中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-08-003

Power-Rate-Distortion Optimized Resource Allocation for Low-Complexity Multiview Distributed Video Coding

Li-Wei Kang (康立威) and Chun-Shien Lu (呂俊賢)

Mar. 31, 2008 || Technical Report No. TR-IIS-08-003
http://www.iis.sinica.edu.tw/page/library/LIB/TechReport/tr2008/tr08.html

Power-Rate-Distortion Optimized Resource Allocation for Low-Complexity Multiview Distributed Video Coding[+]

Li-Wei Kang and Chun-Shien Lu*

Institute of Information Science, Academia Sinica

Taipei, Taiwan 115, R.O.C

Abstract**—**Wireless visual sensor networks are potentially applicable for several emerging applications. Since the data size of the video captured from multiple sensors increases in proportion to the number of video sensors, the efficient compression of video data from multiple sensors is important and still challenging. However, most current multiview video coding approaches extended from single-view video coding standards perform both interview and temporal predictions at the encoder with very high computational complexity, which is not suitable for resource-limited video sensors. In this paper, a resource-scalable low-complexity multiview distributed video coding scheme is proposed. We study efficient exploitation of interview correlation by exchanging the media hash data extracted from video frames of adjacent video sensor nodes at the encoder and using the global motion parameters estimated and fed back from the decoder to improve coding efficiency. In addition, we present a power-rate-distortion (PRD) model to characterize the relationship between the available resources (*e.g.*, power supply and target bit rate) and the RD performance. More specifically, an RD function in terms of the percentages for different coding modes of blocks and the target bit rate under the available resource constraints is derived for optimal block coding mode decision. Analytic results are provided to verify the resource scalability and accuracy of the proposed PRD model, which can provide a theoretical guideline for performance optimization in low-complexity video coding under limited resource constraints. The coding efficiency of the proposed low-complexity video codec is demonstrated via simulation results to outperform three known low-complexity video codecs, especially at high power and low bit rates.

Index Terms**—**Low-complexity video coding, multiview distributed video coding, resource-scalable video coding, power-rate-distortion analysis, wireless visual sensor networks.

+This work was supported in part by National Science Council, Taiwan, ROC, under Grants NSC95-2422-H-001-031 and NSC 95-2221-E-001-022. Preliminary results of this paper were published in [33]-[35].

* Corresponding author: lcs@iis.sinica.edu.tw, Tel: 886-2-2788-3799 ext. 1513.

I. Introduction

A. Background

With the availability of low-cost hardware, such as CMOS cameras, wireless visual sensor networks (WVSNs) have potential to promote several emerging applications, such as security monitoring and environmental tracking [1]. As in a WVSN shown in Fig. 1, several battery-powered video sensor nodes (VSNs) are usually scattered in a sensor field. Each VSN equipped with a camera can capture and encode visual information along with delivering the compressed video data to the aggregation and forwarding node (AFN). The AFNs aggregate and forward the video data to the remote control unit (RCU), which can usually support a powerful decoder for video decoding and further information processing [1]. Compared with traditional network systems, WVSN operates under several resource constraints (*e.g.*, lower computational capability, limited power supply, and narrow transmission bandwidth). Hence, in a WVSN, low-complexity and high-efficiency video compression is critically desired for a VSN.

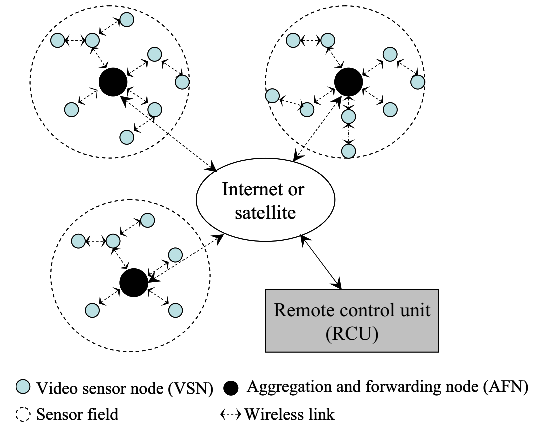

Internet or satellite
Remote control unit (RCU)
○ Video sensor node (VSN)
● Aggregation and forwarding node (AFN)
◌ Sensor field
↔ Wireless link

Fig. 1. A wireless visual sensor network (WVSN) architecture.

To achieve efficient video compression, current single-view video coding standards (*e.g.*, MPEG-X and H.26X [2]-[3]) usually perform complex interframe encoding operations (*e.g.*, block-based motion estimation) at the encoder to exploit temporal correlation of successive frames in a video sequence. On the other hand, current multiview video coding approaches [4]-[8], extended from the single-view coding standards, usually perform both interview (*e.g.*, disparity estimation) and temporal (*e.g.*, motion estimation) predictions at the encoder with very high computational complexity, which is not permitted in a resource-limited VSN. In addition, to perform interview video coding at the encoder, one must perform inter-VSN communication, *i.e.*, uncompressed frames exchanges, which is difficult for a WVSN.

