Transform-domain Wyner-Ziv Codec for Video

Anne Aaron, Shantanu Rane, Eric Setton, and Bernd Girod

Information Systems Laboratory, Department of Electrical Engineering
Stanford University
350 Serra Mall, Stanford, CA 94305

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. We propose a transform-domain Wyner-Ziv coding scheme for motion video that uses intraframe encoding, but interframe decoding. In this system, the transform coefficients of a Wyner-Ziv frame are encoded independently using a scalar quantizer and turbo coder. The decoder uses previously reconstructed frames to generate side information to conditionally decode the Wyner-Ziv frames. Simulation results show significant gains above DCT-based intraframe coding and improvements over the pixel-domain Wyner-Ziv video coder.

Keywords: Wyner-Ziv coding, distributed source coding, video compression, transform coding, low-complexity encoders, Slepian-Wolf, turbo codes

1. INTRODUCTION

Current video compression standards perform interframe predictive coding to exploit the similarities among successive frames. Since predictive coding makes use of motion estimation, the video encoder is typically 5 to 10 times more complex than the decoder. This asymmetry in complexity is desirable for broadcasting or for streaming video-on-demand systems where video is compressed once and decoded many times. However, some future systems may require the dual scenario. For example, we may be interested in compression for mobile wireless cameras uploading video to a fixed base station. Compression must be implemented at the camera where memory and computation are scarce. For this type of system what we desire is a low-complexity encoder, possibly at the expense of a high-complexity decoder, that nevertheless compresses efficiently.

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 come close to 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.

In [3], we apply Wyner-Ziv coding to the pixel values of a video signal. The even frames of the sequence are compressed by an intraframe encoder that does not know the odd frames. The compressed stream is sent to a decoder which uses the odd frames as side information to conditionally decode the even frames. This intraframe encoder - interframe decoder system can be extended to a more practical and general framework as described

Send correspondence to Anne Aaron: amaaron@stanford.edu; phone 1-650-723-3476; fax 1-650-724-3648

in [4]. A subset of frames from the video sequence are designated as *key frames* which 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 side information. Interframe decoding of the Wyner-Ziv frames is performed by exploiting the inherent similarities between the Wyner-Ziv frame and the side information.

A similar video compression system, using distributed source coding principles, was proposed independently by Puri and Ramchandran [5][6]. Sehgal et al. also propose Wyner-Ziv coding for a state-free causal video encoder [7].

In this work, we extend the Wyner-Ziv video codec outlined in [3] and [4], to a transform-domain Wyner-Ziv coder. The spatial transform enables the codec to exploit the statistical dependencies within a frame, thus, achieving better rate-distortion performance. We also apply different methods of generating the side information at the decoder and investigate how the decoder complexity affects the compression performance.

In Section 2, we describe the proposed Wyner-Ziv video codec. In Section 3, we discuss the simulation details and compare the performance of the proposed coder to DCT-based intraframe coding and to conventional interframe 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.

