Title

Low Complexity Spatio-Temporal Key Frame Encoding for Wyner-Ziv Video Coding

Authors

Esmaili, G
Cosman, P C

Publication Date

2009-03-01

Peer reviewed

2009 Data Compression Conference

Low Complexity Spatio-Temporal Key Frame Encoding for Wyner-Ziv Video Coding

Ghazaleh Esmaili and Pamela Cosman
Department of Electrical and Computer Engineering,
9500 Gilman Drive, University of California, San Diego, La Jolla, CA - 92093-0407
Email: {gesmaili,pcosman}@ucsd.edu

Abstract

In most Wyner-Ziv video coding approaches, the temporal correlation of key frames is not exploited since they are simply intra encoded and decoded. In this paper, using the previously decoded key frame as the side information for the key frame to be decoded, we propose new methods of coding key frames in order to improve the rate distortion performance. These schemes which are based on switching between intra mode and Wyner-Ziv mode for a given block or a given frequency band attempt to make use of both spatial and temporal correlation of key frames while satisfying the low complexity encoding requirement of Distributed Video Coding (DVC). Simulation results show that using the proposed methods, one can achieve up to 5 dB improvement over conventional intra coding for relatively low motion sequences and up to 1.3 dB improvement for relatively high motion sequences.

1 Introduction

Traditional video coding standards such as MPEG-x and H.26x which rely on motion compensation algorithms are not suitable for some recent applications such as sensor networks, video surveillance, and mobile camera phones which require many simple and low cost encoders. Wyner-Ziv video coding is based on the Slepian-Wolf [1] and Wyner-Ziv [2] theorems which prove that distributed compression of correlated sources can be as efficient as joint compression. In this scheme, the encoder complexity is shifted to the decoder by encoding individual frames independently (intraframe encoding) but decoding them conditionally (interframe decoding). DVC attempts to achieve the same coding performance as conventional video coding systems.

The first practical implementations of DVC were presented in [3] and [4]. In [3], Puri and Ramchandran introduced a syndrome-based video coding scheme which deployed block-level coding primitives and no feedback was required. It was upgraded to a more practical solution in [5]. In [4], Aaron and Girod proposed a feedback-required Wyner-Ziv video coding algorithm based on Turbo codes in the pixel domain. This technique was extended to the transform domain in [6] to exploit spatial correlation between neighboring pixels, thus achieving better performance. In [7], Brites and Ascenso outperformed [6] by adjusting the quantization step size and applying an advanced frame interpolation for side information generation. Later on, in [8] and [9], enhanced frame interpolation techniques were proposed to achieve better performance. In [10], Jinrong et al. proposed a transform domain classification method to

1068-0314/09 $25.00 © 2009 IEEE
DOI 10.1109/DCC.2009.57

differentiate low motion blocks from high motion blocks to exploit additional video statistics.

In most existing Wyner-Ziv schemes, frames are grouped into two different classes: Wyner-Ziv and key frames. In a GOP size of 2, key frames occur usually every other frame, and Wyner-Ziv frames are the frames in between. Key frames are intra coded using conventional video coding. At the decoder, they are decoded and used to generate side information that is statistically correlated with the Wyner-Ziv frame in between. If the correlation between the Wyner-Ziv frame being encoded and the side information is high, then fewer bits need to be sent from the encoder to the decoder to have a reliable decoding. The statistical dependence between two consecutive key frames is not as high as the statistical dependence between a Wyner-Ziv frame and its corresponding generated side information, because in the latter case the side information comes from frames that are only one frame-time distant, and comes from frames on both sides temporally. Nonetheless, extending the Wyner-Ziv coding method to key frames as well can help to exploit the temporal correlation and improve the rate-distortion performance. We also know that temporal correlation varies for each block. Sometimes temporal correlation between corresponding blocks is so low that intra coding can outperform the Wyner-Ziv technique.

In this paper, we propose two different methods of coding for key frames. In both methods, the previously decoded key frame plays the role of side information for the key frame to be encoded. In the first method, blocks within a frame are classified in three different groups where each one uses a different coding technique. In the second method, frequency bands of each block are divided into two different classes of coding.

