CORRELATION NOISE CLASSIFICATION BASED ON MATCHING SUCCESS FOR TRANSFORM DOMAIN WYNER-ZIV VIDEO CODING

Ghazaleh R. Esmaili and Pamela C. Cosman

Department of Electrical and Computer Engineering
University of California, San Diego, La Jolla, CA, 92093-0407
gesmaili@ucsd.edu, pcosman@ucsd.edu

ABSTRACT

Distributed source coding strongly depends on the knowledge of statistical dependency between source and side information. In transform domain Wyner-Ziv video coding (TDWZ) this statistical dependency (also known as correlation noise) has been usually modeled by a unique Laplacian distribution for each frequency band. In this paper, we propose a method to define different classes of correlation noise for each frequency band based on the accuracy of the side information. With this approach the correlation between source and side information is estimated separately for each frequency band of each class. Therefore, the decoder can discriminate blocks in order to estimate the correlation noise of their frequency bands. Simulation results show that applying the proposed method improves rate-distortion performance.

Index Terms— Wyner-Ziv coding, Distributed source coding, Correlation noise.

1. INTRODUCTION

Some recent applications such as multimedia sensor networks and mobile camera phones require many simple and low cost encoders but might have a small number of higher complexity decoders. Wyner-Ziv video coding is founded on the principles of distributed source coding. The complexity is shifted from the encoder to the decoder by encoding individual frames independently (intraframe encoding) but decoding them conditionally (interframe decoding). Based on the results of the Slepian-Wolf [1] and Wyner-Ziv [2] theorems, almost the same performance as the interframe-encoding interframe-decoding systems is expected.

Recently several practical Wyner-Ziv video codecs were proposed. In [3] Puri and Ramchandran introduced a power-efficient robust high compression syndrome-based video coding scheme which was upgraded to a more practical solution in [4]. In [5] Aaron and Girod proposed Wyner-Ziv video coding based on Turbo codes in the pixel domain. It 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.

Despite recent progress, Wyner-Ziv rate-distortion performance has not met the performance of predictive coding yet. One of the challenges is how to estimate the correlation noise between source and side information. The decoder needs some model for the statistical dependency between source and side information to make use of side information. The model is necessary for the conditional probability calculations in the Slepian-Wolf decoder and the conditional expectation in the reconstruction block. The dependency between source and side information is modeled by $Y = X + Z$ where Y denotes the side information and X denotes the source. Z is called the correlation noise. The conditional pdf of $f_{Y|X}(y|x)$ can be found equal to $f_{Y|X}(y|x) = f_Z(y - x)$. In most approaches the pdf of Z is approximated by a Laplacian distribution.

In [6] and most existing methods, the parameters are approximated by plotting the residual histogram of several sequences. In [8] and [9] some methods at different granularity levels were suggested for online parameter estimation of pixel and transform domain Wyner-Ziv coding. In existing methods based on the proposed approach in [6], blocks within a frame are treated uniformly in order to estimate the correlation noise, while the success of motion compensated frame interpolation (MCFI) methods to estimate the source is different for every block. It is therefore natural to differentiate the statistical dependency of each block based on the side generation success.

In this paper, we propose a simple and effective method to estimate the correlation noise based on MCFI success at the decoder. We define a practical criterion to evaluate the block matching success at the decoder, since in Wyner-Ziv video coding, having low complexity encoders is necessary. Also, we make use of block matching information in an efficient way to elevate the accuracy of correlation noise estimation.

The rest of this paper is organized as follows. In Section 2 we briefly review TDWZ and the method to estimate the correlation channel. In Section 3 we describe our proposed method in detail. In Section 4 the performance of the proposed method is evaluated and conclusions are presented.

