

Band-Level Correlation Noise Modeling for Wyner–Ziv Video Coding with Gaussian Mixture Models

Ming Yin · Junbin Gao · Daming Shi ·
Shuting Cai

Received: 5 February 2014 / Revised: 24 November 2014 / Accepted: 25 November 2014 /
Published online: 5 December 2014
© Springer Science+Business Media New York 2014

Abstract As one of the most adopted distributed video coding approaches in the literature, Wyner–Ziv (WZ) video coding is not yet on par with the motion-compensated predictive coding solutions with respect to rate–distortion (RD) performance. One of the essential reasons lies in the absence of reliable knowledge of the correlation statistics between source and side information. Most of the existing works assume a probability distribution of the statistical dependency to be Laplacian, which is not accurate but computationally cheap. In this paper, a correlation estimation based on Gaussian mixture model is proposed for the band-level correlation noise of discrete cosine transform domain Wyner–Ziv codec. The statistics of the correlation noise between WZ frame and corresponding side information is analyzed by considering the temporal correlation and quantization distortion. Accordingly, the model parameters for correlation noise are estimated offline and utilized online in consequent decoding. The simulation results of Kullback–Leibler divergence show that the proposed model has higher accuracy than the Laplacian one. Experimental results demonstrate that the

M. Yin (✉) · S. Cai
School of Automation, Guangdong University of Technology, Guangzhou 510006, China
e-mail: yiming@gdut.edu.cn

S. Cai
e-mail: shutingcai@gdut.edu.cn

J. Gao
School of Computing and Mathematics, Charles Sturt University,
Bathurst, NSW 2795, Australia
e-mail: jbgao@csu.edu.au

D. Shi
School of Engineering and Information Sciences, Middlesex University,
London NW4 4BT, UK
e-mail: d.shi@mdx.ac.uk

WZ codec incorporated with the proposed model can achieve very competitive RD performance, especially for the sequence with high motion contents and large group of picture (GOP) size.

Keywords Correlation noise · Quantization · Wyner–Ziv coding · Rate–distortion · Gaussian mixture model · Wireless sensor network

1 Introduction

Due to the limited battery power, individual nodes in wireless multimedia sensor network (WMSN) [3] have low processing capability, which calls for lightweight signal processing and compression algorithms [28]. However, the traditional video coding solutions, represented by the ISO/IEC MPEG and ITU-T H.26x standards, rely on a highly complex encoder and cannot meet the processing capability requirements of nodes [4,5]. Meanwhile, the traditional predictive coding architecture is prone to suffering from distortion in wireless environment. Confronting the challenges of multimedia communications in wireless sensor network (WSN), there is an urgent demand for decreasing the computational energy consumption involved in compressing video streams in WMSN applications.

To meet the requirements of WMSN, distributed video coding (DVC) is the right paradigm of video coding which can provide such a low complexity encoder. Based on Slepian–Wolf and Wyner–Ziv information theorems [38], Wyner–Ziv video coding, as one of the popular techniques in DVC, can achieve efficient data compression by incorporating source statistics partially or fully in decoder. Recently, WZ video coding has attracted considerable attention [2,16,22] as its relative low complexity encoder. Specifically, in WZ codec, a flexible allocation of complexity [17,34] between the encoder and the decoder is exploited which is well suitable for WMSN applications. Two major approaches toward WZ video coding have been reported as pixel- and transform-domain schemes in literature, e.g., [1]. Although the pixel-domain codec is simpler regarding to the computational complexity, the transform-domain codec performs better in terms of coding efficiency despite a cost of slightly higher complexity.

According to its principle, coding efficiency of DVC is generally achieved by utilizing the correlation statistics between source and side information (SI) [8,37,41]. Either Turbo [10] or low-density parity-check (LDPC) [39] in WZ codec needs exploit the correlation noise statistics to initialize the decoding algorithm by providing likelihood estimates for the source bits. In particular, high coding efficiency critically relies on the capability of modeling the statistics of the correlation noise. Usually, the dependency is characterized by the residue $N = X - Y$ modeled by a Laplacian distribution with zero mean [8,11,18,25,37], where X is an input frame to be recovered and Y is SI at the decoder. However, it will be much challenging to accurately model the correlation noise statistics since the source is unavailable at the decoder and the absence of SI usually occurs at the encoder. Moreover, the non-stationary characteristics of video signals and occlusions or illumination changes are the major factors impacting on the correlation noise statistics. In fact, the current assumption that the distribution of residuals complies with a Laplacian is not always satisfied and, more often, the Lapla-

cian model significantly differs from the true correlation noise distribution [23]. To improve the coding performance, thus, correlation noise statistics between the source and SI needs to be estimated as accurately as possible. To this end, this paper aims to apply a more general model to fit the correlation noise.

It has been proved that Gaussian mixture model (GMM) is general enough to approximate different kinds of probabilistic distributions [26,29] and good at fitting multivariate signals as well. In transform domain WZ codec, the model parameters can be estimated at band-level or coefficient-level before Slepian–Wolf decoding. Yet it is more often of adopting band-level estimation since it can offer powerful statistical support [8]. That is, we assume all the coefficients in the same band follow the same probabilistic distribution.

In this context, the major contribution of this paper is to improve the performance for discrete cosine transform (DCT)-domain Wyner–Ziv coding (TDWZ) where the statistical dependency between the source and side information can be accurately characterized by means of GMMs. By considering the temporal correlation and quantization term, our proposed model, termed as correlation estimation based on GMM (CEGMM), can better characterize the correlation noise statistics. In particular, a GMM for band-level correlation noise is trained offline and thereafter utilized online to compute the conditional probability during LDPC decoding. A more reasonable two-component GMM model across all the DCT bands is trained by CEGMM for its strong capability of representing the relationship among these bands. Moreover, CEGMM takes the quantizing term into account so that the model is more practical than some of existing methods [8, 11, 18, 25, 37]. Then, we modify the DISCOVER [6] decoder by integrating the proposed CEGMM instead of Laplacian assumption. For a fair comparison, we do not change the processing on the side information generation. That is, no change at the encoder is made by our proposed DVC scheme w.r.t. classic DVC. Accordingly, the proposed DVC maintains low encoding complexity as well.

The rest of this paper is organized as follows. Section 2 reviews the related works on correlation noise statistics estimation and also provides the motivation of this work. In Sect. 3, we present the band-level correlation estimation based on GMM. The performance of CEGMM integrated in WZ codec is evaluated on estimation accuracy and RD performance in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Review on Correlation Noise Modeling

In DVC, the coding efficiency is mainly determined by the quality of side information and the accuracy in modeling the dependency between the source and the corresponding side information [8, 38]. The finer model for the dependency means that fewer accumulated syndrome bits are required to be sent to the decoder, resulting in better RD performance. In this section, we first briefly review the background and then introduce the motivation on characterizing the dependency called correlation noise model (CNM).

