IMPROVED VIRTUAL CHANNEL NOISE MODEL FOR TRANSFORM DOMAIN WYNER-ZIV VIDEO CODING

Xin Huang and Søren Forchhammer

DTU Fotonik, Technical University of Denmark, Building 343, Lyngby 2800, Denmark
Email:{xhua, sofo}@fotonik.dtu.dk

ABSTRACT

Distributed Video Coding (DVC) has been proposed as a new video coding paradigm to deal with lossy source coding using side information to exploit the statistics at the decoder to reduce computational demands at the encoder. A virtual channel noise model is utilized at the decoder to estimate the noise distribution between the side information frame and the original frame. This is one of the most important aspects influencing the coding performance of DVC. Noise models with different granularity have been proposed. In this paper, an improved noise model for transform domain Wyner-Ziv video coding is proposed, which utilizes cross-band correlation to estimate the Laplacian parameters more accurately. Experimental results show that the proposed noise model can improve the Rate-Distortion (RD) performance.

***Index Terms*—** DVC, virtual channel, noise model, cross-band correlation

1. INTRODUCTION

Distributed Video Coding (DVC) [1] aims at avoiding complex motion estimation and compensation at the encoder and only explore the video statistics at the decoder side. According to the Slepian-Wolf theorem [2], it is possible to achieve the same rate as a joint encoding system by independent encoding but joint decoding of two statistically dependent signals. The Wyner-Ziv theorem [3] extends the Slepian-Wolf theorem to a lossy case, which becomes the key theoretical basis of DVC. One approach to DVC is to use a feedback channel based transform domain Wyner-Ziv video coding scheme. This was first proposed by the Stanford group in [4], then improved by the DISCOVER group (DIStributed COding for Video sER-vices) [5]. The DISCOVER codec improved coding performance by including a better side information generation scheme [6], an optimal reconstruction [7] and a realistic online noise model [8] at the decoder side. The coding efficiency of DVC is highly dependent on the error correcting capability of the channel code. A more accurate virtual channel noise model between the side information frame and the original frame will lead to improved channel coding performance.

A Laplacian distribution is usually utilized to model the difference of the transformed coefficients between the original frame and the side information in DVC. Accurate estimation of the Laplacian parameter is a complex task in DVC, because the side information frame is not reconstructed at the encoder side and the original frame is not available at the decoder side. Recently, different granularity online models [8][9] have been proposed to estimate the Laplacian distribution, i.e. from band (frame) level to coefficient (pixel) level for transform (pixel) domain Wyner-Ziv video coding. The results indicate that including finer granularity in the noise model improves the Rate-Distortion (RD) performance. In order to further improve the RD performance of transform domain Wyner-Ziv video coding, an improved noise model with a more accurate estimation of the Laplacian parameters is proposed. In the proposed model, a category map is generated based on previous successfully decoded bands, which are utilized to divide transformed coefficients of the current band into two categories. Different parameter estimators are applied for these two categories to locally calculate the Laplacian parameters. Finally, each transformed coefficient is assigned a Laplacian parameter based on its corresponding category and reliability.

The rest of this paper is organized as follows: Section 2 briefly describes the architecture of transform domain Wyner-Ziv video coding. In Section 3, noise models with different granularity are first described. Thereafter the proposed model is introduced. Test conditions and results are presented in Section 4.

2. ARCHITECTURE OF TRANSFORM DOMAIN WYNER-ZIV VIDEO CODING

The architecture of a transform domain Wyner-Ziv video codec [4][5] is depicted in Fig. 1. A fixed Group of Pictures (GOP=2) is adopted. The video sequence is first split into odd (key) frames and even (Wyner-Ziv) frames. The odd frames are intra coded by using a conventional video coding like H.264/AVC while the even frames are Wyner-Ziv coded.

In the encoder, Wyner-Ziv frames are partitioned into non-overlapped 4x4 blocks and an integer discrete cosine transform (DCT) is applied on each of these. The transform coefficients within a given band $b_k, k \in \{0...15\}$, are grouped together and then quantized [4]. DC coefficients and AC coefficients are uniformly scalar quantized and dead zone quantized, respectively. After quantization, the coefficients are binarized, each bitplane is transmitted to a rate-compatible LDPC accumulate encoder [10] starting from the most significant bitplane. For each encoded bitplane, the corresponding accumulated syndrome is stored in a buffer together with an 8-bit Cyclic Redundancy Check (CRC). CRC is used to aid the decoder detecting the convergence. The amount of bits to be transmitted depends on the requests from the decoder through a feedback channel.