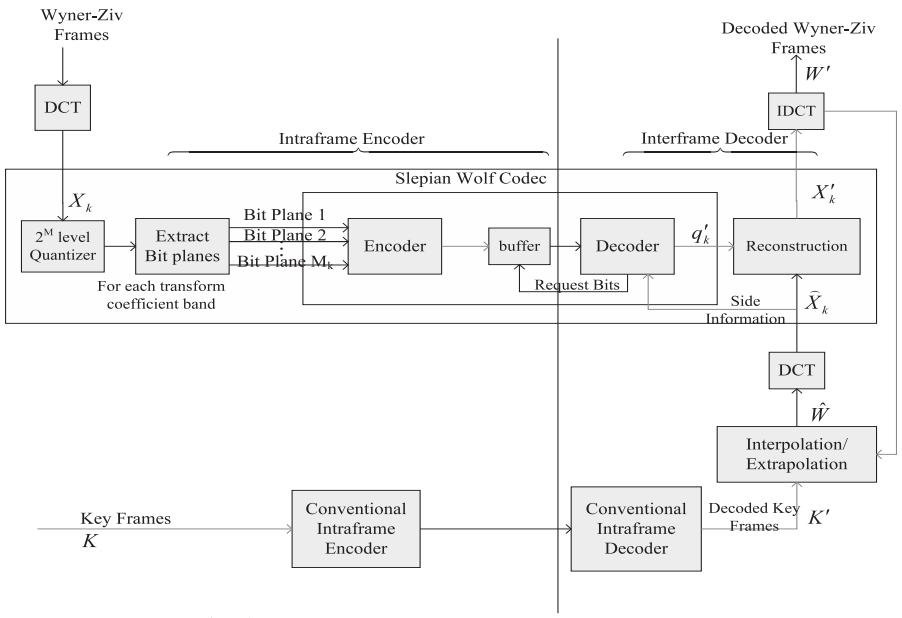

Fig. 1. Transform domain Wyner-Ziv video codec

2. TRANSFORM DOMAIN WYNER-ZIV CODING

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. 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 $\hat{X}$. As mentioned before, the decoder and reconstruction block assume a Laplacian distribution to model the statistical dependency between X_k and $\hat{X}_k$. The distribution of d can be approximated as $f(d) = \frac{\alpha}{2} e^{-\alpha|d|}$ where d denotes the difference between corresponding elements in X_k and $\hat{X}_k$. In most existing approaches, the α parameter of each DCT band is estimated by plotting the residual histogram of several sequences using MCFI for the side information. For example, for frequency band k the difference between corresponding elements in X_k and $\hat{X}_k$ of several sequences are grouped to form vector F_k. The α parameter is calculated by $\frac{\sqrt{2}}{\sigma_k}$ where σ_k is the square root of the variance of the F_k elements. In this way, we have a 16 element lookup table at the decoder shown by the last row of Table 1 where each element represents the α parameter of the corresponding DCT band.

3. THE PROPOSED METHOD

General MCFI methods are based on the assumption that the motion is smooth, continuous and translational. This assumption for high motion regions tends to give a poor estimation. The matching success of MCFI estimation depends on the motion characteristic of the block, and the statistical dependencies between source and side information vary based on that. It is therefore reasonable to use a different correlation estimation for each block within a frame to model its statistical dependency.

We need to define a criterion at the decoder to evaluate the matching success of MCFI. One of the best candidates for matching criterion is the residual energy between forward (FMCFI) and backward (BMCFI) interpolation of a given block. If v is the final motion vector obtained by MCFI for a given block B in F_t which is the skipped frame, forward and backward interpolations are $F_{t-1}(s+\frac{v}{2})$ and $F_{t+1}(s-\frac{v}{2})$ respectively where $s \in B$ and t is the time index. The residual energy between FMCFI and BMCFI for a given block is computed by $E = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} [FMCFI(x,y) - BMCFI(x,y)]^2$ where M and N represent the block size (in our case $M = N = 4$). Although E is a good candidate for evaluating MCFI success there is no linear dependency between that and the residual energy between source and side information. So, the idea of classification seems reasonable since for each class of matching success we can have a better estimation of dependency between source and side information.

