Improving performance of distributed video coding by consecutively refining of side information and correlation noise model

Tien Vu Huu, Thao Nguyen Thi Huong, San Vu Van
Posts and Telecommunications Institute of Technology
Email: tienvh@ptit.edu.vn, thaonth@ptit.edu.vn, sanvv@ptit.edu.vn

Xiem HoangVan
VNU-University of Engineering and Technology
Email: xiemhoang@vnu.edu.vn

***Abstract*—Distributed video coding (DVC) is built on distributed source coding (DSC) principles where the video statistics are exploited, partly or fully, at the decoder instead of the encoder. In theory, DVC scheme is proved that there is no performance loss when compared to predictive video coding. However, its practical implementation has a large gap to achieve the theoretically optimum performance. The DVC coding efficiency depends mainly on creating the side information (SI) - a noisy version of original Wyner-Ziv frame (WZF) at the decoder, and modeling the correlation noise - the difference between the original WZF and corresponding SI. Performance of the DVC scheme will be improved if the SI and correlation noise are estimated as accurately as possible. So, this paper proposes a method to enhance the quality of SI and also correlation noise model by using information in decoded WZFs during the decoding process. Initial SI which is generated by Motion-Compensated Temporal Interpolation (MCTI) and previously decoded KFs will be used as reference frames to consecutively refine the side information after each bitplane is decoded. The experimental results show that performance of the distributed video coder is significantly improved by using this method.**

I. INTRODUCTION

In predictive video coder, the encoder is responsible for exploiting the source statistics and therefore, this architecture has the encoder more complex than the decoder. Inversely, distributed video coder shift the complexity from the encoder to the decoder because the statistical redundancies are exploited at the decoder. Such a setup is suitable for a range of emerging applications such as wireless video surveillance and multimedia sensor networks in which there is a high number of encoders and only one or few decoders.

Based on results of the two theorems Slepian-Wolf and Wyner-Ziv, practical implementations of DVC have been proposed [1,2]. In Stanford WZ video coding architecture [2], video sequence is split into keyframe (KF) and WZF. KFs are encoded by conventional encoder and transmitted to the decoder. WZFs are quantized and fed to channel encoder in order to create parity bits. Then, systematic bits are eliminated while partiy bits are transmitted to the encoder upon the request from the decoder during the decoding process. At the decoder, SI, an estimation of original WZF, is generated from some decoded KFs. Channel encoder uses the estimated correlation noise, the difference between the original WZF and corresponding SI, and parity bits sent in order to correct 'errors' in SI and hence recover the original WZF. Clearly, the quality of SI and also the correlation noise model CNM have strong impact on the coding efficiency of the DVC scheme. If the SI and CNM are estimated more accurately, the amount of parity information requested by the decoder would be reduced and the quality of the decoded WZF would be improved when reconstructed.

Among works on DVC, works about SI generation account for the most. Basically, SI generation is based on motion trajectories estimated between multiple reference frames. The obtained SI, created by motion compensation, is kept unchanged along the whole decoding process. In early transform domain WZ video codec, SI is guessed by interpolation[3,4] or extrapolation [5,6] techniques on decoded KFs. Interpolation methods use past and future previously decoded KFs while extrapolation methods use only the past previously decoded KFs. The results showed that interpolation techniques provide better quality of SI with additional delay when compared to extrapolation techniques. This approach fails when the reference frames are temporally far from each other or motion characteristics of video sequence are irregular and high. In 2005, Ascenso et al. developed a motion compensation temporal interpolation (MCTI) framework to create the SI [3] which has been largely used in the literature and also adopted in this paper. MCTI technique is detailed later in Section 2.

To enhance the quality of estimated SI, authors proposed SI generation methods which are not only based on the previously decoded KF frames but also decoded WZF and some auxiliary data sent from the encoder [7]. Advantage of this approach is the improved quality of SI but encoder complexity could be increased. The encoder is no longer Intra frame due to the (limited) Inter frame operations. Another approach is learning during the decoding process [8,9,10]. This approach is characterized by consecutively refining initial SI by exploiting decoded data and applying some motion estimation techniques after each decoded bitplane or band. Authors proposed to improve the quality of the decoded frame each time a bitplane/band is successfully decoded and use it as enhanced SI to help in the decoding of future bitplanes/bands. Proposals in [8,10] uses the adjacent KFs in order to refine the motion vectors while proposal [9] searches for SI candidates within a given window on the initial SI

to reduce computational complexity. This approach usually provides better final SI quality and reduces the requested bits when compared to above mentioned SI generation methods, however the decoder complexity might be increased because of the motion estimation strategy used in the SI refinement operation.