In the decoder, an Overlapped Block Motion Compensation (OBMC) based interpolation algorithm [11] is adopted to create a side information frame Y_{2i} and a motion estimated residual frame R_{ME} based on two intra coded frames X_{2i-1} and X_{2i+1}. Y_{2i} and R_{ME} undergo the same 4x4 integer DCT to obtain coefficients $C_{Y_{2i}}$ and $C_{R_{ME}}$. $C_{R_{ME}}$ is utilized to model the noise distribution between corresponding DCT bands of the side information and Wyner-Ziv frames (i.e. $C_{Y_{2i}}$ and $C_{X_{2i}}$). By using the noise distribution obtained, coefficient values of the side information frame $C_{Y_{2i}}$ and the previous successfully decoded bitplanes, soft information (conditional bit probabilities P_{cond}) for each bitplane is estimated. With a given soft-input information P_{cond}, the LDPC decoder starts to process the corresponding bitplanes to correct the bit errors. Convergence is tested based on the 8-bit CRC and the Hamming distance between the received syndrome and the one obtained by the decoded bitplane: If the Hamming distance is different from zero after a certain amount of iterations, the LDPC decoder requests more accumulated syndrome bits from the encoder buffer via the feedback channel. If the Hamming distance is equal to zero, then the 8-bit CRC sum is requested from the buffer to verify successful decoding. A decoded bitplane with correct CRC sum is sent to a reconstruction module, a bitplane with incorrect CRC sum requests more accumulated syndrome bits from the encoder buffer to correct the existing bit errors until a low error probability is guaranteed.

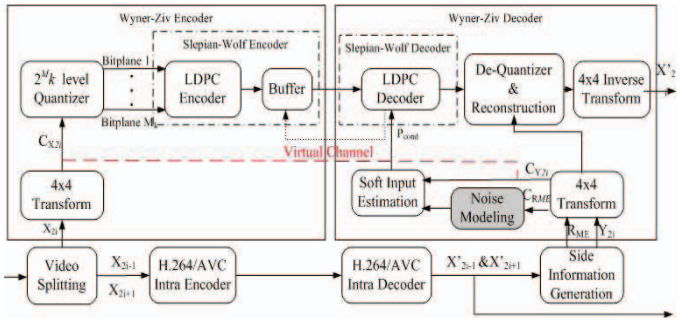


Fig. 1. Diagram of transform domain Wyner-Ziv video codec architecture

3. ONLINE NOISE MODELS

In order to take advantage of side information for decoding, the Wyner-Ziv decoder needs reliable information describing the noise distribution between the original frame and the side information frame R_{XY}. As a realistic solution in [8][9], a motion compensated residual R_{ME} between two key frames X_{2i-1} and X_{2i+1} is used (instead of an unrealistic offline residual R_{XY}) to estimate the Laplacian distribution parameter at the decoder side. Based on the work in [11], OBMC based side information generation is applied, therefore the motion compensated residual R_{ME} is obtained by:

$$R_{ME}(m_0, n_0) = \Sigma_{j=0}^{k} \omega_j \hat{R}_j / \Sigma_{j=0}^{k} \omega_j \tag{1}$$

$$\begin{aligned} \hat{R}_j = (X_{2i-1}(m_0 + \Delta m_j, n_0 + \Delta n_j) - \\ X_{2i+1}(m_0 - \Delta m_j, n_0 - \Delta n_j)) \end{aligned} \tag{2}$$

where (m_0, n_0) is the position within the current block, $(\Delta m_j, \Delta n_j)$ is the motion vector of the neighboring block j ($Block_j$) and k denotes the number of the neighboring blocks. ω_j is the weight of $Block_j$ obtained by:

$$\begin{aligned} \omega_j = (E_j[(X_{2i-1}(m_j + \Delta m_j, n_j + \Delta n_j) \\ -X_{2i+1}(m_j - \Delta m_j, n_j - \Delta n_j))^2])^{-1} \end{aligned} \tag{3}$$

where E_j is the expected value over $(m_j, n_j) \in Block_j$.

