2, we describe the proposed Wyner-Ziv video codec and the use of the hash bits. In Section 3, we compare the performance of the proposed codec to conventional intraframe coding and conventional interframe predictive coding, using a standard H.263+ video coder.

2. WYNER-ZIV VIDEO CODEC

We propose an intraframe encoder and interframe decoder system for video compression as shown in Fig. 1. A subset of frames from the sequence are designated as key frames. The key frames, K, are encoded and decoded using a conventional intraframe codec. In between the key frames are Wyner-Ziv frames, W, which are intraframe encoded but interframe decoded.

2.1. Intraframe Encoder

At the encoder, a blockwise DCT is applied to the Wyner-Ziv frame W to generate X. The transform coefficients are grouped together to form coefficient bands X_k, where k denotes the coefficient number. Each transform coefficient band is then encoded independently.

For each band X_k, the coefficients are quantized using a uniform scalar quantizer with 2^{M_k} levels. The quantized symbols, q_k, are converted to fixed-length binary codewords, and corresponding bit-planes are blocked together forming M_k bit-plane vectors. Each bit-plane vector is then sent to the Slepian-Wolf encoder. The Slepian-Wolf coder is implemented using a rate-compatible punctured turbo code (RCPT) [10][11]. The RCPT, combined with feedback, provides rate flexibility which is essential in adapting to the changing statistics between the side information and the frame to be encoded. The parity bits produced by the turbo encoder are stored in a buffer which transmits a subset of these parity bits to the decoder upon request. The parity bits sent from the encoder buffer constitute the Wyner-Ziv bits.

The encoder also generates hash bits to aid the decoder in estimating the motion. In the current implementation, the robust hash code for an image block simply consists of a small subset of the quantized DCT coefficients of the block. Since the hash is much smaller than the original data, the encoder is allowed to keep the hash codewords for the previous frame in a small hash store. For each block of the current Wyner-Ziv frame, the distance from the corresponding hash of the previous frame is calculated. If the distance is smaller than a threshold, a "no hash bits" codeword is sent. If the distance exceeds the threshold, the block's hash

is sent, along with the Wyner-Ziv bits. Strictly speaking, the encoder is no longer an intraframe coder because of the hash store. However, storing hash codewords of the previous frame is a negligible burden, compared to conventional frame store and encoder-based motion estimation.

The proposed codec has an encoder complexity similar to that of conventional intraframe encoding. For the Wyner-Ziv frames, turbo coding (composed of interleaving and convolutional coding) replaces conventional entropy coding. The generation of the hash information requires minimal memory and computation.

2.2. Interframe Decoder

The hash codewords enable us to implement a low-delay system where only the previous reconstructed frame and the current hash bits are used to generate the side information of the current Wyner-Ziv frame. For a given block of W, if no hash codeword is sent, the co-located block is used as side information. If hash bits are sent, the decoder performs a motion search based on the hash to generate the best side information block from the previous frame. This process generates side information, $\hat{W}$, which is an estimate of W.

The decoder applies a blockwise DCT on $\hat{W}$ to generate Y. The transform coefficients from Y are grouped together to form coefficient bands Y_k, the side information corresponding to X_k. To be able to use Y_k at the turbo decoder and reconstruction block, the decoder assumes a statistical dependence model between X_k and Y_k.

Given a coefficient band, the turbo decoder successively decodes the bit-planes starting with the most significant bit-plane. It takes the received subset of parity bits corresponding to the bit-plane and the side information Y_k to decode the current bit-plane. If the decoder cannot reliably decode the bits, it requests additional parity bits from the encoder buffer through feedback. The request and decode process is repeated until an acceptable probability of bit error is guaranteed. The probabilities generated for the current bit-plane are used for decoding the less significant bit-planes. By using the side information Y_k and successively decoding the bit-planes, the decoder needs to request $R_k \leq M_k$ bits to decode which of the 2^{M_k} bins a transform coefficient belongs to and so compression is achieved. For the current hash implementation, the quantized coefficients in the hash code can also fix the corresponding probabilities in the turbo decoder, thus further reducing the rate needed for the parity bits. This approach is closely related to the idea of probing the dependence channel for universal Slepian-Wolf coding, proposed in [12].