To make usage of obtained SI, the decoder needs to have a reliable knowledge of the model that characterizes the correlation noise between the original WZF and the corresponding SI frame. So, besides SI generation, correlation noise estimation also receives a great attention from researchers. Clearly, this is difficult task because the original information is only available at the encoder and the SI is computed at the decoder and the SI quality varies along the sequence and within each frame or in other words, it is non-stationary in both temporal and spatial directions. In DVC, correlation noise can be estimated at the encoder using the original data and a copy of the SI; at the decoder assuming that the original data is available together with the SI or at the decoder without using the original data. However, the third case corresponds to a realistic and practical scenario where CNM parameter is estimated at the decoder in real time without making use of the original data.

In most of DVC works, the Laplacian distribution [11] is widely used to model the correlation noise because of the good tradeoff between model accuracy and complexity. Many observations show that the Laplacian distribution is not always satisfied and so, some works look for other distributions. In [12], an exponential power model, sometimes named "Generalized Gaussian", is used. Authors in [13] use mixture correlation noise model to describe the error distribution depending on the motion characteristic of the frame. Gaussian distribution is selected for DC coefficients of the low motion frames and Laplacian distribution is chosen for DC coefficients of the high motion frames. With the same approach of SI generation, the accuracy of the online CNM could be improved by progressively refining the correlation noise after decoding a bitplane/band [14,15].

In this paper, a new approach to enhance the SI is proposed. This solution successively refines the SI after each decoded bitplane and then CNM parameter is also adapted depending on the current SI. Initial SI is generated by MCTI method using the backward and forward reference frames. After each bitplane is decoded, DCT coefficients are reconstructed and a temporary WZ frame is created and used as a new SI for the decoding of next bitplane. This new SI is named partial decoded WZ (PDWZ). The PDWZ frame is further refined if it is considerd to be high and/or complex motion. Each block in the PDWZ frame is motion estimated in previously decoded KFs and the initial SI with a larger search area in order to create a new SI for the next bitplane. For CNM, after each band is decoded, CNM parameter is reestimated and updated for decoding the next band.

This paper is organized as follows: Section II presents the transform domain WZ video codec architecture. Section III describes the proposed solution while Section IV discusses the performance in comparison with relevant works. Finally, section V gives some conclusions and future works.

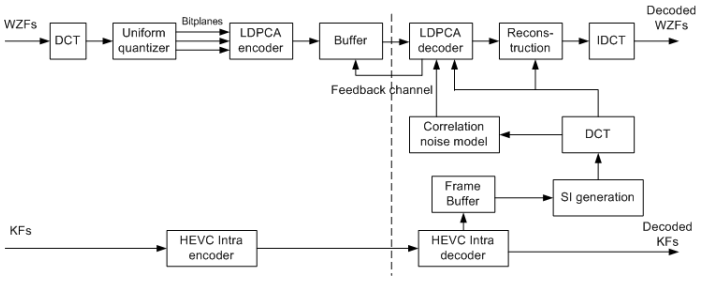

Fig. 1. Architecture of DVC - HEVC video codec

II. TRANSFORM DOMAIN WZ VIDEO CODEC

In this section, we briefly present the transform domain WZ video codec based on the DISCOVER architecture [16]. Different from the original DISCOVER, the KFs in this codec are encoded using HEVC Intra coding. Therefore, it is named DVC-HEVC and is illustrated in Fig. 1. The WZ frame encoding and decoding procedures are detailed in the following.

A. Encoder

First, each WZF is partitioned into non-overlapping block of 4×4 and a discrete cousine transform (DCT) is applied to each block to form the DCT coefficients in zigzag scanning order. DCT coefficients are mapped into 16 bands where each band contains the coefficients in the same positions of different blocks. Then, these DCT bands are uniformly quantized with a number of levels. Quantization matrices corresponding to eight different rates $Q_i = 1, 2, ..., 8$ are chosen as in [9]. Obtained bits are fed into channel encoder LDPCA to generate parity bits. These bits are stored in a buffer and depending on the request from the decoder, parity bits are transmitted to the decoder.