Figure 1 illustrates the TDWZ codec architecture [1] with our proposed model of the correlation noise, referred in the later discussion. At the encoder, the video sequence is partitioned into WZ frames and key frames. As the conventional way, the key frames are processed by exploiting H.264/AVC intra codec through DCT transform

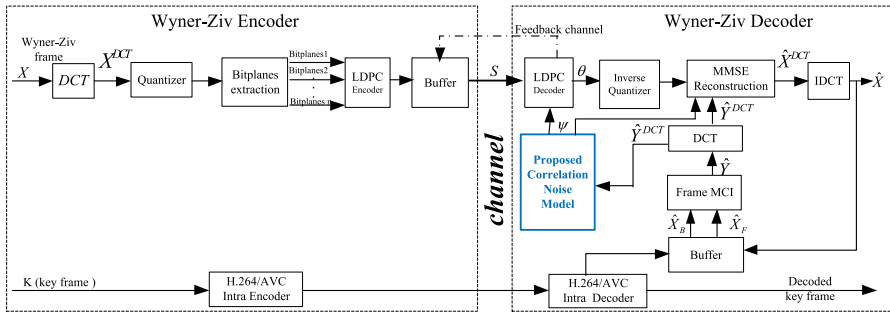


Fig. 1 Architecture of the DCT-domain Wyner–Ziv video coding with CEGMM

and quantization processing. For WZ frames, a 4×4 block-wise DCT is first applied and the DCT coefficients are grouped into different bands, which are quantized into the symbol stream. Then, the bit-planes of each band quantized symbols are extracted and fed into LDPC encoder, where the parity information is obtained. These syndrome bits are stored into a buffer to be sent on fixed amounts upon the decoder request. On the other hand, motion compensated frame interpolation (MCFI) is adopted to generate the SI (i.e., Y) with the previously decoded key frame and the WZ frame [41] at the decoder. Given a group of picture (GOP) size of 2, we denote the previous and next temporally adjacent key frames by \hat{X}_B and \hat{X}_F , respectively. Y^{DCT} is obtained by applying DCT over Y , which is the estimate corresponding to \hat{X}^{DCT} . In the optimal reconstruction given the output of LDPC decoder and SI [37], the CNM plays a key role in WZ video decoding. Y^{DCT} is not only fed into the module of correlation estimation to help LDPC decoder to decode the compressed stream, but also used to reconstruct the WZ frame in a minimum-mean-squared-error (MMSE) optimal way.

Some early works aim at modeling the probability distribution of correlation noise (CN) in TDWZ. In an earlier work [17], the error or residue $N = X - Y$ is offline exploited to model the dependency. However, it is unrealistic in practice because the original information is unavailable at the decoder and the lack of SI occurs at the encoder whilst. Then, the residue between the adjacent motion compensated frames replaces the error to model the correlation noise statistics [8].

Let (x, y) be the spatial coordinate within the current frame, (dx_B, dy_B) and (dx_F, dy_F) denote the associated backward and forward motion vector, respectively. According to the previous work [8, 11, 18, 25, 37], a motion compensated residue R between forward and backward interpolation is regarded as a realistic solution for modeling correlation noise instead of an unrealistic offline residue,

$$R(x, y) = \hat{X}_B(x + dx_B, y + dy_B) - \hat{X}_F(x + dx_F, y + dy_F). \tag{1}$$

In general, the motion-compensated residue DCT coefficient is modeled as Laplacian random variables [21] and its probability distribution is defined as follows:

$$f(R) = \exp\left(-\sqrt{2}|R - \mu|/\sigma\right) / \sqrt{2}\sigma, \tag{2}$$

where μ is the mean value and σ^2 is the variance of the residue.

Since the luminance of each pixel varies along with time and space, the parameter σ is likely to change over different parts of an image. This means that the actual DCT coefficient distribution will be affected by the change of σ^2 in different granularity levels [8, 11, 18, 21, 25, 37]. Then, in order to obtain a more accurate model, many methods have been proposed to estimate the parameters under the Laplacian assumption. For the pixel- and transform-domain WZ video codec, Brites and Pereira [8] proposed an online estimation of the CN model parameters in the decoder by exploiting the correlation within different levels of granularity: frame-, block-, and pixel-levels. Experiments show that the higher the estimation granularity is, the better the RD performance is. Deligiannis et al. [11] demonstrated that the CN distribution is dependent on the realization of the SI, then the authors further proposed a new technique by incorporating into a unidirectional spatial-domain DVC system. Mys et al. [25] improved the correlation noise estimation by taking into account the effect of quantization noise in intra frames. Results indicate that average Wyner–Ziv bit-rate reductions are up to 19.5% (Bjontegaard delta metric) for coarser quantization. By utilizing cross-band correlation to estimate the model parameters, Huang [18] proposed an improved noise model for TDWZ. Compared with the model at coefficient-level, the new statistics model is more robust and improves the RD performance for high bit-rates amounting to 0.5 dB.

Recently, the progressive refinement methods are proposed to estimate the model parameters. Fan et al. [15] proposed a novel transform-domain adaptive correlation estimation method, in which the model parameter, i.e., the variance at band-level, is learned in a progressive way. In [31], a progressive correlation noise refinement method is proposed for transform-domain Wyner–Ziv coding. The parameter is continuously updated during the decoding process. Based on the accuracy of the side information, the correlation estimation method is proposed by differentiating blocks within a frame [14]. In [36], the correlation noise statistics estimation is processed jointly with belief propagation based LDPC decoding. By exploiting the spatial correlation and quantization distortion, Skorupa et al. [30] developed a correlation model that is able to adapt the changes of the content and coding parameters. Instead of assuming Slepian–Wolf encoded bit-planes to be memoryless source, Toto-Zarasoia et al. [32] considered a predictive correlation model together with a Gilbert–Elliott (GE) memory source. In [12], a side-information-dependent (SID) model, rather than side-information-independent (SII), of correlation channel is proposed to improve the Wyner–Ziv coding performance. A cross-band based adaptive noise model is proposed for TDWZ video coding [19] and the noise residue is successively updated. More recently, Deligiannis et al. proposed a novel correlation channel estimation method designed for generic layered WZ coding [13].