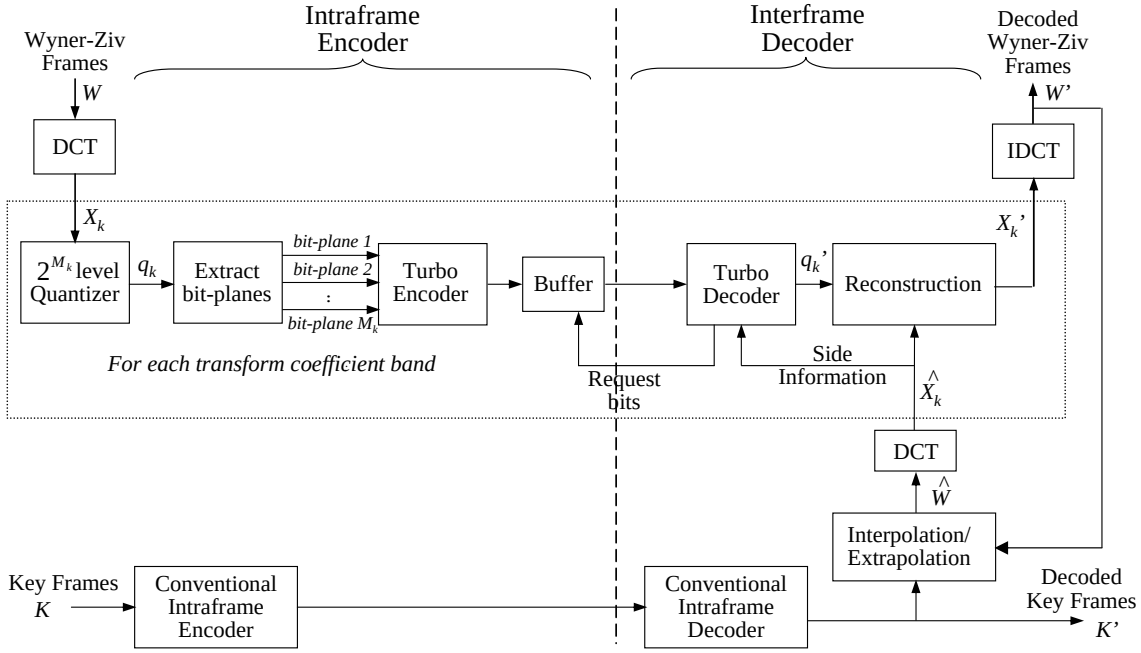

Figure 1. Transform-domain Wyner-Ziv video codec with intraframe encoding and interframe decoding. A turbo encoder-turbo decoder pair is used as a Slepian-Wolf codec.

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) [8][9]. 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.

2.2. Interframe Decoder

For each W, the decoder takes previously reconstructed frames to form the side information, $\hat{W}$, which is an estimate of W. In this work we assume alternating key frames and Wyner-Ziv frames. For the simulations, we investigate different methods for generating $\hat{W}$ from decoded adjacent frames, as described in Sec. 3.1 and Sec. 3.5.

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

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 $\hat{X_k}$ 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 lower significance bit-planes. By using the side information $\hat{X_k}$ 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.

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 $E(X_k|{q_k}', \hat{X_k})$. 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. W' is finally generated by taking the inverse-DCT of the reconstructed coefficient bands.

2.3. Complexity

The proposed transform-domain 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. Compared to standard motion-compensated predictive encoders, Wyner-Ziv encoding is much less complex since motion estimation and prediction is completely eliminated at the encoder. However, since the temporal dependencies are exploited at the decoder, the compression efficiency can approach that of a conventional interframe encoder - interframe decoder system. The proposed Wyner-Ziv video codec allows low-complexity encoding without sacrificing compression efficiency.

The decoder of the proposed system is more complex than standard intraframe or interframe video decoders. The Wyner-Ziv interframe decoder requires iterative decoding which is computationally more intensive than conventional techniques such as Huffman decoding or arithmetic decoding.

Generation of side information also adds complexity to the decoder. If motion-compensated interpolation or extrapolation techniques are employed at the decoder, the added complexity can be seen as shifting the motion estimation from the encoder to the decoder. This complexity can be reduced by using simpler interpolation or extrapolation methods, as will be discussed is Sec. 3.5, at the expense of compression efficiency.

3. SIMULATION RESULTS

We implemented the proposed intraframe encoder - interframe decoder system and assessed the performance for QCIF video sequences. For the simulations we use a simplified set-up where the odd frames are key frames and the even frames are Wyner-Ziv frames. The key frames are encoded as I frames using a standard H.263+ codec.