B. Decoder

Using the decoded KFs, the decoder creates the side information with a motion interpolation technique named MCTI. In this technique, SI is interpolated from backward and forward reference frames by using steps: forward motion estimation, bidirectional motion estimation, spatial smoothing and bidirectional motion compensation. A 4×4 block-based DCT is then carried out over the SI in order to obtain an estimate of WZ frame DCT coefficients.

In order to model the error distribution between corresponding DCT bands of SI and WZ frame, the DISCOVER codec uses a Laplacian distribution as follows:

$$f_{X|y}(x) = \frac{\alpha}{2} e^{-\alpha|x-y|} \tag{1}$$

where $f_{X|y}$ is the conditional p.d.f of X given y. α is the Laplacian distribution parameter defined by:

$$\alpha = \sqrt{\frac{2}{\sigma^2}} \tag{2}$$

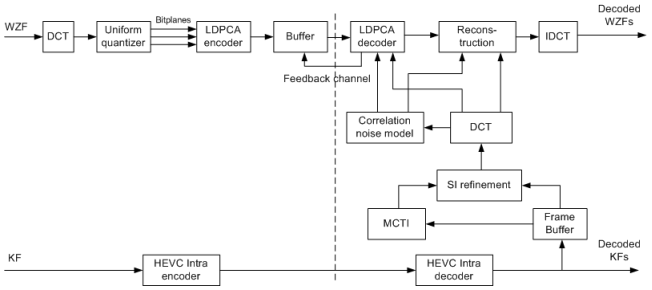

Fig. 2. Proposed transform domain WZ video codec

In Eq.(2), σ^2 is the variance of the residual between the original WZF and corresponding SI frame. The Laplacian parameter is estimated offline at the encoder or online at the decoder at band or coefficient level.

When the DCT transformed SI and the residual statistics for a given DCT band are known, the channel decoder corrects the bit errors in the DCT transformed SI using the parity bits of WZF requested through the feedback channel. Then, reconstruction module restores the original DCT coefficients from decoded quantized DCT coefficients and the side information. Finally, reconstructed DCT coefficients are inversely DCT transformed (IDCT) to obtain the pixel domain frame.

III. PROPOSED METHOD

Fig. 2 shows the proposed transform domain video codec that is based on the codec structure mentioned in Section II. Initial SI, SI_{MCTI}, is created by MCTI on backward and forward decoded KFs. After each bitplane of a band is fed to the channel decoder to get the corrected quantization bin, the reconstruction module will restore the original DCT coefficient with the help of corresponding SI. The reconstructed DCT coefficients will replace the collocated coefficients in the last SI to create a new SI called partial decoded WZ (PDWZ). In case of low motion content, PDWZ is considered to have a good quality but in case of high motion and irregular, it could be very different from the original WZF and so, PDWZ need to be reestimated with a larger search area. The searching is performed in three reference frames: backward, forward KF and SI_{MCTI} in order to create three block candidates corresponding to three motion vectors. These candidates are fused to create the new block. This procedure is repeated whenever each bitplane is decoded and new SI would become more and more similar to the original WZF. After each band is successfully decoded, the reconstructed coefficients are used by CNM module to update the CNM parameter for decoding the next band. In particularly, the proposed algorithm is described as following:

1) Step 1: Identification of search range for the PDWZ: When each bitplane is decoded, the PDWZ is generated. This PDWZ is motion estimated in different search ranges depending on motion content of that frame. In this paper, we use the averaged motion vector magnitude parameter in Eq. (3) to describe the motion activity.

$$MV_{avg} = \frac{1}{N}\sum_{i=1}^{N} ||\overrightarrow{MV_i}|| \tag{3}$$

where $\overrightarrow{MV_i}$ is the motion vector in the frame and N is the number of motion vectors in the frame. Search range is determined for each PDWZ depending on which condition the MV_{avg} satisfies in Eq.(4).

$$Search\ range = \begin{cases} 14 \times 14 & \text{if } \text{MV}_{avg} < T_1 \\ 16 \times 16 & \text{if } T_1 \leq \text{MV}_{avg} < T_2 \\ 18 \times 18 & \text{if } \text{MV}_{avg} \geq T_2 \end{cases} \tag{4}$$

where T_1, T_2 are thresholds. In this work, T_1, T_2 are empirically selected to be 5 and 10 respectively.