Different granularity online noise models for pixel domain and transform domain Wyner-Ziv video coding are discussed in [8][9]. In the following sub-sections, the band level and coefficient level noise models for transform domain Wyner-Ziv video coding are described first, then the proposed noise model is introduced.

3.1. Band Level

With the motion compensated residual R_{ME}, 16 bands of transformed residual coefficients $C^{b_k}_{R_{ME}}, b_k \in \{0...15\}$ are obtained after the 4x4 DCT transform. For a given band b_k, different Laplacian parameters $\alpha^{|\sigma|}_{b_k}$ are used to online model the distribution between transformed coefficients $C^{b_k}_{X_{2i}}$ and $C^{b_k}_{Y_{2i}}$:

$$f(C^{b_k}_{X_{2i}} - C^{b_k}_{Y_{2i}}) \approx \frac{\alpha^{|\sigma|}_{b_k}}{2} e^{-\alpha^{|\sigma|}_{b_k} |C^{b_k}_{R_{ME}}|} \tag{4}$$

$$\alpha^{|\sigma|}_{b_k} = \sqrt{2/\sigma^2_{|b_k|}}, \sigma^2_{|b_k|} = E(|C^{b_k}_{R_{ME}}|^2) - E(|C^{b_k}_{R_{ME}}|)^2 \tag{5}$$

where $\sigma^2_{|b_k|}$ is the variance of the absolute value of the transformed motion compensated residual ($|C^{b_k}_{R_{ME}}|$) within band b_k. The absolute value is chosen for Laplacian parameter estimation, since it is observed that the distribution with parameter $\alpha^{|\sigma|}_{b_k}$ is in general closer to the histogram of the actual residual $C^{b_k}_{R_{XY}}$ ($= C^{b_k}_{X_{2i}} - C^{b_k}_{Y_{2i}}$) compared with the distribution with the parameter $\alpha^{\sigma}_{b_k}$ obtained by residual ($C^{b_k}_{R_{ME}}$) through experiments [8] (See also Fig. 2).

3.2. Coefficient Level

In the band level noise model, the same Laplacian parameter $\alpha^{|\sigma|}_{b_k}$ is utilized for all the coefficients within band b_k. The spatial variation between different blocks is not explored, thus a coefficients level noise model ($c1$) is proposed in [8] to exploit spatial variation.

$$\alpha^{c1}_{b_k}(u,v) = \begin{cases} \alpha^{|\sigma|}_{b_k}, & \text{if } D(u,v)^2 \le \sigma^2_{|b_k|} \\ \sqrt{2/D(u,v)^2}, & \text{if } D(u,v)^2 > \sigma^2_{|b_k|} \end{cases} \tag{6}$$

$$D(u,v) = C^{b_k}_{R_{ME}}(u,v) - E(|C^{b_k}_{R_{ME}}|) \tag{7}$$

where $\alpha^{c1}_{b_k}(u,v)$ represents the estimated Laplacian parameter for the coefficient located at (u,v) within band b_k. $\alpha^{|\sigma|}_{b_k}$ and $\sigma^2_{|b_k|}$ are estimates of the Laplacian parameter and the variance at band level. $E(|C^{b_k}_{R_{ME}}|)$ represents the average absolute value of coefficients in band b_k. $C^{b_k}_{R_{ME}}(u,v)$ is the coefficients value at position (u,v) within band b_k. This coefficient level noise model divides coefficients into two categories by comparing D^2 and the variance $\sigma^2_{|b_k|}$. If D^2 is smaller than the variance, the band level Laplacian parameter $\alpha^{|\sigma|}_{b_k}$ is applied. Otherwise, the coefficient level parameter $\sqrt{2/D(u,v)^2}$ is assigned [8].