In the above work, people usually assume that the probability distribution of CN belongs to the Laplacian family, however, it has been found that the Laplacian family is inaccurate in modeling dependency because the actual distribution of the DCT coefficients differs from the Laplacian distribution in some cases [21, 33]. The major reason why the Laplacian distribution is used in modeling dependency is due to the simplicity of the Laplacian function in deriving mathematical formulations for the correlation noise statistics. In a nutshell, the inaccurate assumption motivates us to find new models to better characterize the statistics.

During the creation process of side information, MCFI is carried out based on the assumption that the motion is translational and linear over time among temporally adjacent frames [14]. Yet this assumption is not always true. Moreover, the key frames used for interpolating are unavoidably contaminated by quantization noise. Thus, the correlation noise is non-stationary [36], which is mainly caused by two factors such as the incoherent motion field and quantization noise [25] in key frames. These multiple factors motivate this paper to apply a GMM to describe the correlation noise.

3 Band-Level Correlation Estimation Based on Gaussian Mixture Model

The Laplacian model does not fit certain distributions well, while other models, such as generalized Gaussian distribution [21], fail to be a good tradeoff between model accuracy and computational complexity. Nevertheless, it is suggested that the Gaussian family is a realistic model in practice. To this end, GMM is introduced here for the correlation noise statistics with its universality. GMM is a statistically mature model defined as a mixture of components, each of which is a Gaussian probability distribution [40]. We have witnessed the success of this model in different domains [27]. It is naturally expected GMM is able to model the correlation noise well in this work.

3.1 The Motivation

Let (x, y) be the spatial coordinate within one frame and the correlation noise at location (x, y) be $N(x, y) = X(x, y) - Y(x, y)$, which is assumed to be determined by a probabilistic density function

$$p(N(x, y)) = p(N_{MV}(x, y), N_Q(x, y)). \quad (3)$$

Generally, N can be regarded as the result yielded by two factors, N_{MV} and N_Q . Next, we detail why these two factors can affect the distribution of the correlation noise. First, motion compensated interpolation (MCI) is adopted to generate the side information at the decoder [6]. In MCI, a motion vector field is usually obtained by motion estimation. Unfortunately, the motion vector is hard to be accurately estimated without the original frame at the DVC decoder. Specifically, the motion field close to the true motion is estimated by backward \hat{X}_B and forward \hat{X}_F . This case will become worse for the video sequence with high motion. Under such circumstances, the motion vector is definitely different from the true one, which results in larger error in the side information and further affects the correlation noise. Here, we let N_{MV} denote the noise caused by the erroneous motion vector, and N_Q the noise component introduced by the reconstructed key frames with quantization noise during the process of motion compensation interpolation. At the decoder with the absence of original frames, the case will get worse due to the noise sources N_Q during the process of motion compensation interpolation. In most WZ codec solutions, forward and backward interpolations usually do produce the block artifacts which make contributions to N_Q .

Because of the presence of multiple noise sources, many simple noise models naturally fail to characterize the correlation noise statistics well. Laplacian family

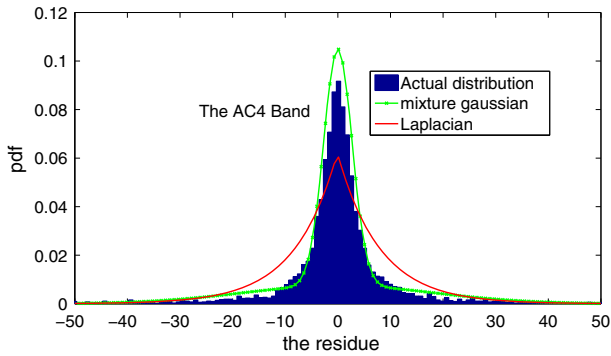


Fig. 2 Residue histogram for AC4 band of *Soccer*

noise models are widely used in modeling noise process in video coding [8,37,38,41]. Figure 2 shows a scenario where a Laplacian distribution wrongly models the correlation noise statistics.

Actually the correlation noise is a complicated random process and dynamically changes over time due to the varying video contents. For example, in a recent work [42], the authors decomposed the transmission error into the error caused by motion vectors and another one caused by prediction residue loss under packet-loss environment. Based on the definition of correlation noise in the previous work [17], we take the ideal channel case into account in this paper, i.e., the error-free channel, for evaluating the accuracy of the new model. It is worth noting that this work is different from our prior work [24] in several aspects. We apply statistical learning theory to accurately analyze the factors influencing the correlation noise. In particular, we propose a more reasonable two-component model to describe the correlation noise statistics. To validate the efficacy of our approach, more extensive experiments of the test video sequences are carried out.

3.2 The Model

Generally, when there are many possible random components, it is more natural to model the probability of the noise generating process as multi-mode by a mixture density function [40] which, for example, depends on the coherence of motion field and the energy of prediction residue,

$$p(N) = p(N_{MV}, N_Q) = \sum_i p(N_{MV}^i | N_Q^i) p(N_Q^i), \quad (4)$$

where i indexes the possible modes or mixture components. Then, the modeling problem is transferred into describing the distribution of multivariate stochastic variables, specifically two stochastic variables, in which each one obeys a Gaussian distribution.

However, the noise process N is unavailable in practice due to the absence of the original frame at the decoder. Thus, we have to use the transformed residue R obtained

from Eq. (1) to model the correlation noise instead. As usual, a 4×4 DCT is applied on WZ frame, thus the transformed coefficients of WZ frame are grouped into 16 bands, denoted by R_k ($k = 0, \dots, 15$). There may exist dependency among these bands. In order to effectively model the noise process across bands, we consider the random vector $\mathbf{R} = (R_0, R_1, \dots, R_{15})^T$ and model its probability distribution by the classic GMM in the following form:

$$p(\mathbf{R}) = \sum_{m=1}^M \pi_m \mathcal{N}(\mathbf{R}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \text{ with } \sum_{m=1}^M \pi_m = 1, \quad (5)$$

where $\{\pi_m \geq 0\}_{m=1}^M$ are the mixing weights of M components which sums to one totally and each $\mathcal{N}(\mathbf{R}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ ($m = 1, \dots, M$) is the Gaussian probability density function defined as

$$\mathcal{N}(\mathbf{R}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_m|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{R} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{R} - \boldsymbol{\mu}_m)\right) \quad (6)$$

with $\boldsymbol{\mu}_m$ as the mean and $\boldsymbol{\Sigma}_m$ the covariance matrix, respectively. d is the dimension of random vector \mathbf{R} in the distribution function.