Motion estimation is performed for each block in PDWZ in three reference frames (RF): backward and forward KFs and initial SI. Matching error is computed as the sum of the error absolute between the block n of PDWZ and block k in reference frame.

$$MAD = \alpha_n(RF) = \sum_{x=0}^{3}\sum_{y=0}^{3} |RF_k^{d(k)}(x,y) - PDWZ_n(x,y)| \tag{5}$$

where $d(k)$ is the displacement corresponding to the candidate block k and (x, y) are the pixel coordinates inside the displaced 4×4 block $PDWZ_n$.

The block k with the lowest MAD in the search range is considered the most similar to block n.

2) Step 2: Fusion of candidate blocks : Let X_f, X_b, SI_{MCTI} are forward and backward decoded KFs and initial SI. After the motion estimation procedure, we have three block candidates in three reference frames corresponding to their MAD values: $\alpha_n(X_f)$, $\alpha_n(X_b)$, $\alpha_n(SI_{MCTI})$. Naturally, lower matching error should imply that the block is a better candidate and it should contribute more to the final SI and vice-versa. So, weighting factors $\beta_n(RF) = 1/\alpha_n(RF)$ are used to determine the contribution of each block candidate in the fusion scheme. We use Eq. (6) to compute the new SI block:

$$SI_n = \frac{X_f.\beta_n(X_f) + X_b.\beta_n(X_b) + SI_{MCTI}.\beta_n(SI_{MCTI})}{\beta_n(X_f) + \beta_n(X_b) + \beta_n(SI_{MCTI})} \tag{6}$$

This new SI is used to decode the next bitplane at the decoder. When all bands are decoded, the final SI is generated to perform the reconstruction again to obtain the final WZ frame.

After each band, the refined SI is also used to adapt the parameter α in correlation noise model. Following an approach similar to the method proposed in [11], transform domain correlation noise modeling is performed at the decoder at the DCT band/frame granularity level. The difference is Laplacian distribution parameter α in Eq.(1) is adapted according to the newly refined SI at each bitplane. In this paper, correlation noise estimation is performed at DCT band level and parameter α is updated after decoding each band. The α value for each DCT band is computed as in Section VII.A of [11].

TABLE I
CHARACTERISTICS OF TEST SEQUENCES

Test sequences	Spatial resolution	Number of frames	Quantization parameters
Akiyo	176x144	300	{25,29,34,40}
Carphone		300	{25,29,34,40}
Foreman		300	{25,29,34,40}
Soccer		300	{25,31,36,44}

TABLE II
RD PERFORMANCE GAIN OF VIDEO SEQUENCES COMPARED TO DISCOVER CODEC USING BJONTEGAARD METRIC

	BD rate (%)	BD PSNR (dB)
Akiyo	-80.31	10.19
Carphone	-67.97	6.07
Forman	-6.92	0.39
Soccer	-2.76	0.15

IV. PERFORMANCE EVALUATION

A. Test conditions

To evaluate the performance of the proposed algorithm, four common video sequences are tested for the simulation: *Aikyo, Carphone, Foreman, and Soccer*. These sequences were selected for their representativeness of motion and texture characteristics. Table I summarizes the main characteristics of these sequences. Four RD points are considered corresponding to the four 4×4 quantization matrices. The values in each 4×4 matrix indicate the number of quantization levels associated to the various DCT coefficients bands. When the Q_i increases, the bitrate and the quality also increase. In order to improve the user subjective impact by reaching a smooth quality variation, KFs are HEVC intra encoded using a quantization parameter which allows reaching a similar quality as WZF for each RD point.

For each RD point, bitrate and PSNR are computed for the luminance component of each frame. These results of the proposed method are compared to DVC-HEVC codec in which SI is generated by MCTI technique. We also use the Bjontegaard metric to measure average difference between the two RD-curves of two methods. The bitrate saving and PSNR gain between two methods are computed for a given objective quality and for a given bitrate, correspondingly.

B. Experimental results

This section presents the RD performance of the proposed codec in comparison with the DVC-HEVC codec mentioned above. From all the RD plots illustrated in Fig 3,4,5,6, it is observed that the proposed solution outperforms DVC-HEVC codec for all Aikyo, Carphone, Foreman and Soccer sequences. From results obtained in Table II, the proposed method gives a gain up to 10.19 dB and rate reduction of 80.31% compared to DVC-HEVC codec for Aikyo sequence. It is clear that the proposed method achieves the best RD performance for low motion sequences such as Aikyo and Carphone. For complex motion sequences such as Foreman and Soccer, the results do not show such big improvements because these sequences contain complex motion which is difficult to estimate and therefore, the quality of refined SI is worse.