To meet the requirements of resource-limited VSNs, it has recently been very popular to study low-complexity video encoding, which can be roughly classified into two categories: (i) distributed video coding (DVC) [9]-[20] and (ii) collaborative video coding and transmission [21]-[22]. Both will be described briefly in Secs. II-A and II-B, respectively. In addition, for resource-limited VSNs, the characteristic (resource scalability or power scalability) of scalable video coding [23] is also very critical [24]-[27]. Based on the power-rate-distortion (PRD) model derived for power-scalable video encoder, the resource allocation can be performed according to the available resources (*e.g.*, power supply and target bit rate) while optimizing the reconstructed video quality [25]-[27], which is briefly described in Sec. II-C.

B. Overview of Our Method

In this paper, a low-complexity hash-based multiview distributed video coding scheme with power-rate-distortion resource allocation is proposed. Our method possesses the characteristics of the aforementioned low-complexity video encoding approaches: (i) similar to DVC [9]-[20], motion estimation is shifted to the decoder; and (ii) similar to [21]-[22], collaborative video coding and transmission is employed for further compression. For an input video sequence, a frame is either a key frame or a non-key frame. Each key frame is encoded using the H.264/AVC intraframe encoder [3]. For each non-key frame, the temporal and interview predictive coding is efficiently achieved without performing motion and disparity estimations by extracting the significant differences between this non-key frame and its reference frames from the same VSN and the adjacent VSN at the encoder for media hash [29]-[30] comparison. Only the extracted significant components will be entropy-encoded. To exploit interview correlation, unlike the current multiview DVC [15]-[19], limited inter-VSN communication during the encoding process is allowed to exchange hash information of relatively small size. The global disparity between the frames from adjacent VSNs is estimated via global motion estimation performed at the decoder, and the estimated motion parameters are fed back to the encoder via a feedback channel. The availability of a feedback channel is a common assumption of most DVC approaches [9], [11]-[14], [16]-[18]. In addition, similar to the concept of the collaborative video coding and transmission approach, the intra-encoded key frames from adjacent VSNs can be further re-encoded via hash information comparison while they are transmitted through the same intermediate VSN.

In addition, the unique characteristic of our method is that we present a PRD model to characterize the relationship between the available resources (*e.g.*, power supply and target bit rate) and the RD performance of the proposed video codec. Based on this PRD model, the proposed video encoding procedure can be roughly viewed as the combination of the intra-mode block encoding, the inter-mode block encoding, and the entropy encoding operations. More specifically, an RD function in terms of the percentages for different coding modes of blocks and the target bit rate under the available resource constraints is derived for optimal block coding mode decision. With this model, the resource allocation can be efficiently performed at the encoder while optimizing the reconstructed video quality.

The remainder of this paper is organized as follows. Literature review about low-complexity video coding is given in Sec. II. Our robust media hashing technique [29]-[30] and hash-based video coding technique are described in Sec. III. The proposed low-complexity resource-scalable multiview distributed video coding scheme is described in Sec. IV. The PRD optimized resource allocation and block coding mode decision for the proposed video coding scheme is addressed in Sec. V. Simulation results are presented in Sec. VI. Finally, conclusions and future works are given in Sec. VII.

II. Related Works

In this section, distributed video coding, and collaborative video coding and transmission will first be described in Secs. II-A and II-B, respectively. Then, resource-scalable video coding and power-rate-distortion optimized resource allocation are described in Sec. II-C.

A. Distributed Video Coding

The major characteristic of the distributed video coding (DVC) approach is that individual frames are conceptually encoded independently, but decoded jointly [9]-[20]. The major computational burden (*e.g.*, complex motion or disparity estimation) at the encoder can be shifted to the decoder while preserving a certain coding efficiency. That is, the objective of DVC is to achieve the coding efficiency as high as that of interframe encoding (*e.g.*, H.264/AVC interframe encoding), with encoder complexity as low as that of intraframe encoding (*e.g.*, JPEG-2000 or H.264/AVC intraframe encoding). Clearly, such techniques are very promising for applications in WVSN where a low-complexity video encoder can be embedded in each VSN, and the complex decoding tasks can be performed at the RCU, which supports a powerful decoder [1], [2], [9]-[20].