When all the bit-planes are decoded, the bits are regrouped and the quantized symbol stream is reconstructed as $q_k{}'$. The reconstructed coefficient band $X_k{}'$ is calculated as $E(X_k|q_k{}', Y_k)$. Assuming that $q_k{}'$ is error-free, this reconstruction function has the advantage of bounding the magnitude of the reconstruction distortion to a maximum value, determined by the quantizer coarseness. This property is desirable since it eliminates large positive or negative errors for a given transform coefficient. These large errors tend to be very perceptible and annoying to the viewer. The inverse-DCT is then applied to the reconstructed coefficient bands.

3. SIMULATION RESULTS

We implemented the Wyner-Ziv video codec proposed in Section 2 and assessed its performance for QCIF video sequences.

For encoding a Wyner-Ziv frame, we use a 4×4 DCT and each coefficient band is quantized with a uniform scalar quantizer. We use the same step size for all the coefficient bands. The number of quantizer bins coded for each band determines the bit allocation between bands.

The turbo encoder is composed of two identical constituent convolutional encoders of rate $\frac{1}{2}$ and generator matrix $[1 \ \frac{1+D+D^3+D^4}{1+D^3+D^4}]$ [10]. The parity bits from the convolutional encoder are stored in the encoder buffer while the systematic bits are discarded. The simulation set-up assumes ideal error detection at the decoder – the decoder can determine whether the current bit-plane error rate, P_e, is greater than or less than 10^{-3}. If $P_e \geq 10^{-3}$ it requests for additional parity bits.

The turbo decoder and reconstruction block assume a Laplacian residual distribution between X_k and Y_k. Let d be the difference between corresponding elements in X_k and Y_k. We observe that the distribution of d can be approximated as $f(d) = \frac{\alpha}{2} e^{-\alpha|d|}$. Each coefficient band has a different α parameter which was approximated by training from various sequences.

We generate a hash codeword, consisting of quantized DCT coefficients, for each 4×4 block of the frame. As explained in Section 2.1 we do not send the hash codeword of every block. Instead we calculate the distance of the block from the co-located hash of the previous frame. For this hash code implementation, storing the codewords corresponding to the previous frame requires a hash storage the size of about 20% the original frame. In the simulations we send about 5% to 20% of the hash codewords, depending on the sequence and the quantization parameters.

The results for the *Salesman* and *Hall Monitor* QCIF sequences at 10 fps are shown in Fig. 2 and 3. We varied the number of Wyner-Ziv frames between key frames (resulting in different GOP lengths) to generate the different Wyner-Ziv plots. The key frames were encoded as I frames using a standard H.263+ codec. For each GOP length, we changed the quantization parameter of the key frames as well as the quantization of the Wyner-Ziv frames. The plots show the total rate and the average frame PSNR of both the key frames and Wyner-Ziv frames. We compare the rate-distortion performance to that of conventional DCT-based intraframe coding and H.263+ interframe coding (I-P-P) with the same GOP size. The H.263+ interframe coding plots were generated by choosing the best combination of quantization parameters for the I and P frames.

From the plots we observe impressive gains (up to 9 dB) over conventional intraframe DCT coding for all the GOP lengths. There is a performance gap relative to H.263+ interframe coding, with the gap widening for longer GOP's. For the *Salesman* and *Hall Monitor* sequences, the gap from H.263+ interframe coding ranges from 1 to 4 dB. It can be seen from the results that increasing the GOP size does not necessarily improve the compression performance of the Wyner-Ziv video codec. With more Wyner-Ziv frames between key frames, we reduce the bit rate for the key frames