The rest of this paper is organized as follows. In Section 2 we briefly review conventional transform domain Wyner-Ziv video coding. In Section 3 the two proposed methods of key frame coding are described in detail. Simulation results and conclusions are presented in Section 4 and Section 5, respectively.

2 Transform domain Wyner-Ziv codec

Fig. 1 shows the transform domain Wyner-Ziv video codec architecture [6]. Key frames are encoded and decoded by a conventional intraframe codec. The frames between them which are Wyner-Ziv frames are intraframe encoded but interframe decoded. A blockwise 4×4 discrete cosine transform (DCT) is applied on Wyner-Ziv frames. X_k is a vector obtained by grouping together the k^{th} DCT coefficient from all blocks. The coefficients of X_k are uniformly quantized to form quantized symbols q_k. After representing the quantized values q_k in binary form, bit planes are extracted and blocked together to form M_k bit plane vectors. Each bit-plane vector is then fed to the Slepian-Wolf encoder. Let W denote a Wyner-Ziv frame, and $\hat{W}$ is the estimate of W generated from previously reconstructed key frames. A blockwise 4×4 DCT is applied on $\hat{W}$ to provide $\hat{X}$. $\hat{X}_k$, the side information corresponding to X_k, is generated by grouping the transform coefficients of $\hat{X}$. In the block diagram of Figure 1, the decoder and reconstruction blocks assume a Laplacian distribution

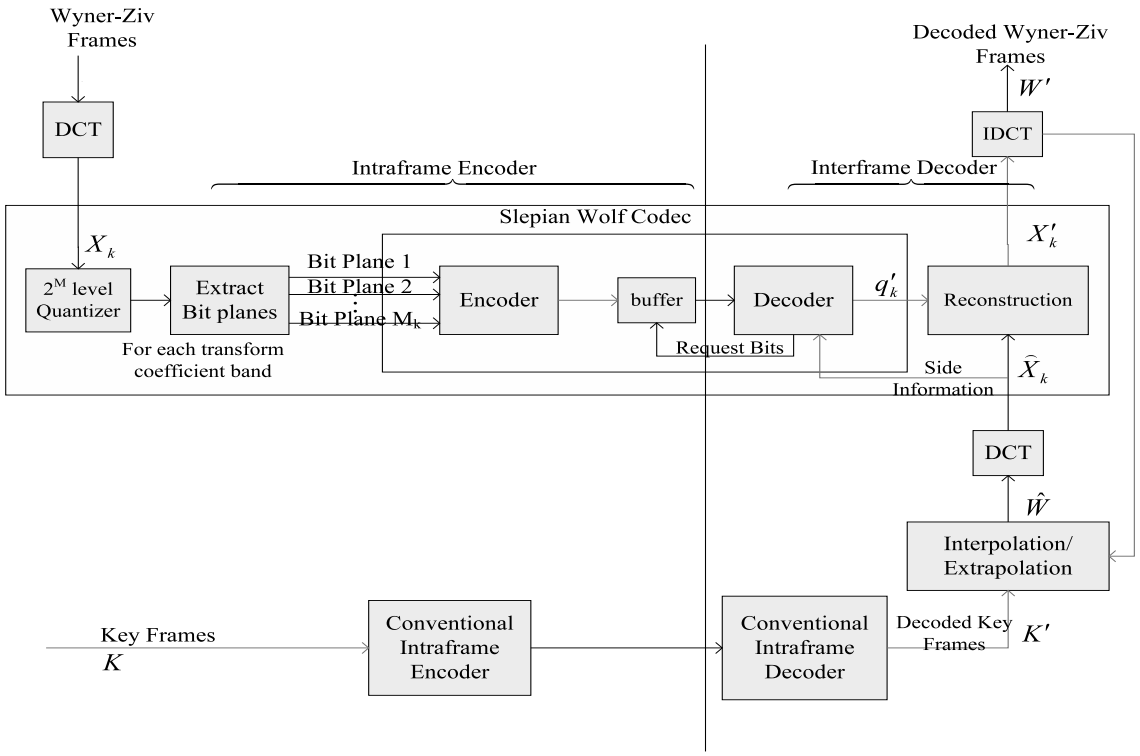

Figure 1: Transform domain Wyner-Ziv video codec [6]