As argued in (3), there assumably exist two major noise sources N_{MV} and N_Q . Thus, it is reasonable to use a mixture model (5) with only two Gaussian components, each of which models one of major noise sources. For the sake of simplicity, we write out the proposed model of the Gaussian mixture distribution as follows:

$$p(\mathbf{R}) = \pi_1 \mathcal{N}(\mathbf{R}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{R}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2). \quad (7)$$

The model parameters defined in (7) are $\Theta = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$. Given a set of K training data $\mathcal{D} = \{\mathbf{R}^i\}_{i=1}^K$, the logarithm of the likelihood function [35] can be written as

$$\mathcal{L}(\Theta) = \sum_{i=1}^K \log\left(\pi_1 \mathcal{N}(\mathbf{R}^i|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{R}^i|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right). \quad (8)$$

Maximizing $\mathcal{L}(\Theta)$ w.r.t. Θ according to (8) is very hard due to the log of a summation, however, in order to guarantee an increasing log-likelihood from iteration to iteration, the expectation-maximum (EM) algorithm [35] is usually applied to GMM parameter estimation.

3.3 Model Fitting Using Generalized EM

To avoid the sensitivity to initialization of EM algorithm, a greedy learning technique, similar to [35], is adopted to obtain model (7). In our case, the missing data is the Gaussian cluster, i.e., the band coefficients cluster, to which the data points belong. We predict values to fill in for the missing data (the E-step), calculate the maximum



Fig. 3 Sample frames for testing sequences

likelihood parameter estimates with this data (the M-step), and repeat until a stopping criterion is met. Considering the paper limitation, we here omit the intermediate EM process. (Please refer to [35] for the details.)

In order to decrease the computational complexity in practical decoding, we fit the model at offline stage and then the model parameters are used to calculate the soft-input information [6]. In our model training, the training data are collected from several WZ frames. For example, for frequency band k , the residue of WZ frames coming from different sequences are grouped to form the training set. As illustrated in Fig. 3, the training sequences in this research include *Foreman*, *Soccer*, *Football*, *Hall Monitor*, *Carphone*, and *Coastguard*, which covers low and high motion contents. To figure out the model parameters, we randomly select 10 WZ frames from each sequence and 60 frames in total are used as training set. Each band data is grouped by collecting the DCT coefficients within the same frequency band along all the training samples. Given 16 bands, we ensure each band includes approximate 100,000 data to train GMM parameters. For the initialization of the model parameter, we apply the mean value of source data for initializing the mean of the mixture components. As for the covariances of the components, the individual covariance matrices for the components are created by adding different small positive numbers to the eigenvalues of the estimated source covariance matrix. The weight is initialized to be 0.5. When using EM algorithm to estimate model parameter, 20 iterations are run to achieve convergence at least. The iteration threshold of EM algorithm is set $1e-5$ for stopping the iteration.

Once the model has been trained, we build up a 16-element lookup table at the decoder, which can be done offline. For the sake of simplicity, we only use the diagonal elements of Σ_1 and Σ_2 as the variances of individual bands. When calculating the

conditional probability within LDPC decoding, the model parameter of correlation noise in each frequency band is found through the lookup table.

4 Experimental Results

In this section, we perform several experiments on some classic video sequences, such as *Foreman* (without the Siemens logo), *Soccer*, *HallMonitor*, *Carphone*, and *Coastguard*, to evaluate the performance of CEGMM. These video sequences cover a variety of contents which are helpful to obtain meaningful and representative results. Their spatial resolution is Quarter common intermediate format (QCIF) and the GOP length is typically selected as 2, 4, and 8, respectively. The frames of test sequences are chosen as follows, 299 frames for *Foreman*, 329 frames for *HallMonitor*, 299 frames for *Coastguard*, 382 frames for *Carphone*, and 299 frames for *Soccer*. The frame rate of all test sequences is 30 Hz.

First, we verify the accuracy of correlation estimation via applying CEGMM and Laplacian model, respectively. Then, the RD performance is compared with the different correlation estimation methods integrated into a DVC system. The state-of-the-art DISCOVER codec and several standard video coding solutions, such as H.263+ Intra, H.264/AVC Intra, and H.264/AVC Inter No Motion, are exploited as benchmark for performance comparison. Although H.263+ Intra is not with the best performance among the available Intra codecs, it is still widely used as an acceptable benchmark in WZ coding literature. H.264/AVC Intra is the most efficient Intra coding standard which encodes H.264/AVC main profile without exploiting temporal redundancy, while the H.264/AVC Inter No Motion does exploit the temporal redundancy in IBIB structure without any motion estimation.

To fairly compare the performance of our method to DISCOVER codec, we adopt the similar process of side information and reconstruction as specified in [7, 9, 20] in our experiments. Similar to the previous literature, only the luminance data is considered for evaluation in all of the experiments.

4.1 Evaluation of The Model Accuracy

The difference between two probability distributions can be evaluated by the Kullback–Leibler (K–L) divergence, which is popular for verifying the accuracy of different models [33],

$$KL(p||q) = \int p(x) \log_2(p(x)/q(x))dx, \quad (9)$$

where p is the actual probability density function (PDF) and q is the modeling PDF. In this experiment, we use the histogram of the residue instead of true PDF because the original frame is unavailable at the decoder. The GOP length is fixed as 2. As the lower band coefficients conserve majority energy in a frame, we only give the K–L results for the bands from DC to AC7. A small K–L divergence means a good modeling. Table 1 shows the divergence between the real PDF and the proposed GMM modeling

Table 1 K–L divergences of two models for the bands DC to AC7 of the sequences

Model	DC	AC1	AC2	AC3	AC4	AC5	AC6	AC7
<i>Foreman</i>								
Laplacian	0.089874	0.076229	0.121264	0.156922	0.147241	0.204272	0.536842	0.073582
GMM	0.029159	0.041795	0.037046	0.059698	0.049863	0.062125	0.176621	0.04774
<i>Soccer</i>								
Laplacian	0.07329	0.180744	0.132247	0.119882	0.17737	0.076989	0.177669	0.113129
GMM	0.027391	0.031723	0.036303	0.049572	0.060764	0.033322	0.079919	0.058289
<i>HallMonitor</i>								
Laplacian	0.229777	0.259666	0.103524	0.1255	0.138912	0.205425	0.201605	0.183249
GMM	0.022357	0.038524	0.026195	0.042771	0.032082	0.051645	0.061955	0.056873
<i>Carphone</i>								
Laplacian	0.050924	0.142395	0.205634	0.203662	0.237339	0.398695	0.37894	0.394379
GMM	0.017598	0.036001	0.051244	0.055992	0.062186	0.083519	0.096483	0.098783

for several test cases. We can observe, from the table, that the proposed model matches the real PDF much better than Laplacian model in most bands.