In our method, the residual energy between Forward and Backward interpolation of every block within a frame for all Wyner-Ziv frames of several sequences is calculated to form a long vector R. We classify elements of this vector into m

Table 1. Lookup table of α parameters for 16 DCT bands of different classes

class	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$
1	1.59	2.46	2.58	2.71	2.38	2.55	2.59	2.81	2.57	2.62	2.69	2.89	2.82	2.93	3.00	3.23
2	1.36	2.02	2.03	2.18	1.88	2.03	2.05	2.19	2.02	2.14	2.18	2.34	2.27	2.41	2.42	2.60
3	1.19	1.67	1.70	1.86	1.52	1.63	1.73	1.88	1.70	1.79	1.87	2.03	1.92	2.05	2.09	2.28
4	0.56	1.00	1.21	1.40	0.88	1.19	1.30	1.45	0.95	1.29	1.44	1.64	1.20	1.46	1.60	1.90
5	0.39	0.71	0.85	1.05	0.49	0.82	0.93	1.14	0.62	0.85	1.05	1.25	0.82	1.09	1.22	1.50
6	0.27	0.45	0.55	0.72	0.37	0.54	0.63	0.80	0.46	0.60	0.70	0.90	0.62	0.75	0.87	1.05
7	0.19	0.31	0.39	0.51	0.26	0.36	0.45	0.58	0.32	0.42	0.49	0.65	0.44	0.57	0.60	0.75
8	0.09	0.16	0.20	0.26	0.15	0.19	0.23	0.31	0.18	0.22	0.25	0.34	0.25	0.30	0.33	0.39
Unique	0.21	0.34	0.41	0.53	0.31	0.41	0.50	0.65	0.39	0.48	0.55	0.72	0.51	0.62	0.69	0.83

different groups using a set of $m-1$ thresholds T_i where $i \in \{1, ..., m-1\}$ corresponding to m classes labeled 1 to m. Class i is chosen when $T_{i-1} < r < T_i$ where $r \in R$. Threshold values are set such that classes have almost the same number of elements. So, all coefficients corresponding to frequency band j of all blocks labeled with class i are grouped together to form vector $v_{i,j}$. The α parameter of vector $v_{i,j}$ is calculated by $\frac{\sqrt{2}}{\sigma_{i,j}}$ where $\sigma_{i,j}$ is the square root of the variance of $v_{i,j}$ elements.

Based on the above procedure, there are m different classes of correlation estimation for each frequency band. We have therefore an m by 16 (since 4×4 DCT is applied) lookup table of α parameters at the decoder. The component i, j of this table represents the α parameter of frequency band j of class i where $i \in \{1, ..., m\}$ and $j \in \{1, ..., 16\}$. For a given block of the skipped frame, the decoder evaluates the matching success of MCFI by calculating the residual energy between forward and backward interpolation and chooses one of the defined m classes by comparing to the threshold values. Once the block class is determined, the α parameter of each frequency band is found through the lookup table. In our simulation, the number of classes is set to 8 since in that case we can have enough elements in each class to have a reliable distribution model. Threshold values are $[T_1, T_2, ..., T_7] = [0.4, 0.65, 1.1, 2.7, 8, 25, 90]$. Table 1 shows the computed lookup table. Each row represents the α parameter of different DCT bands of a given class. The last row represents the calculated α parameter of different DCT bands based on the existing method where there is no classification. As we can see, the α parameter of each DCT band is a monotonic descending function of residual energy satisfying our expectation. Also, the α parameter of each class is an increasing function of frequency in each direction meaning that the α parameters of $f_{i,j}, f_{i,j+1}, ..., f_{i,j+3}$ and $f_{i,j}, f_{i+1,j}, ..., f_{i+3,j}$ are monotonically increasing. This confirms our classification reliability since the α parameters behave as they do when there is no classification.

As shown in Table 1, the α parameters of the last row corresponding to no classification method lie somewhere between class 6 and class 7. So for high motion sequences with most blocks classified to class 6 or higher we expect less improvement than for low motion sequences with most blocks classified to class 5 or lower.