3.1. Motion-Compensated Side Information

In the first set of experiments we apply two motion-compensated techniques to generate the side information:

1. Motion-compensated Interpolation (*MC-I*) - The side information for an even frame at time index t is generated by performing motion-compensated interpolation using the decoded key frames at time $t-1$ and $t+1$. This interpolation technique involves symmetrical bidirectional block matching, smoothness constraints for the estimated motion and overlapped block motion compensation. Since the next key frame is needed for interpolation, the frames have to be decoded out-of-order, similar to the decoding of B frames in predictive video coding.
2. Motion-compensated Extrapolation (*MC-E*) - To generate the side information for the even frame at time t, we estimate the motion between the decoded Wyner-Ziv frame at time $t-2$ and the decoded key frame at time $t-1$ using block matching and a smoothness constraint. The estimated motion is extrapolated to time t and the side information is formed by performing overlapped motion compensation using the pixel values from the previous key frame. Since a decoded Wyner-Ziv frame is used for motion estimation, reconstruction errors from the Wyner-Ziv frame can degrade the reliability of the motion compensation. However, unlike MC-I, the frames can be decoded sequentially.

3.2. Quantizers

For encoding a Wyner-Ziv frame W, we use a 4x4 discrete cosine transform (DCT) and each coefficient band is quantized with 2^{M_k} uniform quantization levels, where $2^{M_k} \in \{0, 2, 4, 8, 16, 32, 64, 128, 256\}$. $2^{M_k} = 0$ means that no bits are sent for coefficient band k and the side information $\hat{X}_k$ is used as the reconstruction $X_k{}'$. The combination of quantizers ($\bar{M} = M_1, M_2..., M_{16}$) determines the bit allocation between bands and is an important optimization parameter given a sequence to be encoded. However, for a practical system it is desirable to have a set of $\bar{M}$'s which work well for most sequences.

To design the set of good generic quantizers, we trained on several sequences as follows: First, we fixed the range of the quantizers for each band based on the histogram of all the sequences. Then for each sequence, we encoded the coefficient bands with all possible number of quantization levels. The sequence was decoded using MC-I for side information and the rate-distortion pairs per band given the chosen M_k were recorded. The Lagrangian cost function was applied to determine the sequence-optimized $\bar{M}^{\lambda}$ for different values of λ. By observing the resulting $\bar{M}^{\lambda}$'s for the different sequences, we designed a group of $\bar{M}$'s similar to the optimal ones. The designed set of $\bar{M}$'s used in the simulations are shown in Fig. 2. For each $\bar{M}$, the upper-leftmost value corresponds to the number of quantization levels used for the DC coefficient from the 4x4 DCT. The bottom right-most value corresponds to the highest AC coefficient. The quantizers in Fig. 2 are ordered in decreasing rate and increasing distortion. Using the fixed generic quantizers instead of the optimal set results in at most 0.2 dB drop in the rate-PSNR curve.

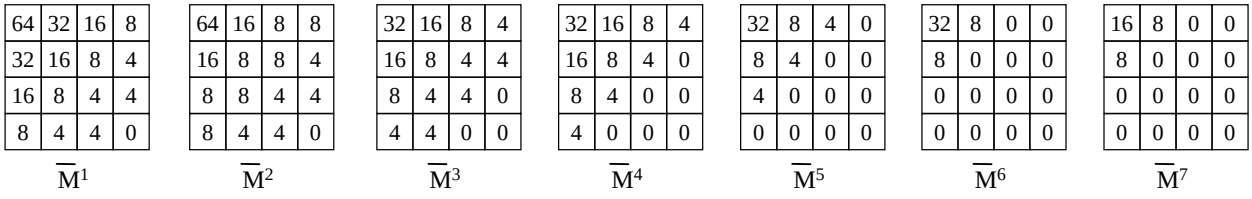

Figure 2. Fixed set of $\bar{M}$'s used for the simulations. An $\bar{M}$ block describes the number of quantization levels used for each coefficient band.

3.3. Turbo Coding

The turbo encoder is composed of two identical constituent convolutional encoders of rate $\frac{1}{2}$ and generator matrix $[1 \frac{1+D+D^3+D^4}{1+D^3+D^4}]$ [8]. The parity bits from the convolutional encoder are stored in the encoder buffer while the systematic part is 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 $\hat{X}_k$. Let d be the difference between corresponding elements in X_k and $\hat{X}_k$. We observed 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 plotting the residual histogram of several sequences using MC-I for the side information.