Figures 4, 5, and 6 show the PDF of the residual bands and their approximations given by Laplacian and the proposed GMM for *Carphone*, *Soccer*, and *Foreman*, respectively. The tested frames contain relative complex motion contents. The PDF fitted by Laplacian has long tails, leading to a slowly decaying compared to the actual distribution at each band. From the figures, the actual DCT coefficient band has a small tail and its distribution cannot be well approximated by a zero-mean Laplacian distribution, however, this is the assumption used in many previous works [8, 11, 37, 38]. In contrast, the PDF using CEGMM presents similar behavior as the actual distribution, for which the tail of the density decays faster.

4.2 Evaluation on RD Performance

After verifying the model accuracy, we will test the RD performance of our DVC system incorporating CEGMM in this section. The RD scores are computed corresponding to different values of quantization matrices [9] used at the encoder. In particular, the quantization parameters (QP) used to quantize the key frames are chosen as in Table 2. In addition, the plotted RD curves include the average rate and distortion measured for both the key frames and the WZ frames.

The experimental results for the sequences are shown in Figs. 7, 8, and 9. Compared to the Laplacian model employed in DISCOVER codec, CEGMM achieves a better and more consistent performance in all the cases. Figure 7 presents the RD charts for *Coastguard* and *Foreman*, respectively, whose contents include the well behaved and high motion scenes. When the GOP length is 2, the RD gain obtained by the proposed model is relatively small for high bit-rate case up to 0.48 dB since the quality of side information after frame interpolation is good enough. With the GOP length increasing,

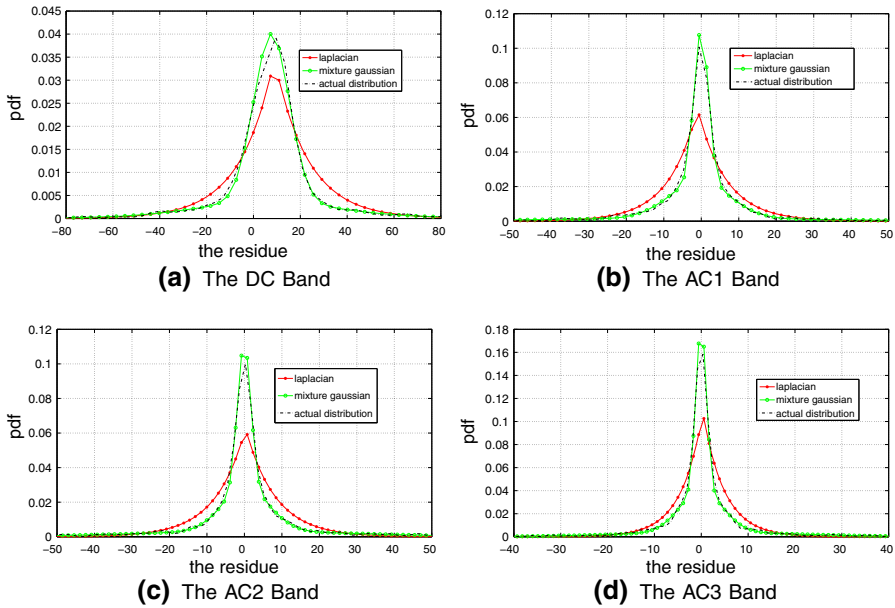


Fig. 4 The PDF of the residual band and their approximations by Laplacian and the proposed GMM for *Carphone* (frame 182)

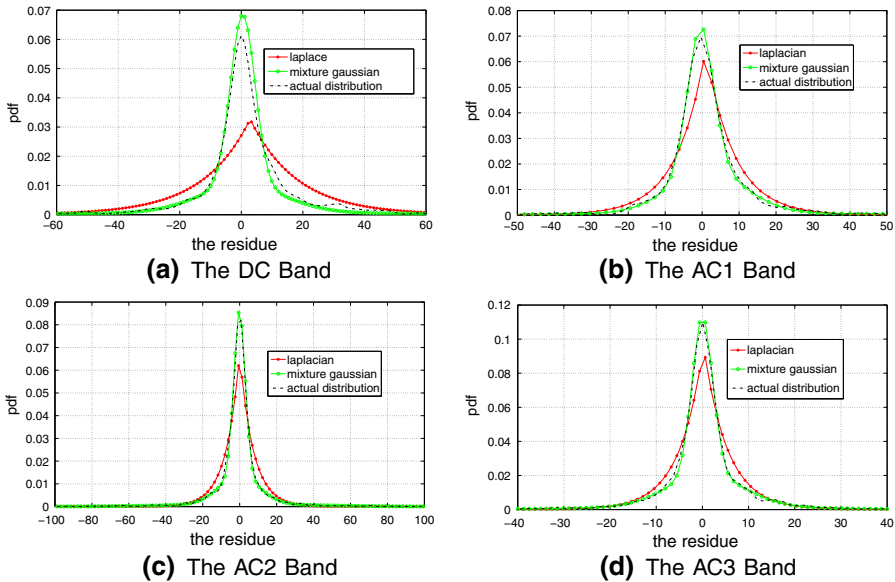


Fig. 5 The PDF of the residual band and their approximations by Laplacian and the proposed GMM for *Soccer* (frame 286)

the quality of side information degrades and leaves much room for improvements of RD performance. From Figs. 8 and 9, it is observed that the good RD gains can be achieved by our proposed method. As for the more complex motion sequences, such as

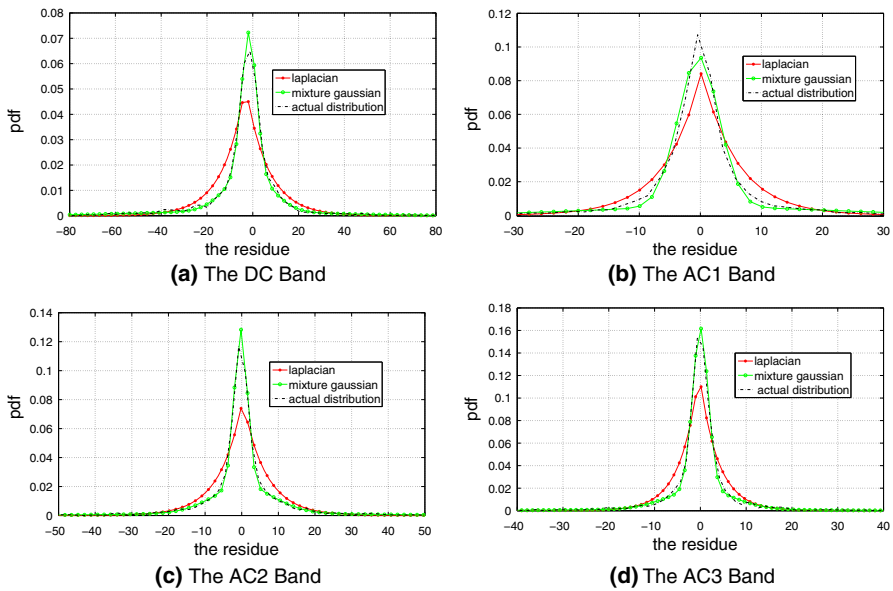


Fig. 6 The PDF of the residual band and their approximations by Laplacian and the proposed GMM for *Foreman* (frame 152)