3.3. Proposed noise model

A pixel level noise model is proposed in [9] for pixel domain Wyner-Ziv video coding. This work is here extended to a coefficient level noise model ($c2$) for transform domain Wyner-Ziv video coding which weights band level and coefficient level statistics.

$$\alpha^{c2}_{b_k}(u,v) = \frac{\beta \cdot E(|C^{b_k}_{R_{ME}}|) \cdot \alpha^{|\sigma|}_{b_k}}{(\beta - 1) \cdot |C^{b_k}_{R_{ME}}(u,v)| + E(|C^{b_k}_{R_{ME}}|)} \tag{8}$$

where parameter β determines the amplitude of the deviations of $\alpha^{c2}_{b_k}(u,v)$ from $\alpha^{|\sigma|}_{b_k}$. $\beta = 2$ was chosen experimentally [9]. Generally, this noise model assigns Laplacian parameters adaptively based on the absolute magnitude of the transformed motion compensated residual. The larger the absolute transformed residual $|C^{b_k}_{R_{ME}}(u,v)|$ is, the less reliable it is, and therefore a smaller Laplacian parameter $\alpha_{b_k}(u,v)$ is assigned.

As in [8][9], the variance $\sigma^2_{|b_k|}$ is utilized to estimate the Laplacian parameter at band level (Eq. 5) which in turn influences the estimated coefficient level (Eqs. 6 and 8). The maximum likehood estimator can also be used to estimate the Laplacian parameter:

$$\alpha^{|b|}_{b_k} = ((\sum ||C^{b_k}_{R_{ME}}| - E(|C^{b_k}_{R_{ME}}|)|)/N)^{-1} \tag{9}$$

Assuming a Laplacian distribution, these two different estimators (Eqs. 5 and 9) should give the same parameter value. However, as shown in Fig. 2, the experiments indicate that $\alpha^{|b|}_{b_k}$ is generally larger than $\alpha^{|\sigma|}_{b_k}$. The histogram of the actual residual $C^{b_k}_{R_{XY}}$ is more peaked and has longer tails than the assumed Laplacian distribution. $\alpha_{b_k}^{|b|}$ is closer to the histogram close to zero while the $\alpha_{b_k}^{|\sigma|}$ is closer at the high values. Therefore it is reasonable to classify coefficients into two categories and apply the estimators $\alpha_{b_k}^{|b|}$ (Eq. 5) and $\alpha_{b_k}^{|\sigma|}$ (Eq. 9) for each category, respectively. Further, these estimators will be based on the coefficients within the respective category.

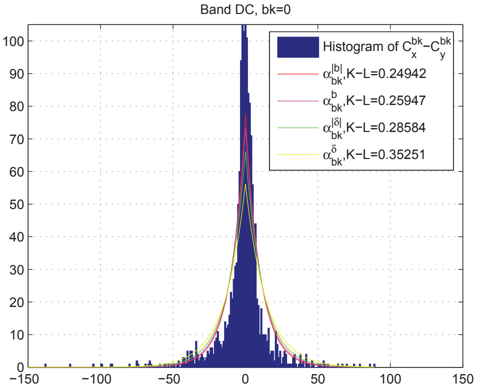


Fig. 2. Histogram of the actual residual $C_{R_{XY}}^0 = C_{X_{2i}}^0 - C_{Y_{2i}}^0$ and the estimated distributions with different estimators (DC coefficients, frame 22 of Foreman). Kullback-Leibler distances (KL) are calculated to compare the distance between the true distribution and modeling distribution.

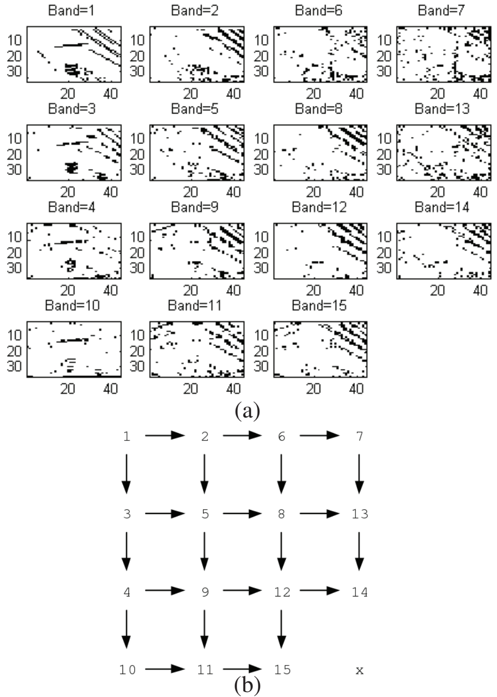
The coefficient level noise model proposed in [8] classifies coefficients by comparing $D(u,v)^2$ and the variance $\sigma_{|b_k|}^2$ as shown in Eq. 6. However, this calculation is only based on $C_{R_{ME}}^{b_k}$, which may be unreliable in some regions. Only using $C_{R_{ME}}^{b_k}$ (Eq. 6) may lead to inaccurate local parameter calculation. The correlation between classifications of different bands is tested in Fig. 3(a) based on comparing $D(u,v)^2$ and $\sigma_{|b_k|}^2$ of the actual residual $C_{R_{XY}}^{b_k}$.Therefore cross-band correlation can be utilized.