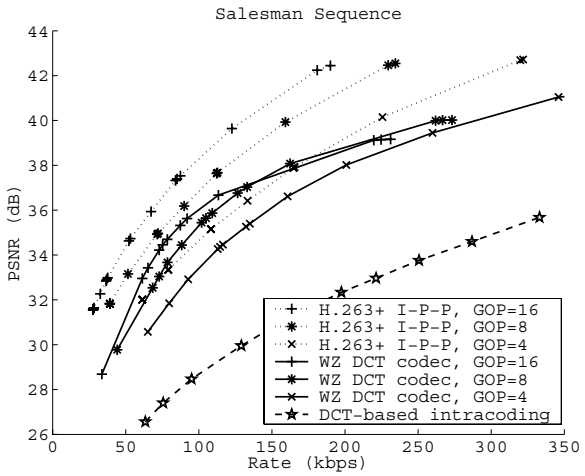

Fig. 2. Bit rate vs. Ave. Frame PSNR for *Salesman*

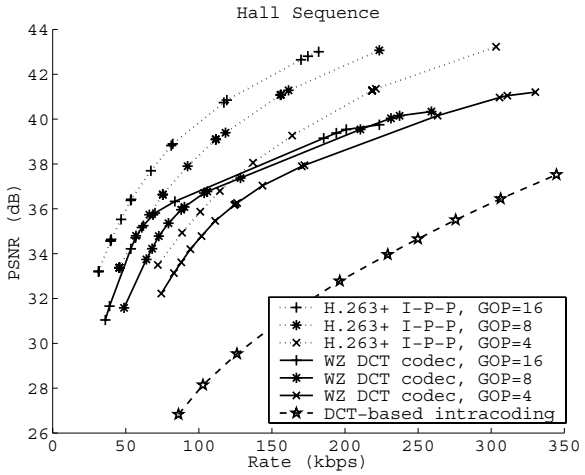

Fig. 3. Bit rate vs. Ave. Frame PSNR for *Hall Monitor*

but we increase the distortion propagation from a decoded frame to the side information of the next Wyner-Ziv frame.

4. CONCLUSIONS

In this work we propose to send robust hash codewords in a Wyner-Ziv video codec to achieve more accurate decoder motion estimation. This improvement enables us to implement a low-delay system which recursively decodes a series of Wyner-Ziv frames by performing hash-based motion compensation of the previous frame to generate the side information. This is analogous to the I-P-P structure used in conventional interframe video coding. Note that the I-P-P dependency is only meaningful at the decoder because the frames are still encoded independently at the encoder.

The Wyner-Ziv video codec shows impressive gains over conventional DCT-based intraframe coding while having comparable encoding complexity. There is still a performance gap from H.263+ interframe coding especially at larger GOP lengths.

For Wyner-Ziv coding, the optimal bit rate is determined by the statistical dependence between the source and the side information. Rate control is a special challenge since the side information is exploited only at the decoder but not at the encoder. At present our rate control depends entirely on a feedback channel from the decoder to the encoder. In future work we can compare the stored hash codewords of the previous frame and those of the current frame to estimate the statistics and the resulting bit rate.

5. REFERENCES

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.

[2] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[3] R. Puri and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles," in *Proc. Allerton Conference on Communication, Control, and Computing*, Allerton, IL, Oct. 2002.

[4] R. Puri and K. Ramchandran, "PRISM: An uplink-friendly multimedia coding paradigm," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, Apr. 2003.

[5] R. Puri and K. Ramchandran, "PRISM: A 'reversed' multimedia coding paradigm," in *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.

[6] A. Sehgal, A. Jagmohan, and N. Ahuja, "A causal state-free video encoding paradigm," in *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.

[7] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Proc. Asilomar Conference on Signals and Systems*, Pacific Grove, CA, Nov. 2002.

[8] A. Aaron, E. Setton, and B. Girod, "Towards practical Wyner-Ziv coding of video," in *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.

[9] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *Proc. SPIE Visual Communications and Image Processing*, San Jose, CA, Jan. 2004.

[10] D. Rowitch and L. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo codes," *IEEE Transactions on Communications*, vol. 48, no. 6, pp. 948–959, June 2000.

[11] A. Aaron and B. Girod, "Compression with side information using turbo codes," in *Proc. IEEE Data Compression Conference*, Snowbird, UT, Apr. 2002, pp. 252–261.

[12] J. García-Frías, "Compression of correlated binary sources using turbo codes," *IEEE Communications Letters*, vol. 5, no. 10, pp. 417–419, Oct. 2001.