Table 2 Quantization parameters for key frames in different RD points, QCIF at 30Hz

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
<i>Foreman</i>	40	38	38	34	34	32	29	25
<i>Soccer</i>	44	43	41	36	36	34	31	25
<i>HallMonitor</i>	37	36	36	33	33	31	39	24
<i>Coastguard</i>	38	37	37	34	33	31	30	26

Soccer and *Foreman*, with the Laplacian assumption, the decoder usually needs more bits to correct the error to achieve a satisfactory error ratio. Thus, compared to the Laplacian model integrated in DISCOVER codec, the proposed CEGMM outperforms up to 2.23 dB for *Foreman* and 2.41 dB for *Soccer* at a GOP size of 8, respectively. For the low motion sequence, such as *Hall monitor*, our RD gain is consistently achievable though the improvement is not as significant as for complex video sequences. This is reasonable due to, as the high quality of the SI is achieved, the effect of accurate correlation estimation on the RD performance decreases. As for the gain against the DISCOVER codec, it can be explained by the reduction of the LDPC decoder in requiring fewer accumulated syndrome bits to correct the estimation errors.

In addition, RD performance of the proposed method is also evaluated by comparing with that of other three standard coding solutions. From the results, we conclude that the WZ codec based on CEGMM outperforms H.264/AVC Intra for the sequences with the low motion content, especially for the longer GOP length. However, it is only better than the H.264/AVC No Motion codec for *Coastguard* approximately 1 dB at most. The gap between WZ codec and the standard video codecs has been much

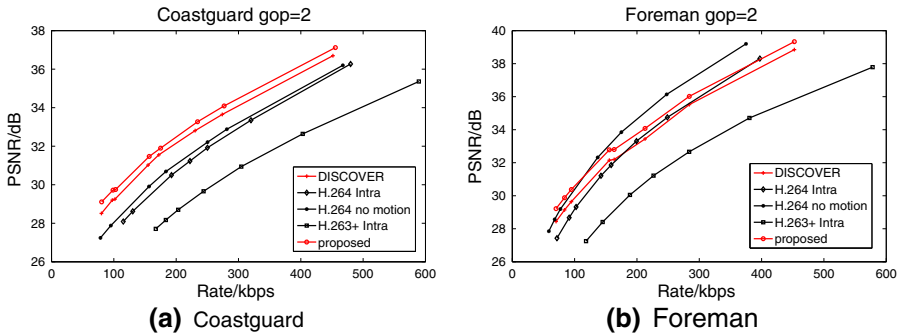


Fig. 7 RD performance for *Coastguard* and *Foreman* at GOP = 2

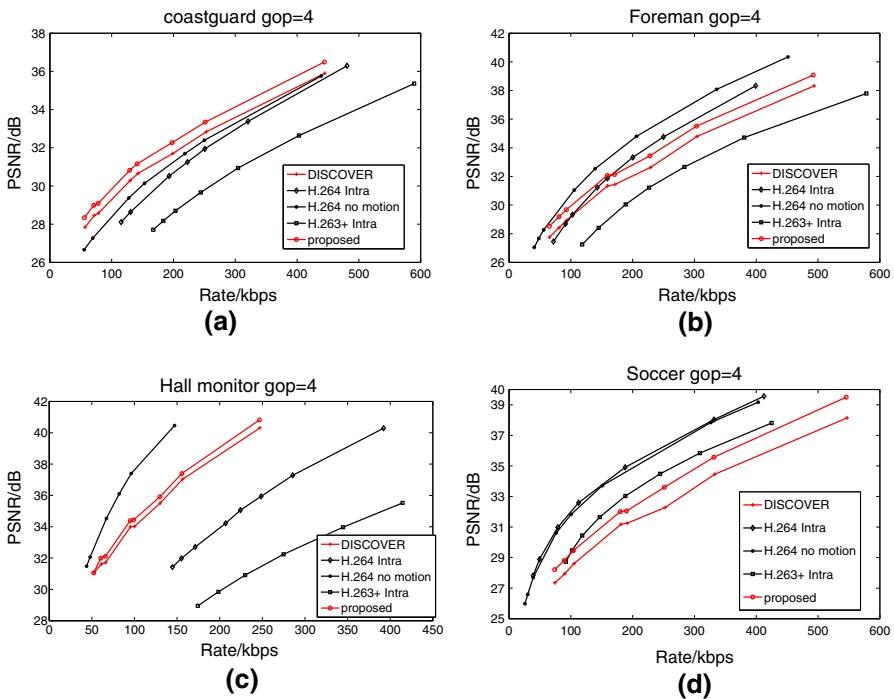


Fig. 8 RD performance for the different sequences at GOP = 4

shortened for the sequences with high motion content, such as *Foreman* and *Soccer*. As for *Foreman* at GOP 8, our RD curve is close to H.264/AVC Intra, while that of *Soccer* is close to H.263+ Intra. However, our RD performance for high motion content sequence is inferior to H.264 No Motion, especially with a larger GOP size.