3.4. Compression Performance

The results for the first 100 frames of *Mother and Daughter* and *Foreman* QCIF sequences are shown in Fig. 3 and Fig. 4. For the plots, we only include the rate and distortion of the luminance of the even frames. The even frame rate is 15 frames per second. We compare our results to (i) DCT-based intraframe coding (the even frames are encoded as I frames) and (ii) H.263+ interframe coding with an I-B-I-B predictive structure, counting only the rate and PSNR of the B frames. We also plot the compression results of the pixel-domain Wyner-Ziv codec. For the pixel-domain results, we eliminate the transform and inverse transform blocks in Fig. 1 and quantize and encode the pixel values directly. This is similar to the system presented in [3] except that in the previous work, turbo coding was performed on the symbol level instead of bit-planes. Bit-plane coding and symbol-based coding of the quantized pixel values yield similar results. Note that the simulations assume the same rate and quality for the odd frames for all the schemes so these values are not included in the plots.

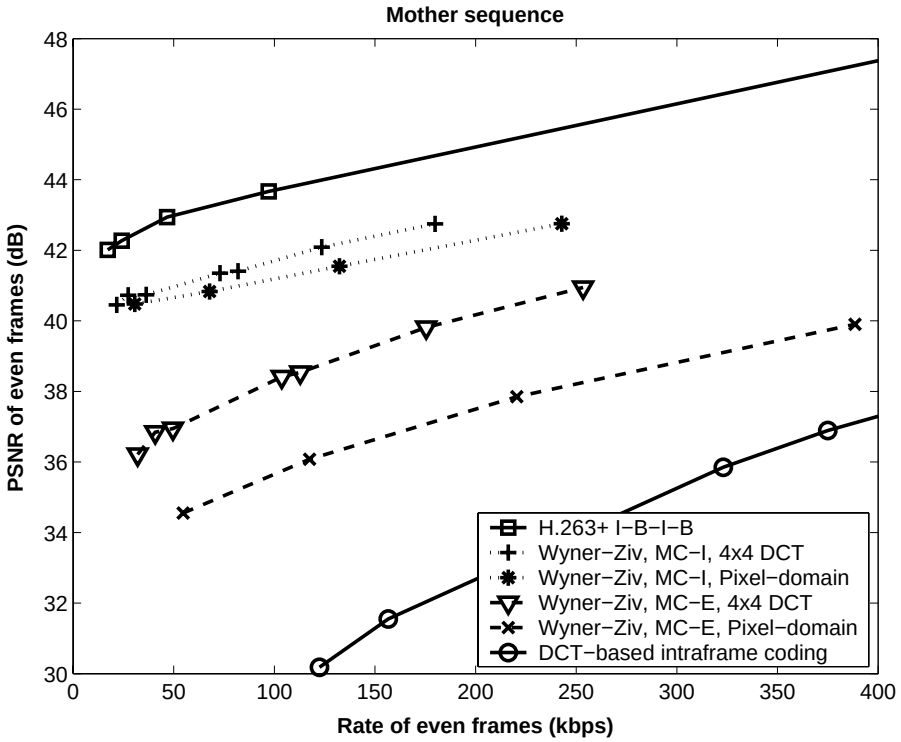

Figure 3. Rate and PSNR comparison of Wyner-Ziv codec vs. DCT-based intraframe coding and H.263+ I-B-I-B coding. *Mother and Daughter* sequence.