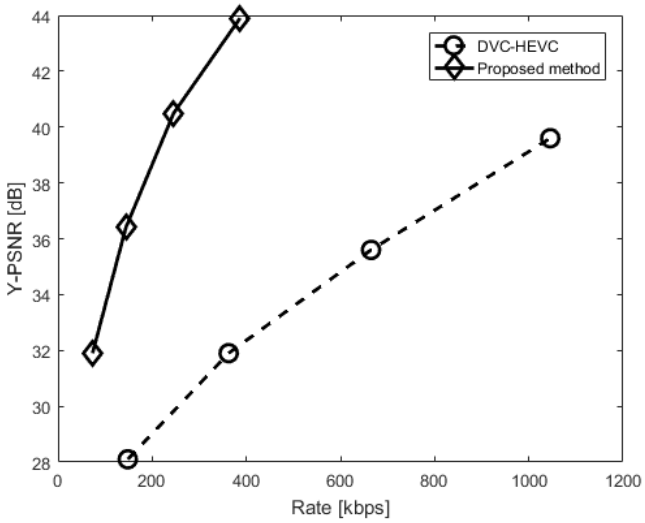

Fig. 3. RD performance for Akiyo sequence

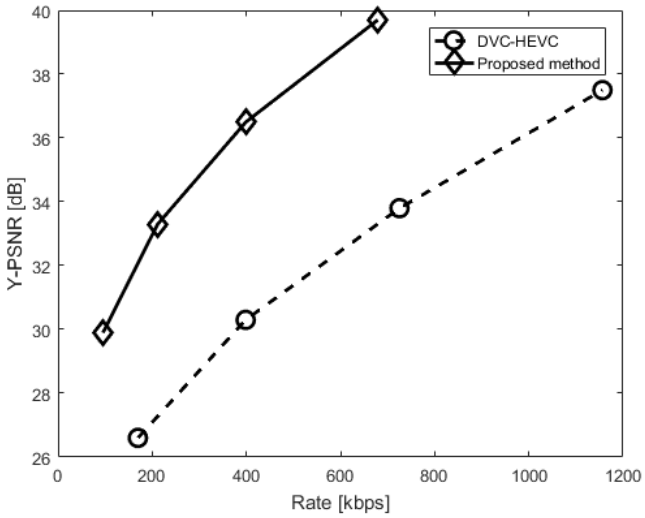

Fig. 4. RD performance for Carphone sequence

V. CONCLUSION

In this paper, a method to enhace the quality of SI and CNM is proposed for distributed video coding. In the proposed method, three candidate SI blocks are generated by motion estimation in three reference frames. Then the fusion is performed to generate the final SI block. The new SI is used for the decoding of next bitplane and for computing the α parameter of the correlation noise model. The experimental results show that the proposed method can significantly improve the SI quality comparing with previous algorithms. The