to model the statistical dependency between X_k and $\hat{X}_k$. For each coefficient band, the Slepian-Wolf decoder successively decodes bit-planes beginning from the most significant bit-plane. The received bits of the bit-plane and the side information $\hat{X}_k$ are the decoder tools to decode the current bit-plane. If the decoder can not meet the desired bit error rate, it asks for additional bits through feedback. At the end, the reconstructed coefficient band X'_k is calculated as $E(X_k|q_k, \hat{X}_k)$.

3 The proposed key frame coding methods

In conventional transform domain Wyner-Ziv coding, key frames are encoded and decoded by a conventional intraframe coder. So, the spatial correlation within a block is exploited by applying a Discrete Cosine Transform (DCT) but the temporal correlation between adjacent key frames is not exploited. In this section, we introduce two new methods of encoding for key frames to jointly exploit the spatial and temporal correlation. In both methods, the side information is generated from the previously decoded key frame and the basic idea involves switching between Wyner-Ziv coding and Intra coding in an efficient way. When the spatial correlation is high, Intra coding is more likely to outperform, and when the temporal correlation is high Wyner-Ziv coding is likely the better choice. In this paper, the LDPCA code reported in [11] is adopted for the Slepian-Wolf coding of the Wyner-Ziv codec.

3.1 Key frame coding based on block classification

Blocks within a frame can have very different characteristics and these characteristics have a great impact on the coding performance. Often in a frame there are some regions where blocks are highly correlated with the past due to the low motion activity and some regions where blocks are almost independent of the past because of high motion activity. So, it seems reasonable to differentiate blocks within a frame based on temporal correlation. As depicted in Fig. 2, we define three classes of correlation. Blocks with a very low motion which are highly correlated are classified to the skip class and are not sent to the decoder. They are substituted by the co-located blocks in the previously decoded key frame at the decoder. Low-medium motion blocks which are moderately correlated are classified to the inter class and are coded by Wyner-Ziv codec. High motion blocks which are almost independent are classified to the intra class and are intra coded.

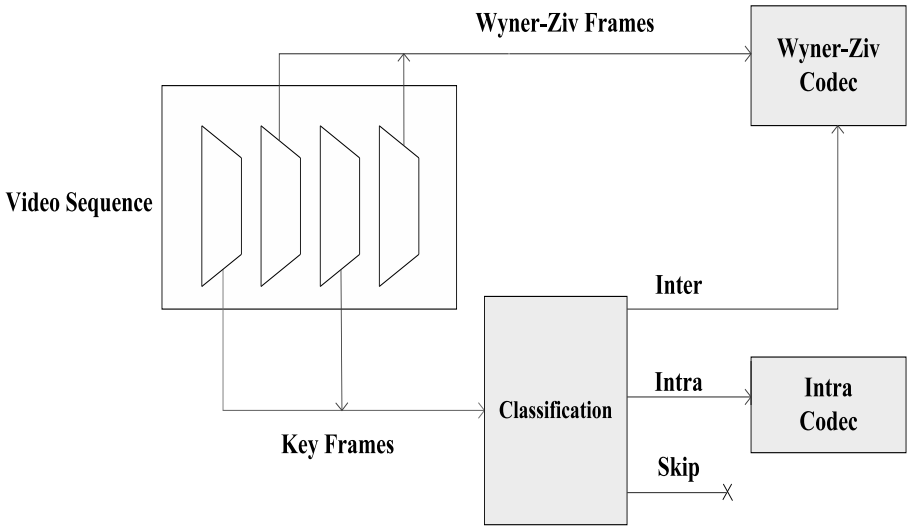

Figure 2: Wyner-Ziv video codec with encoding key frames based on block classification