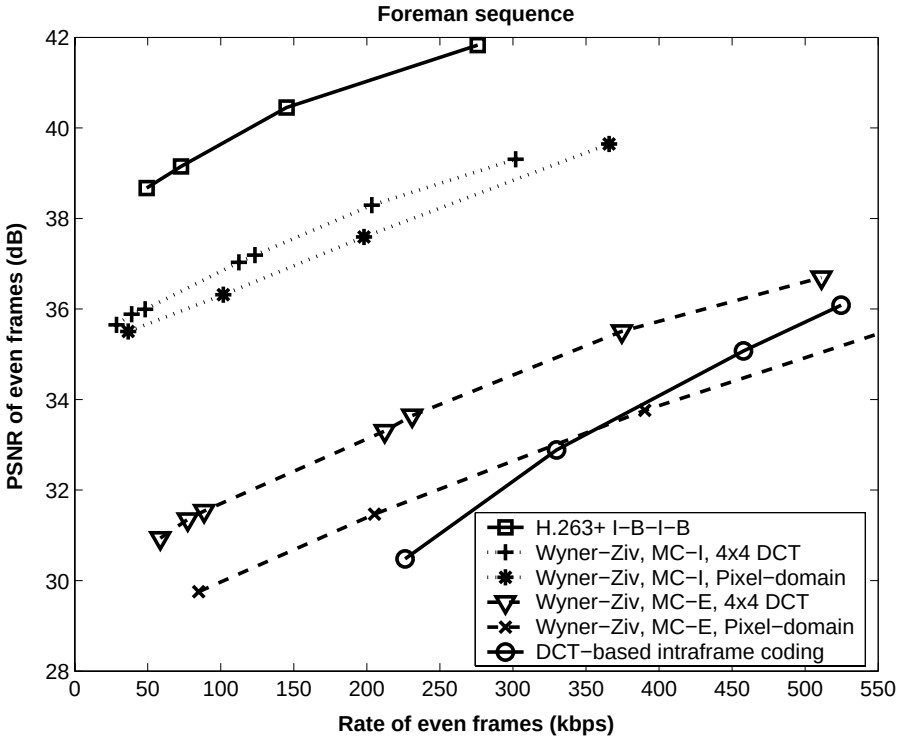

Figure 4. Rate and PSNR comparison of Wyner-Ziv codec vs. DCT-based intraframe coding and H.263+ I-B-I-B coding. *Foreman* Sequence.

(a) (b) (c)

Figure 5. Sample frame for *Foreman*, encoded using (a) DCT-based intraframe coding, (b) DCT-domain Wyner-Ziv coding, using MC-I and (c) H.263+ interframe coding (B frames). Bit-rate for even frames of sequence is 300 kbps.

As it can be observed from the plots, when the side information is highly reliable, such as when MC-I is used, the transform-domain codec is only 0.5 dB better than the pixel-domain Wyner-Ziv codec. With less reliable MC-E, using a transform before encoding results in a 2 to 2.5 dB improvement.

For the lower motion *Mother and Daughter* sequence, the Wyner-Ziv DCT-based codec is about 10 to 12 dB (with MC-I) and 7 to 9 dB (with MC-E) better that DCT-based intraframe coding. The gap from H.263+ interframe coding is 2 dB for MC-I and about 5 dB for MC-E. For *Foreman*, which has more motion and occlusions, the Wyner-Ziv DCT-based codec is 7 to 8 dB (with MC-I) and 1 to 3 dB (with MC-E) better that DCT-based intraframe coding. The gap from H.263+ interframe coding is 2.5 dB for MC-I and about 7 dB for MC-E.

It can be seen that using motion-compensated extrapolation instead of interpolation shifts down the Wyner-Ziv DCT-domain codec rate-PSNR plot by about 3 dB for *Mother and Daughter*. For *Foreman*, the compression efficiency loss is greater (about 5 dB) since the motion and occlusions make it more difficult to extrapolate the succeeding frames.

From visual inspection (Fig. 5), sequences compressed using Wyner-Ziv coding exhibit superior quality compared to DCT-based intraframe coding at the same bit rate. In spite of the PSNR gap (around 3 dB for the sample frames in Fig. 5), the difference in quality between H.263+ interframe coding and Wyner-Ziv coding at the same bit rate is nearly imperceptible.