Since the Wyner-Ziv frames can be decoded successively band by band, after successfully decoding one (lower frequency) band b_k, an unfinished decoded frame (Z) can be reconstructed. By calculating the coefficients difference between $C_Z^{b_k}$ and $C_{Y_{2i}}^{b_k}$, an updated residual $C_{R_{ZY}}^{b_k}$ in band b_k is obtained, which is closer to the actual residual $C_{R_{XY}}^{b_k}$ than the motion compensated residual $C_{R_{ME}}^{b_k}$. The $\sigma_{|b_k|}^2$ and $D(u,v)^2$ in Eqs. 5 and 7 are recalculated based on the updated residual $C_{R_{ZY}}^{b_k}$, the classification map of band b_k is obtained as:

$$map_{b_k}^{out} = \{(u,v)|D(u,v)^2 > \sigma_{|b_k|}^2\} \qquad (10)$$

$$map_{b_k}^{in} = \{(u,v)|D(u,v)^2 \leq \sigma_{|b_k|}^2\} \qquad (11)$$

Due to the existing cross-band correlation, classification map of band b_k can be utilized to estimate the classification of the next (higher frequency) band $b_l, l > k$. The classification estimation follows the decoding order as shown in Fig. 3(b). For instance, after the first band is successfully decoded, the classification map of band 1 (map_1^{out},map_1^{in}) is obtained as described in Eqs. 10 and 11. The classification maps of band 2 and band 3 are simply estimated by copying the map of the neighboring band 1, i.e. $map_3^{out} = map_2^{out} = map_1^{out}$ and $map_3^{in} = map_2^{in} = map_1^{in}$. Similarly, the classification map of band 5 is estimated by using band 2 and band 3 by $map_5^{out} = map_2^{out} \cup map_3^{out}$ and $map_5^{in} = map_2^{in} \cup map_3^{in}$ etc. With the estimated classification, $\alpha_{b_k}^{|b|}$ and $\alpha_{b_k}^{|\sigma|}$ can be calculated within the coefficient sets $map_{b_k}^{in}$ and $map_{b_k}^{out}$, respectively.



(a)

(b)

Fig. 3. (a) Coefficient classification within different bands tested on the actual residual $C_{R_{XY}}^{b_k}$ (Frame 22 of Foreman). (b) The classification estimation from lower frequency band to higher frequency band

$$\alpha_{map_{b_k}^{in}}^{|b|} = ((\sum ||C_{R_{ME}}^{map_{b_k}^{in}}| - E(|C_{R_{ME}}^{map_{b_k}^{in}}|)|)/N)^{-1} \qquad (12)$$

$$\alpha_{map_{b_k}^{out}}^{|\sigma|} = \sqrt{2/(E(|C_{R_{ME}}^{map_{b_k}^{out}}|^2) - E(|C_{R_{ME}}^{map_{b_k}^{out}}|)^2)} \qquad (13)$$