4.3 Complexity Evaluation

Finally, the decoding algorithm complexity measured by CPU running time for CEGMM is compared to that of the benchmark DISCOVER [6] in Table 3. The

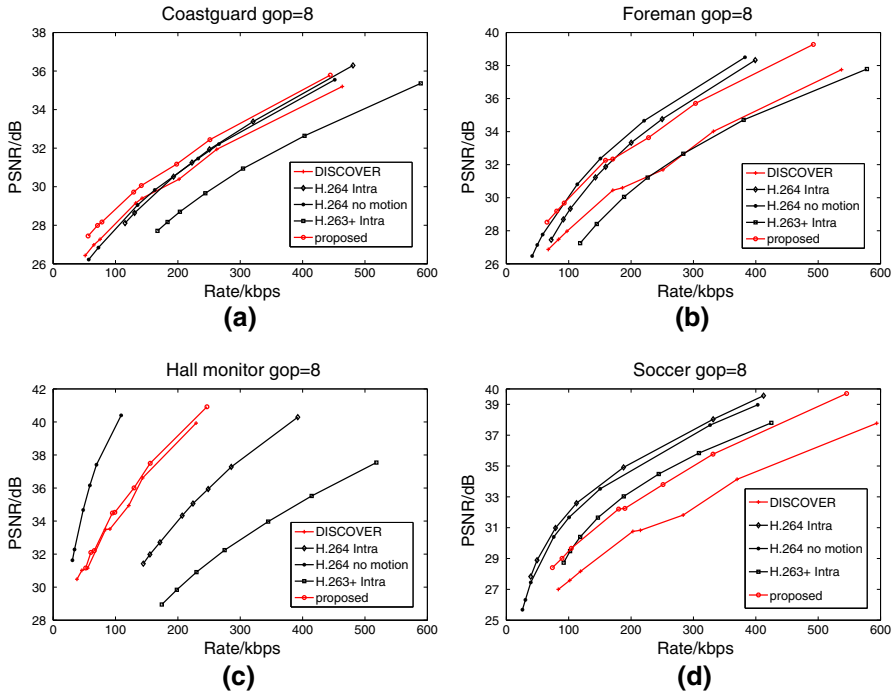


Fig. 9 RD performance for the different sequences at GOP = 8

execution time of DISCOVER is obtained using the executable code, as found on the DISCOVER website. Since our proposed method is only intergraded into the decoder, the encoding complexity is as same as DISCOVER. It is well known that the execution time is highly dependent on the conducted hardware and software platforms. In our experiments, the simulation is carried out on an $\times 86$ machine with a Intel core processor at 2.13 GHz with 3.0 GB of RAM. In Table 3, three decoding complexity comparison is provided according to different quantization matrices at GOP 2 at a frame rate of 15 Hz. In addition, a suitable optimized for speed is applied compared with the traditional DISCOVER decoder.

From the results, it is observed that the proposed method consumes less execution time than DISCOVER codec. This is due to the fact that the parameter learning is conducted offline, so that the time taken by the decoding is not much higher than DISCOVER. Contrary to the demanding operations of side information creation and LDPC decoding, the computational time of the estimation of correlation noise is relatively limited [12]. In light of this point, the decoding time mainly relies on the LDPC decoding. In other words, one can expect a less computational complexity can be achieved by fewer LDPC decoding iterations. In our work, the reduction of the decoding time can be owed to the two factors, i.e., the quantization term and the mixture model for the residual data. As a result, the LDPC decoder requires fewer feedback channel requests and decoding operations.

Table 3 Decoding time for test sequences at 15 Hz (unit: s)

Index	<i>Foreman</i>		<i>HallMonitor</i>		<i>Soccer</i>		<i>Coastguard</i>	
	DISCOVER	CEGMM	DISCOVER	CEGMM	DISCOVER	CEGMM	DISCOVER	CEGMM
Q1	508.56	413.97	295.15	279.21	240.67	210.83	723.66	659.98
Q3	709.72	577.71	475.67	449.03	371.01	352.00	1,010.78	921.83
Q5	1,290.25	1,058.01	820.48	772.07	605.84	530.72	1,771.77	1,617.62

5 Conclusion and Future Work

In this paper, a GMM-based correlation estimation is proposed to characterize the band-level correlation noise statistics in TDWZ by simultaneously considering the temporal correlation and quantization distortion. The difference between WZ frame and its corresponding side information is first analyzed, and the noise in transformed coefficients in each band is characterized by a two-component Gaussian mixture distribution. The results of K–L divergence show that CEGMM has better accuracy than the Laplacian version. Experimental results also show that an improvement on RD performance is achieved against the Laplacian model in DISCOVER codec, especially for high motion content sequences and longer GOP length. Compared with the alternative standard video coding solutions, the WZ codec with CEGMM outperforms H.264/AVC Intra coding for the sequences with low motion content. However, the RD performance for high motion content sequence is inferior to H.264 No Motion coding, especially with a larger GOP size. Therefore, our future work includes further improvements targeting at higher coding efficiency.

Acknowledgments The work of the second author was partially supported by the Commonwealth of Australian under the Australian-China Science and Research Fund (ACSRF01222) and the Australian Research Council (ARC) under Discovery Project Grant DP130100364. The work was also supported by the National Science Foundation of China, under Grant 61201392.

References

1. A. Aaron, S. Rane, E. Setton, B. Girod, Transform-domain Wyner–Ziv codec for video, in *Proceedings of SPIE Visual Communications and Image Processing*, pp. 520–528, 2004
2. A. Abou-Elailah, F. Dufaux, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, Fusion of global and local motion estimation for distributed video coding. *IEEE Trans. Circ. Syst. Video Technol.* **23**(1), 158–172 (2013)
3. F. Akyildiz, T. Melodia, K.R. Chowdhury, A survey on wireless multimedia sensor networks. *Comput. Netw.* **51**(4), 921–960 (2007)
4. I.F. Akyildiz, T. Melodia, K.R. Chowdhury, Wireless multimedia sensor networks: a survey. *IEEE Wirel. Commun. Mag.* **14**(6), 32–39 (2007)
5. G. Anastasi, M. Conti, M. Francesco, A. Passarella, Energy conservation in wireless sensor networks: a survey. *Ad Hoc Netw.* **7**(3), 537–568 (2009)
6. X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, M. Ouaret, The DISCOVER codec: architecture, techniques and evaluation, in *Proceedings of Picture Coding Symposium*, 2007
7. J. Ascenso, C. Brites, O. Pereira, Content adaptive Wyner–Ziv video coding driven by motion activity, in *IEEE International Conference on Image Processing*, 2006