3.5. Low-Complexity Side Information

We can simplify the interpolation or extrapolation scheme to reduce the decoder complexity at the expense of compression efficiency. In the simulations we used two simplified schemes, which do not employ motion compensation, to generate the side information:

1. Average Interpolation (*Ave-I*) - The side information for the Wyner-Ziv frame is generated by averaging the pixel values from the key frames at $t-1$ and $t+1$.
2. Previous Frame Extrapolation (*Prev-E*) - The previous key frame is used directly as side information.

The simulation results can be seen in Fig. 6 and Fig. 7. For *Mother and Daughter*, eliminating the motion compensation does not reduce the compression performance by a significant amount. By simply using the previous frame as side information we still achieve a 6 to 8 dB gain above DCT-based intraframe coding. However, for *Foreman*, there is a 1 to 2 dB drop in PSNR when motion compensation is not performed. In the case where we use the previous frame as side information for *Foreman*, the Wyner-Ziv codec is better than DCT-based intraframe coding only at lower bit rates.

In Fig. 8 we see how our coding scheme can remove the artifacts introduced by non-sophisticated interpolation techniques. Fig. 8(a) shows the interpolated frame as a result of scheme Ave-I. After Wyner-Ziv coding (Fig. 8(b)), the image is sharpened and the Foreman's facial features are corrected.

4. CONCLUSION

In this paper we propose a transform-domain Wyner-Ziv video codec which uses intraframe encoding and interframe decoding. This type of codec is useful for systems which require simple encoders but can handle more complex decoders. Encoding is composed of a spatial transform, scalar quantization and rate-compatible turbo coding. The decoder generates side information from previously decoded frames and performs turbo decoding using the side information. The proposed system has an encoder complexity similar to current intraframe video coders while coming close to the compression efficiency of interframe coders.

We showed that the transform-domain Wyner-Ziv coder performs 0.5 to 2.5 dB better than our previous pixel-domain implementation. The current system shows significant gains (up to 12 dB) above DCT-based intraframe coding while having the same encoder complexity. The PSNR gain depends on the degree of motion in the sequence and the complexity of the interpolation or extrapolation technique used at the decoder.