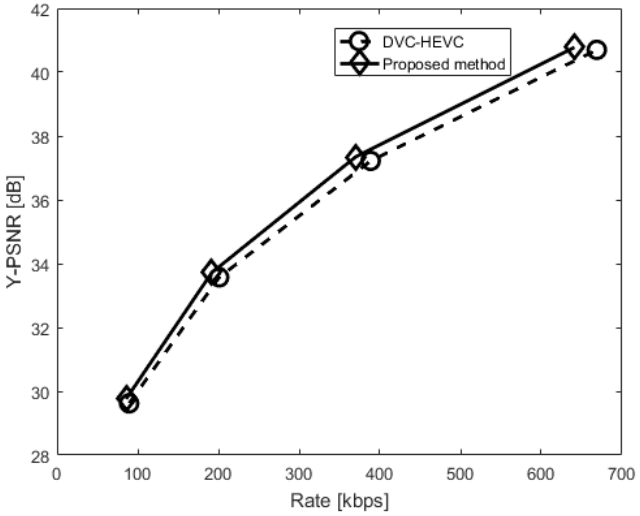

Fig. 5. RD performance for Foreman sequence

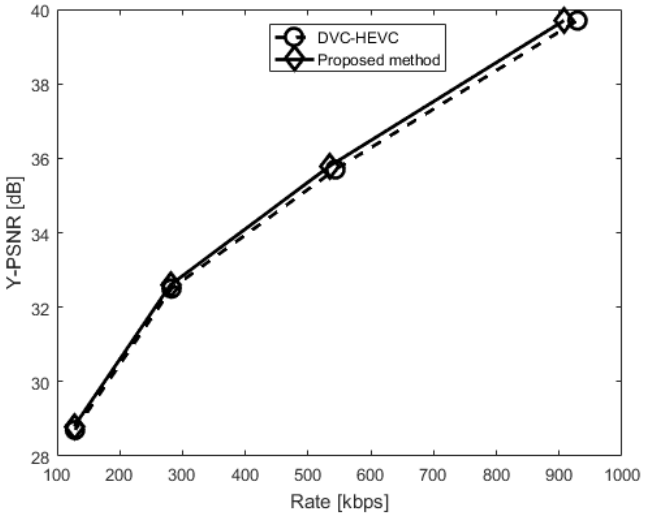

Fig. 6. RD performance for Soccer sequence

future works may improve motion esimation techniques and optimal selection scheme to create better SI frame.

REFERENCES

[1] R.Puri, A.Majumdar, and K.Ramchandran, PRISM: a video coding paradigm with motion estimation at the decoder, IEEE Transactions on Image Processing, vol.16, no.10, pp.2436-2448, Oct.2007.

[2] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, Distributed video coding, Proc. IEEE, vol. 93, no. 1, pp. 7183, Jan. 2005

[3] J. Ascenso, C. Brites, and F. Pereira, Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding, in Proc. 5th EURASIP Conf. Speech Image Process., Multimedia Commun. Services, Slovak, Jul. 2005, pp. 16

[4] D. Kubasov, C. Guillemot, Mesh-based motion-compensated interpolation for side information extraction in distributed video coding, in: Proceedings of IEEE International Conference on Image Processing (ICIP), Atlanta, GA, USA, October 2006X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret,

[5] A. Aaron, S. Rane, E. Setton, B. Girod, Transform-domain WynerZiv codec for video, in: Proceedings of SPIE Visual Communications and Image Processing (VCIP), San Jose, CA, USA, January 2004

[6] L. Natrio, C. Brites, J. Ascenso, F. Pereira, Extrapolating side information for low-delay pixel-domain distributed video coding, in: Proceedings of International Workshop on Very Low Bitrate Video Coding (VLBV), Sardinia, Italy, September 2005

[7]] J. Ascenso, F. Pereira, Adaptive hash-based side information exploitation for efficient WynerZiv video coding, in: Proceedings of IEEE International Conference on Image Processing (ICIP), San Antonio,TX, USA, September 2007

[8] J. Ascenso, C. Brites, F. Pereira, "Motion compensated refinement for low complexity pixel based distributed video coding", International Conference on Advanced Video and Signal-Based Surveillance, Como, Italy, September, 2005.

[9] R. Martins, C. Brites, J. Ascenso, F. Pereira, Refining side information for improved transform domain WynerZiv video coding, IEEE Transactions on Circuits and Systems for Video Technology 19 (9)(2009) 13271341

[10] A. Abou-Elailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, F. Dufaux, Improved side information generation for distributed video coding, in: Proceedings of the European Workshop on Visual Information Processing (EUVIP), Paris, France, July 2011.

[11] C. Brites and F. Pereira, Correlation noise modeling for efficient pixel and transform domain WynerZiv video coding, IEEE Trans. Circuit Syst. Video Technol., vol. 18, no. 9, pp. 11771190, Sep. 2008

[12] JThomas Maugey, Jerome Gauthier, Beatrice Pesquet-Popescu, et al.. Using an exponential power model for Wyner Ziv video coding. IEEE International Conference on Acoustics Speech and Signal Processing, Dallas, TX, March 2010, pp. 23382341.

[13] Hao Qin, Bin Song, Yue Zhao, and Haihua Liu, Adaptive Correlation Noise Model for DC Coefficients in Wyner-Ziv Video Coding. ETRI Journal, Volume 34, Number 2, April 2012, pp. 190-198.

[14] J. Park, B. Jeon, D. Wang, and A. Vincent, WynerZiv video coding with region adaptive quantization and progressive channel noise modeling, in Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB), Bilbao, Spain, May 2009, pp. 16.

[15] Brites, C., Ascenso, J., Pereira, F. (2012). Learning based decoding approach for improved Wyner-Ziv video coding. 2012 Picture Coding Symposium, pp. 165-168.

[16] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M.Ouaret, The DISCOVER codec: Architecture, techniques and evaluation, Proc. of Picture Coding Symposium, Oct. 2007.