We now describe how the blocks are classified. The residual energy or the mean squared error between a given block B in the key frame to be encoded, F_t, where t is the time index, and the co-located block B_{ref} in the reconstructed previous key frame at the encoder, $\hat{F}_{t-1}$, is computed by $E = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} [B(x,y) - B_{ref}(x,y)]^2$ where M and N represent the block size (in our case $M = N = 4$). E is a good and very simple candidate for evaluating the temporal correlation of each block. If E is less than a threshold value, T_1, the block is classified to the skip class. If $T_1 \leq E < T_2$ the block belongs to the Wyner-Ziv class and if $E \geq T_2$, it is categorized to the intra class. Since in the Wyner-Ziv codec, the number of bit planes for each frequency band is a fixed value and the encoder lets the decoder know the dynamic range of each frequency band, we can set the threshold value of T_1 based on the quantization step size. If we select $T_1 = \frac{\Delta_{min}^2}{16}$ where Δ_{min} is the minimum of the quantization step sizes over all the frequency bands, then $|B(x,y) - B_{ref}(x,y)|$ will never be greater

than Δ_{min} since $\frac{1}{16}\sum_{x=1}^{M}\sum_{y=1}^{N}[B(x,y)-B_{ref}(x,y)]^2 \leq \frac{\Delta_{min}^2}{16}$. Therefore the distortion penalty of the skip mode is going to be small.

3.2 Key frame coding based on frequency band classification

For the proposed method in Section 3.1, the encoder must tell the decoder the class of each block. For the sequences where high motion blocks are dominant, the number of bits which are saved by exploiting the temporal correlation turns out empirically to be less than or comparable to the number of bits spent in transmitting the classification information. A solution to this problem would be to have a classification approach which does not need to be conveyed to the decoder.

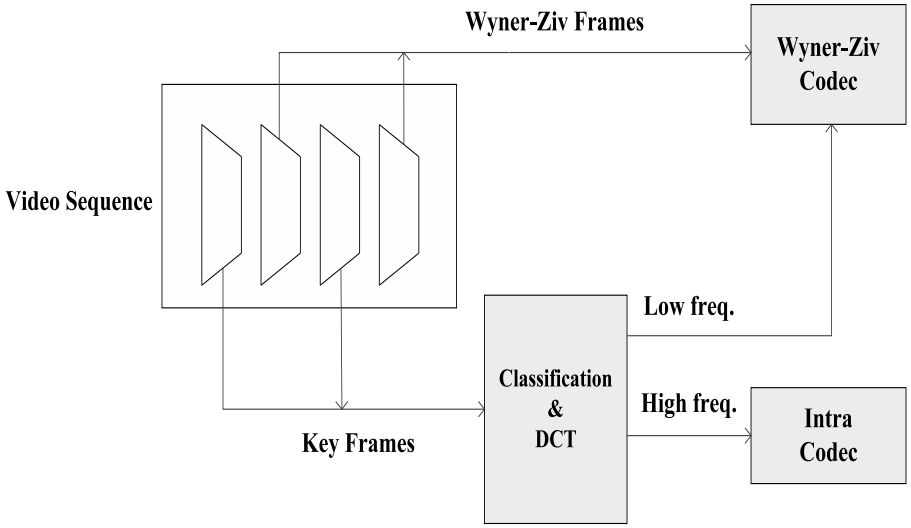

Figure 3: Wyner-Ziv video codec with encoding key frames based on frequency bands classification

Fig. 3 shows our proposed method. Frequency bands of each block are divided into two classes. Wyner-Ziv coding is used for the low frequency bands of each block, while high frequency bands are intra coded. Exploiting the temporal correlation of low frequency bands which tends to be high can result in outperforming the conventional intra method. Since low frequency bands and high frequency bands are known at both encoder and decoder, no bits need to be sent for classification. In our simulation based on a 4×4 block size, frequency bands $f(1,1)$, $f(1,2)$ and $f(2,1)$ are considered low frequency bands.

For relatively low motion sequences where many other frequency bands are highly correlated with their corresponding side information as well, lower rate-distortion performance is expected compared with the method proposed in Section 3.1.

4 Simulation results

Fig. 4 and Fig. 5 show the results for the first 99 frames of the Mother and Foreman sequences. For both plots, only the rate distortion performance of the luminance of

key frames is included. Odd frames are considered as key frames. So, this can be thought of as a GOP structure of size 2.