8. C. Brites, F. Pereira, Correlation noise modeling for efficient pixel and transform domain Wyner–Ziv video coding. *IEEE Trans. Circ. Syst. Video Technol.* **18**(9), 1117–1190 (2008)
9. C. Brites, J. Ascenso, J. Quintas Pedro, F. Pereira, Evaluating a feedback channel based transform domain Wyner–Ziv video codec. *Signal Process. Image Commun.* **23**(4), 269–297 (2008)
10. H. Chen, E. Steinbach, Wyner–Ziv video coding based on turbo codes exploiting perfect knowledge of parity bits, in *IEEE International Conference on Multimedia & Expo, ICME 2007*, Beijing, China, 2007
11. N. Deligiannis, A. Munteanu, T. Clerckx, J. Cornelis, P. Schelkens, Correlaiton channel estimation in pixel-domain distributed video coding, in *Proceedings of 10th International Workshop on Image Analysis for Multimedia Interactive Services*, 2009
12. N. Deligiannis, J. Barbarien, M. Jacobs, A. Munteanu, A.N. Skodras, P. Schelkens, Side-information-dependent correlation channel estimation in hash-based distributed video coding. *IEEE Trans. Image Process.* **21**(4), 1934–1949 (2012)
13. N. Deligiannis, A. Munteanu, S. Wang, S. Cheng, P. Schelkens, Maximum likelihood laplacian correlation channel estimation in layered Wyner–Ziv coding. *IEEE Trans. Signal Process.* **62**(4), 892–904 (2014)
14. G. Esmaili, P. Cosman, Wyner–Ziv video coding with classified correlation noise estimation and key frame coding mode selection. *IEEE Trans. Image Process.* **20**(9), 2463–2474 (2011)
15. X. Fan, O.C. Au, N.M. Cheung, Transform-domain adaptive correlation estimation (TRACE) for Wyner–Ziv video coding. *IEEE Trans. Circ. Syst. Video Technol.* **20**(11), 1423–1436 (2010)
16. S. Fang, Z. Li, L.W. Zhang, Distributed video codec modeling correlation noise in wavelet coarsest subband. *Electron. Lett.* **43**(23), 1266–1267 (2007)
17. B. Girod, A. Aaron, S. Rane, D. Rebollo-Monedero, Distributed video coding. *Proc. IEEE Spec. Issue Video Coding Deliv.* **93**(1), 71–83 (2005)
18. X. Huang, Improved virtual channel noise model for transform domain Wyner–Ziv video coding, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 921–924, 2009
19. X. Huang, S. Forchhammer, Cross-band noise model refinement for transform domain Wyner–Ziv video coding. *Signal Process. Image Commun.* **27**(1), 16–30 (2012)
20. D. Kubasov, J. Nayak, C. Guillemot, Optimal reconstruction in Wyner–Ziv video coding with multiple side information, in *Proceedings of the IEEE 9th Workshop on Multimedia Signal Processing*, pp. 183–186, 2007
21. E.Y. Lam, J.W. Goodman, A mathematical analysis of the DCT coefficient distributions for images. *IEEE Trans. Image Process.* **9**(10), 1661–1666 (2000)
22. H.V. Luong, L.L. Raket, X. Huang, S. Forchhammer, Side information and noise learning for distributed video coding using optical flow and clustering. *IEEE Trans. Image Process.* **21**(12), 4782–4796 (2012)
23. T. Maugey, J. Gauthier, B. Pesquet-Popescu, C. Guillemot, Using an exponential power model for Wyner–Ziv video coding, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2338–2341, 2010
24. M. Yin, S. Cai, Y. Xie, Wyner–Ziv video coding based on Gaussian mixture model. *Chin. J. Comput.* **35**(1), 173–182 (2012)
25. S. Mys, J. Skorupa, P. Lambert, R. Van de Walle, C. Grecos, Accounting for quantization noise in online correlation noise estimation for distributed video coding, in *Proceedings of Picture Coding Symposium*, pp. 1–4, 2009
26. S. Nadarajah, Gaussian DCT coefficient models. *Acta Appl. Math.* **106**, 455–472 (2009)
27. D. Persson, T. Eriksson, P. Hedelin, Packet video error concealment with Gaussian mixture models. *IEEE Trans. Image Process.* **17**(2), 145–154 (2008)
28. R. Puri, A. Majumdar, P. Ishwar, K. Ramchandran, Distributed video coding in wireless sensor networks. *IEEE Signal Process. Mag.* **23**(4), 94–106 (2006)
29. C. Sanderson, K.K. Paliwal, Fast features for face recognition under illumination direction changes. *Pattern Recogn. Lett.* **24**(14), 2409–2419 (2003)
30. J. Skorupa, J. Slowack, S. Mys, N. Deligiannis, J.D. Cock, P. Lambert, A. Munteanu, R.V. de Walle, Exploiting quantization and spatial correlation in virtual-noise modeling for distributed video coding. *Signal Process. Image Commun.* **25**(9), 674–686 (2010)
31. J. Song, K. Wang, H. Liu, Y. Li, C. Wu, Progressive correlation noise refinement for transform domain Wyner–Ziv video coding, in *IEEE International Conference on Image Processing*, pp. 2625–2628, 2011

32. V. Toto-Zarasoia, A. Roumy, C. Guillemot, Source modeling for distributed video coding. *IEEE Trans. Circ. Syst. Video Technol.* **22**(2), 174–187 (2012)
33. C.Y. Tsai, H.M. Hang, ρ -GGD source modeling for wavelet coefficients in image/video coding, in *IEEE International Conference on Multimedia & Expo (ICME)*, pp. 601–604, 2008
34. X. Van Hoang, B. Jeon, Flexible complexity control solution for transform domain Wyner–Ziv video coding. *IEEE Trans. Broadcast.* **58**(2), 209–220 (2012)
35. J.J. Verbeek, N. Vlassis, B. Kröse, Efficient greedy learning of Gaussian mixture models. *Neural Comput.* **15**(2), 469–485 (2003)
36. S. Wang, L. Cui, L. Stankovic, V. Stankovic, S. Cheng, Adaptive correlation estimation with particle filtering for distributed video coding. *IEEE Trans. Circ. Syst. Video Technol.* **22**(5), 649–658 (2012)
37. R.P. Westerlaken, R.K. Gunnewiek, R.L. Lagendijk, The role of the virtual channel in distributed source coding of video. *IEEE Int. Conf. Image Process.* **1**, 581–584 (2005)
38. Z. Xiong, A.D. Liveris, S. Cheng, Distributed source coding for sensor networks. *IEEE Signal Process. Mag.* **21**(5), 80–94 (2004)
39. Y. Yang, S. Cheng, Z. Xiong, W. Zhao, Wyner–Ziv coding based on TCQ and LDPC codes. *IEEE Trans. Commun.* **57**(2), 376–387 (2009)
40. G. Yazbek, C. Mokbel, G. Chollet, Video segmentation and compression using hierarchies of Gaussian mixture models, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I-1009, 2007
41. S. Ye, M. Ouaret, F. Dufaux, T. Ebrahimi, Improved side information generation for distributed video coding by exploiting spatial and temporal correlations. *EURASIP J. Image Video Process.* **1–15**, 2009 (2009)
42. Y. Zhang, H. Xiong, Z. He, S. Yu, C.W. Chen, An error resilient video coding scheme using embedded Wyner–Ziv description with decoder side non-stationary distortion modeling. *IEEE Trans. Circ. Syst. Video Technol.* **21**, 498–512 (2011)