In order to combine the advantages of the two coefficient level noise models described in the subsections 3.2 and 3.3, the Laplacian parameters for lower frequency bands and higher frequency bands are assigned differently. Let $\alpha_{b_k}^{c2}[(u,v)|C_{R_{ME}}^{map_{b_k}^{\bullet}}, \alpha_{b_k}^{|\sigma|}]$ denote the function in Eq. 8. For coefficients $C_{R_{ME}}^{b_k}, b_k \in \{0,1,2\}$,

$$\alpha_{b_k}(\mathrm{u,v}) = \begin{cases} \alpha_{b_k}^{c2}[(u,v)|C_{R_{ME}}^{map_{b_k}^{in}}, \alpha_{map_{b_k}^{in}}^{|b|}] & (u,v) \in map_{b_k}^{in} \\ \alpha_{b_k}^{c2}[(u,v)|C_{R_{ME}}^{map_{b_k}^{out}}, \alpha_{map_{b_k}^{out}}^{|\sigma|}] & (u,v) \in map_{b_k}^{out} \end{cases} \qquad (14)$$

For coefficients $C_{R_{ME}}^{b_k}, b_k \in \{3...15\}$,

$$\alpha_{b_k}(\mathrm{u,v}) = \begin{cases} \alpha_{map_{b_k}^{out}}^{|\sigma|} & \text{if } \sqrt{2/D(u,v)^2} \geq \alpha_{map_{b_k}^{out}}^{|\sigma|} \\ & \cup (u,v) \in map_{b_k}^{out} \\ \alpha_{map_{b_k}^{in}}^{|b|} & \text{if } \sqrt{2/D(u,v)^2} \geq \alpha_{map_{b_k}^{in}}^{|b|} \\ & \cup (u,v) \in map_{b_k}^{in} \\ \sqrt{2/D(u,v)^2}, & \text{otherwise} \end{cases} \qquad (15)$$

4. EXPERIMENTAL RESULTS

The following test conditions are used to obtain the RD performance results: The test sequences (available on [5]) are 149 frames of "Foreman", "Soccer", "Coast-guard" and "Hallmonitor" at 15 frames per second (fps). The most common GOP length of 2 is used. The key frames are encoded by H.264/AVC intra and the QPs