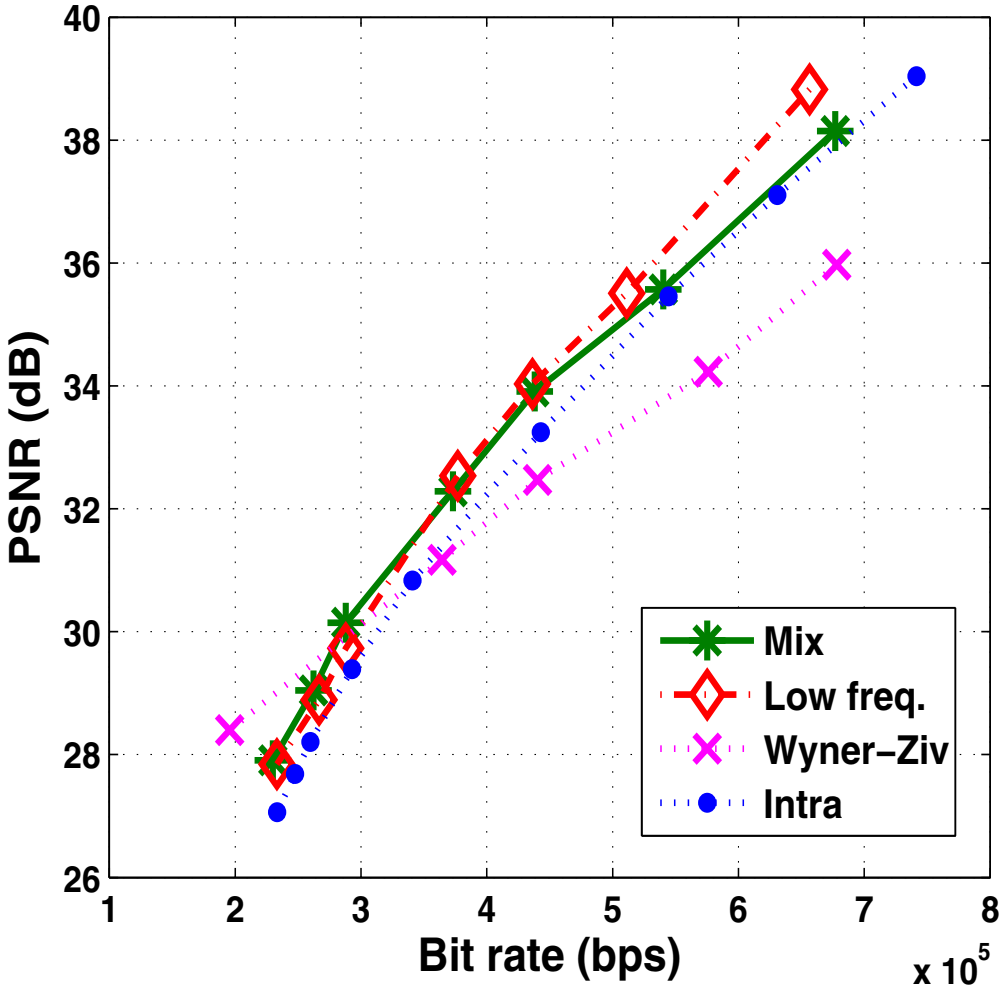

Figure 4: PSNR vs. rate for different methods of coding key frames for the Foreman sequence

For the "Intra" method, all blocks are intra encoded and decoded using a 4×4 DCT transform, Huffman and run length coding. Huffman and run length coding tables are borrowed from the JPEG standard. The DC coefficient is predictively encoded with respect to the DC coefficient in the adjacent block. For the Wyner-Ziv method, all blocks use Wyner-Ziv coding.