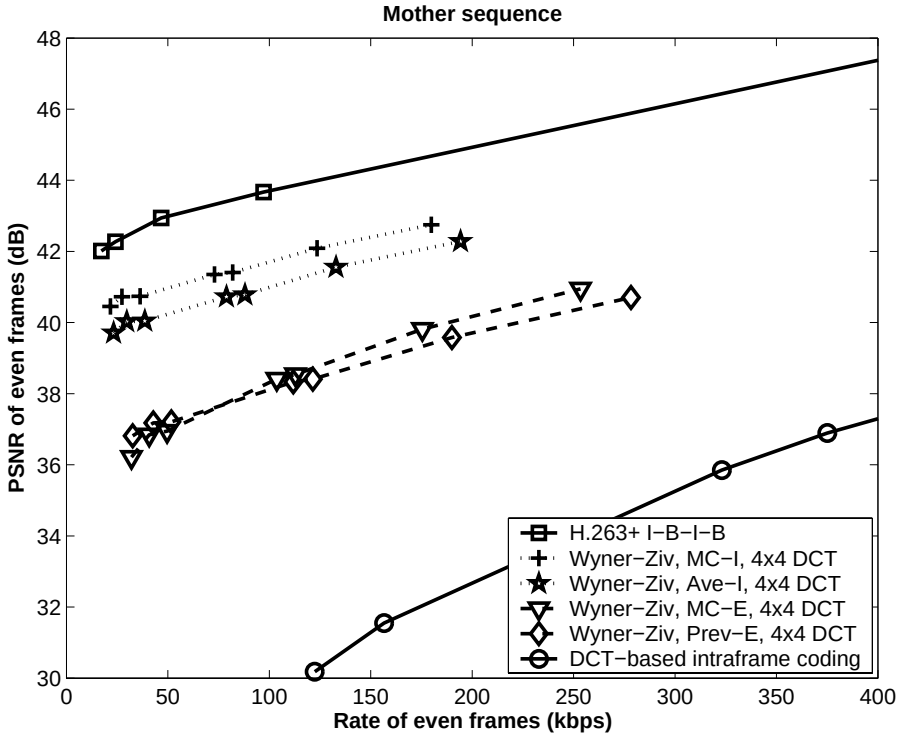

Figure 6. Rate and PSNR comparison for different side information schemes. *Mother and Daughter* Sequence.

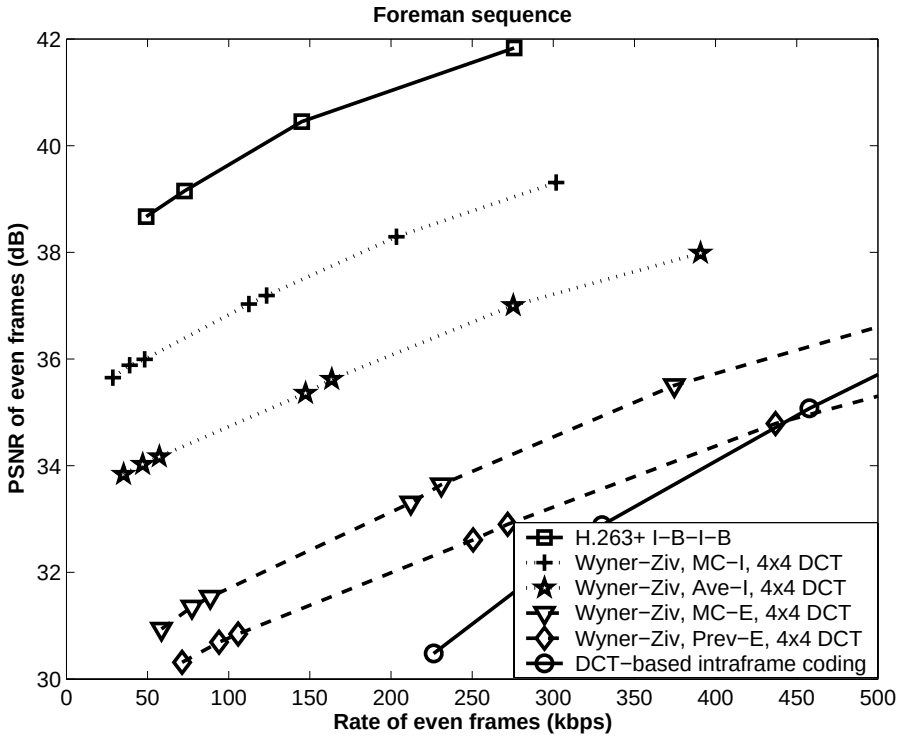

Figure 7. Rate and PSNR comparison for different side information schemes. *Foreman* Sequence.

(a) (b)

Figure 8. (a) Interpolated frame from *Foreman* using Average Interpolation and (b) After Wyner-Ziv coding at 370 kbps

ACKNOWLEDGMENTS

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

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. A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in *Proc. Asilomar Conference on Signals and Systems*, Pacific Grove, California, Nov. 2002.
4. A. Aaron, E. Setton, and B. Girod, "Towards practical Wyner-Ziv coding of video," in *Proc. International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.
5. 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, Illinois, Oct. 2002.
6. R. Puri and K. Ramchandran, "PRISM: A 'reversed' multimedia coding paradigm," in *Proc. International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.
7. A. Sehgal, A. Jagmohan, and N. Ahuja, "A causal state-free video encoding paradigm," in *Proc. International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.
8. 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.
9. A. Aaron and B. Girod, "Compression with side information using turbo codes," in *Proc. IEEE Data Compression Conference*, Snowbird, Utah, Apr. 2002, pp. 252–261.