4. RESULTS

Fig. 2 and Fig. 3 show the rate-distortion performance for the first 99 frames of the $Mother\text{-}daughter$ and $Foreman$ sequences at 30 fps. For both plots, only the rate distortion performance of the luminance of even frames is included. The improved transform domain Wyner-Ziv codec (IST-TDWZ) in [7] is implemented as the baseline. The only exception that we made is removing the availability of original key frames at the decoder. This assumption is not valid from a practical point of view. In our simulation, key frames are intra coded and decoded by H.264. Note that the rate and quality for key frames are the same for all the schemes to have a fair comparison. The result is compared with H.264/AVC intra, H.264/AVC I-B-I structure with no motion search and the best implemented method in [7]. By allowing 4×4 and 8×8 transforms, disabling the intra prediction modes and setting the "SearchRange" and "InterSearch16x16" through "InterSearch4x4" equal to zero, the complexity of the H.264/AVC encoder with the I-B-I structure is going to be comparable to the complexity of the Wyner-Ziv encoder. As we can see, applying the proposed classification method results in up to 1.1 dB improvement for $Mother\text{-}daughter$ and 0.5 dB for $Foreman$ over the best implemented method in [7]. The performance of $Mother\text{-}daughter$ is more improved since it is a relatively low motion sequence in comparison with $Foreman$. Based on our simulation, 54% of blocks of $Mother\text{-}daughter$ are classified to class 4 or lower while in $Foreman$ only 20% of them are classified to class 4 or lower.

In conclusion, we proposed a new method of correlation noise estimation for TDWZ based on block matching classification at the decoder. We were able to exploit additional statistical dependency between source and side information by using the residual energy between forward and backward interpolation as the matching criterion. The proposed approach does not increase the complexity at the encoder. Simulation

results show up to 1.1 dB improvement compared to the best implemented method in [7]. Classifying in 16 classes rather than 8 resulted in slightly higher gain (up to 1.3 dB improvement) at the expense of slightly more complexity at the decoder.

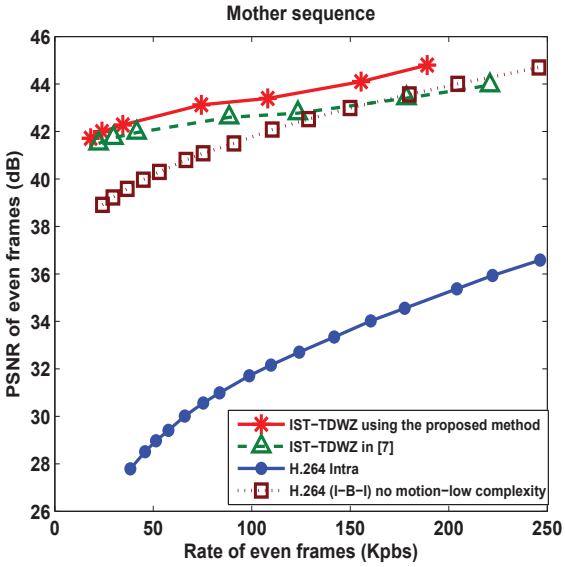

Fig. 2. PSNR vs. Rate of two different Wyner-Ziv codecs and H.264 for Mother-daughter.

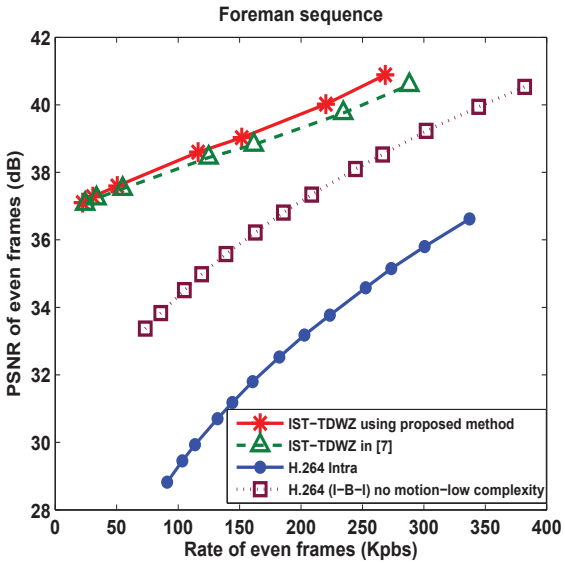

Fig. 3. PSNR vs. Rate of two different Wyner-Ziv codecs and H.264 for Foreman

5. 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] 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.

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

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

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

[8] C. Brites, J. Ascenso, and F. Pereira, "Studying temporal correlation noise modeling for pixel based Wyner-Ziv video coding," *IEEE International Conference on Image Processing*, Oct. 2006.

[9] C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding," *IEEE Trans. on Circ. and Sys. for Video Tech.*, vol. 18, pp. 1177 – 1190, Sept. 2008.