The method proposed in Section 3.1 which is based on block classification into three different classes of skip, intra and Wyner-Ziv coding is called the "Mix" method. In the "Mix" method, the DC coefficient of the blocks classified as intra is predictively encoded with respect to the DC coefficient in the previous intra block. To set T_2 for the "Mix" method, we tried different values of $\{0.8, 50, 100, 300, 500, 700\}$ for several sequences. Although $T_2 = 100$ is not the optimum value for any of these sequences,

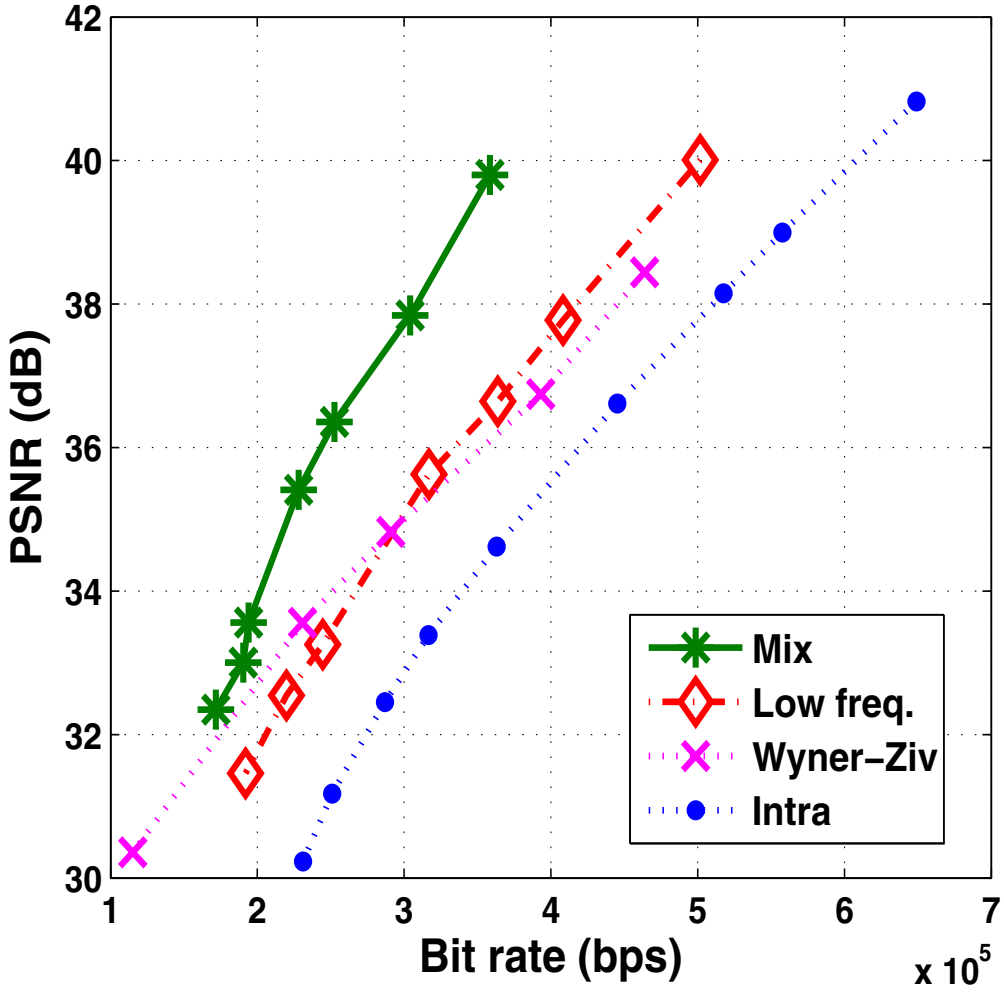

Figure 5: PSNR vs. rate for different methods of coding key frames for the Mother-daughter sequence

we used this value for all sequences because it works well for different sequences with different characteristics.

The method proposed in Section 3.2 which is based on using Wyner-Ziv coding for low frequency bands and intra coding for high frequency bands is called "Low freq". Appropriate quantization parameters are set to have similar quality reconstructed blocks for different classes of coding. For simplicity, 1584×2 bits are sent for the classification part of the "Mix" method, since we have 3 classes and 1584 blocks. This number of bits can certainly be reduced by using more complicated methods.