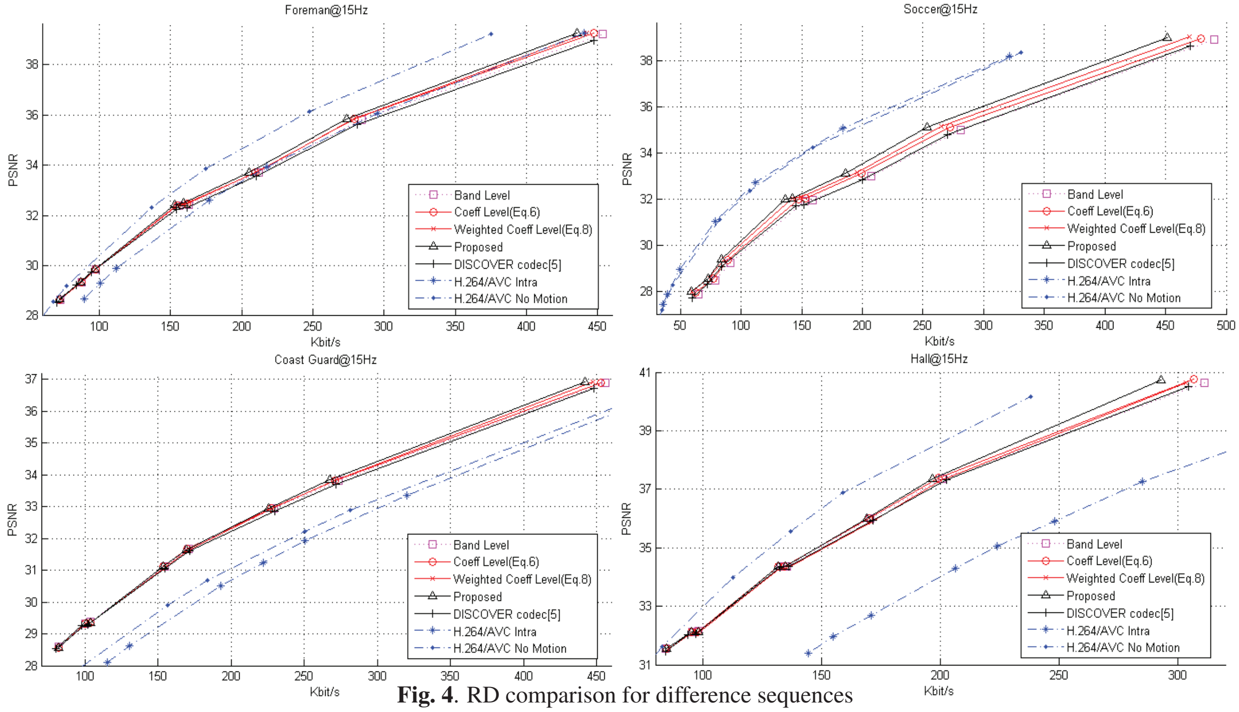


Fig. 4. RD comparison for difference sequences

are chosen so that the average PSNR of Wyner-Ziv frames are similar to the quality of key frames as in [5]. Overlapped Block Motion Compensation (OBMC) based side information generation [11] with half-pixel accuracy is utilized. The RD results are evaluated by the average for the luminance components of key frames and Wyner-Ziv frames. RD performance results of transform domain Wyner-Ziv video coding with different noise models are compared.

The experimental results are depicted in Fig 4. The performance of the DISCOVER executable codec [5]-[8] is depicted for comparison. The performance of H.264/AVC intra coding and H.264/AVC frame difference coding (i.e. No motion estimation with IBI GOP structure) are also included. The band level noise model with side information generation [11] is seen as a baseline. The coefficient level noise models achieve better RD performance than band level noise model. Compared with the coefficient level model [8] (Eq. 6) employed in the DISCOVER codec, the weighted coefficient level model (Eq. 8) gives better RD performance results for sequences ”Foreman”, ”Soccer” and ”Coast-guard”, but worse RD performance for sequence ”Hallmonitor”. The proposed noise model achieves better RD performance than all the other noise models. Compared with the coefficient level noise models, the proposed noise model is more robust and it improves the RD performance for high bit-rates up to 0.5 dB.

5. CONCLUSION

In this paper, an improved virtual channel noise model is proposed for transformed domain Wyner-Ziv video coding. It classifies the transformed coefficients into two categories by using the cross-band correlations, applies different estimators to locally calculate the Laplacian parameters and thus adaptively assigns a parameter value for each coefficient. Experimental results show that the proposed noise model can improve the coding efficiency of transformed domain Wyner-Ziv video coding up to 0.5 dB compared with the other noise models.

6. REFERENCES

[1] A. Aaron, R. Zhang, and B. Girod, “Wyner-ziv coding of motion video,” *Proc. Asilomar Conf. on Signals and Syst.*, pp. 240–244, 2002.

[2] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, July 1973.

[3] A.D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[4] A. Aaron, S. Rane, E. Setton, and B. Girod, “Transform domain wyner-ziv codec for video,” *Proc. SPIE VCIP*, pp. 520–528, Jan. 2004.

[5] Available on: www.discoverdvc.org.

[6] J. Ascenso, C. Brites, and F. Pereira, “Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding,” *5th EURASIP Conf. on Speech and Image Process., Multimedia Commun. and Services*, July 2005.

[7] D. Kubasov, J. Nayak, and C. Guillemot, “Optimal reconstruction in wyner-ziv video coding with multiple side information,” *IEEE Int’l Workshop Multimedia Signal Process.*, pp. 183–186, Oct. 2007.

[8] C. Brites and F. Pereira, “Correlation noise modelling for efficient pixel and transform domain wyner-ziv video coding,” *IEEE Trans. on Circuits Syst. Video Technol.*, 2008.

[9] L. Qing, X. He, and R. Lv, “Distributed video coding with dynamic virtual channel mode estimation,” *Int’l Symposium on Data, Privacy and E-Commerce*, pp. 170–173, 2007.

[10] D. Varodayan, A. Aaron, and B. Girod, “Rate-adaptive distributed source coding using low-density parity-check codes,” *EURASIP Signal Process. Journal, Special Section on Distributed Source Coding*, vol. 86, pp. 3123–3130, Nov. 2006.

[11] X. Huang and S. Forchhammer, “Improved side information generation for distributed video coding,” *IEEE Int’l Workshop Multimedia Signal Process.*, pp. 223–228, Oct. 2008.