The LDPCA assumes a fixed length of input bits (in our case 396). For applying the "Mix" method, the bit value corresponding to the blocks which are classified to the skip and intra classes are set to zero at the encoder and their log-likelihood ratios at the decoder are set to a relatively high value (for example 30) indicating that those bits are zero with a very high confidence. The Wyner-Ziv decoder just cares about

the bits corresponding to the inter class and simply ignores the rest, since the other blocks use their own methods of coding.

As is shown, using the "Mix" method we can gain up to 5 dB improvement over the conventional intra method for *Mother-daughter* and 1 dB improvement for *Foreman* at low bit rates. The gain from applying the "Low freq" method is up to 2.2 dB for *Mother-daughter* and 1.3 dB for *Foreman* at both low and high bit rates.

As expected, applying the "Mix" method for *Foreman* which is a relatively high motion sequence is not very successful. For relatively high motion sequences, few blocks are assigned to Wyner-Ziv coding so almost the same distortion as from the "Intra" method is expected. Since the number of bits spent for transmitting classification information is almost as large as the number of bits saved because of doing the classification at high bit rates, the overall rate-distortion performance is only trivially better than "Intra".

We need to mention that even for *Foreman*, the "Mix" method gains up to 1 dB improvement over the intra method at low bit rates since using a coarse quantizer leads to more correlated source and side information.

As we can see, the "Low freq" method is more successful than the "Mix" method for *Foreman* since the temporal correlation of the low frequency bands of even a high motion frame is usually high. And also, no classification bits need to be sent. But for *Mother-daughter* the improvement is not as high as for the "Mix" method. It is reasonable because for relatively low motion sequences with many blocks classified as inter, the temporal correlation is high for many other frequency bands as well.

5 Conclusion

In this paper, we presented two methods of coding to jointly exploit the spatial and temporal correlation of key frames. Both methods are designed such that the encoder remains low complexity. For this purpose, an extension of the Wyner-Ziv coding method for key frames is considered.

The first method which attempts to classify blocks within a frame based on their motion activity to assign different methods of coding results in up to 5 dB improvement for low motion sequences. Simulation results show that the PSNR gain strongly depends on the motion activity of the sequences. The second method which is designed based on classification of frequency bands results in up to 2 dB improvement for both low motion and high motion sequences.

Both methods do not add anything to the latency of the system, because key frames only exploit frames from the past. Also, both methods have almost no impact on the complexity of the encoder because they make use of elements (transforms, LDPCA) which are already part of the conventional transform domain Wyner-Ziv video codec.

Since our frequency band classification method is very simplistic (always classify bands $f(1,1)$, $f(1,2)$ and $f(2,1)$ as low frequency) future extension of this work will focus on designing a low-complexity method to classify each frequency band based on statistics of each sequence.

References

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. 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 Trans. on Information Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1973.

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

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

[5] R. Puri, A. Majumdar, and K. Ramchandran, "Prism: A video coding paradigm with motion estimation at the decoder," *IEEE Trans. on Image Processing*, vol. 16, pp. 2436–2447, Oct. 2007.

[6] A. Aaron, S. Rane, and B. Girod, "Transform-domain Wyner-Ziv codec for video," *VCIP*, vol. San Jose, January 2004.

[7] C. Brites, J. Ascenso, and F. Pereira, "Improving transform domain Wyner-Ziv video coding performance," *IEEE ICASSP*, vol. 2, May 2006.

[8] S. Argyropoulos, N. Thomosy, N. Boulgourisz, and M. Strintzis, "Adaptive frame interpolation for Wyner-Ziv video coding," *IEEE 9th workshop on Multimedia signal processing*, pp. 159–162, Oct. 2007.

[9] S. Ye, M. Ouaret, F. Dufaux, and T. Ebrahimi, "Improved side information generation with iterative decoding and frame interpolation for distributed video coding," *Proc. ICIP*, 2008.

[10] J. Zhang, H. Li, Q. Liu, and C. W. Chen, "A transform domain classification based Wyner-Ziv video codec," *IEEE International Conference on Multimedia and Expo*, pp. 144–147, July 2007.

[11] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive distributed source coding using low-density parity-check codes," *Proc. Asilomar Conference on Signals and Systems*, 2005.