More specifically, most existing DVC approaches [9]-[20] modeled lossy video coding as a channel coding problem based on Wyner-Ziv information theory [28]. The statistical dependency between two correlated sources *W* and *Y* is modeled as a virtual correlation channel, where the side information *Y* is viewed as a noisy version of the source *W*. At the encoder, the compression of *W* can be achieved by transmitting, via the feedback channel, only part of the parity bits derived from the channel-encoded version of *W* according to the request. Here, the parity bits form the so-called Wyner-Ziv bits. The decoder uses the received Wyner-Ziv bits and the side information *Y* derived from previous decoded video signals at the decoder to perform channel decoding to correct some "errors" in *Y*, *i.e.*, the noisy version of the source *W*, for the reconstruction of *W*.

In DVC, the Wyner-Ziv bits are generated by first transforming (*e.g.*, discrete cosine transform) each input frame to the transformed domain, followed by performing scalar quantization and channel encoding. The computational complexity for the DVC encoder is comparable to that of traditional intraframe encoder consisting of transformation, quantization, and entropy encoding, and hence, is suitable for a VSN. In addition, the side information for each frame to be decoded can be generated at the decoder by interpolating or extrapolating previous decoded frames from the same VSN (single-view DVC [9]-[14]) or those from the same VSN and adjacent VSNs (multiview DVC [11]-[12], [15]-[20]). The side information generation process at the decoder is usually computationally expensive and is similar to the motion or disparity estimation in traditional video encoder. However, the decoder supported by the RCU is usually powerful enough to perform these complex operations. On the other hand, the major characteristic of multiview DVC [15]-[20] is that each VSN can encode its captured video individually, and the compressed bitstreams received at the decoder from multiple VSNs can be jointly decoded. Hence, inter-VSN communication can be avoided during the encoding process to save the power consumed in data communication, and the interview correlation can be exploited at the decoder.

B. Collaborative Video Coding and Transmission

Another popular low-complexity video coding paradigm for WVSN based on collaborating video coding and transmission has been recently presented [21]-[22]. First, each frame captured by a VSN is encoded using a traditional intraframe video encoder. While transmitting the encoded frames from adjacent VSNs toward the AFN through the same intermediate node, this intermediate node will first intra-decode these frames, and perform an image matching procedure to detect the similar/overlapping regions for these frames. Then, the similar/overlapping regions will be encoded once only to further reduce the bit rate. Since these frames from adjacent VSNs may be captured from different viewpoints or different time instants, it usually needs to perform some image registration techniques [1], [21]-[22] to detect the similar/overlapping regions. Image registration usually performs feature detection, feature matching, transform model estimation, and image re-sampling and transformation for the frames from different VSNs to identify their similar/overlapping regions. Finally, those overlapped regions can be encoded only once using an intraframe encoder, while the other regions can be intra-encoded. However, image registration is usually a complex task, which is prohibitive for a resource-limited VSN. Hence, the major challenge is how to efficiently and accurately identify the similar/overlapping regions among the frames from adjacent VSNs in an intermediate VSN under resource-limited constraints.

C. Resource-scalable Video Coding and Power-rate-distortion Optimized Resource Allocation

For a resource-limited VSN, designing a power-scalable video encoder for a VSN, which can adjust its encoding parameters based on current available resources (*e.g.*, power supply and target bit rate) is necessary. In [25]-[27], a parametric complexity-scalable video encoder is developed by adjusting the three encoding parameters; namely, (i) the number of the SAD (sum of absolute difference) computations for motion estimation per frame; (ii) the number of nonzero MBs (macroblocks), *i.e.*, MBs with nonzero DCT coefficients, per frame; and (iii) the encoding frame rate based on the encoding complexity constraint. In addition, using a popular CMOS circuit design technology, called dynamic voltage scaling (DVS), the power scalability is equivalent to the complexity scalability. Therefore, an analytic power-rate-distortion (PRD) model is developed to characterize the relationship between the power consumption of the encoder and its RD performance. Based on the PRD model, the resource allocation can be performed by optimally adjusting the three parameters according to the available resources at the encoder while optimizing the reconstructed video quality.

III. ROBUST MEDIA HASHING AND HASH-BASED VIDEO CODING

A. Robust Media Hashing

In the proposed low-complexity single-view video coding scheme, unlike traditional interframe coding and current DVC, motion estimation cannot be performed at both the encoder and the decoder. On the contrary, temporal correlation is exploited by efficiently comparing the block-based robust media hashing information between two successive frames. In addition, in the proposed multiview video encoder, limited block-based media hash information is allowed to be exchanged among adjacent VSNs to achieve further coding efficiency. Our robust media hashing scheme, called structural digital signature (SDS) [29]-[30], which can extract the most significant components and provide a compact representation for an image (or video frame) or an image block efficiently, meets the aforementioned requirements.

To extract the SDS for a frame, the frame is first decomposed into several non-overlapped blocks. In order to make the SDS extracted from a block be representative, the block size should be large enough. To extract the SDS for an image block of size $n \times n$, a J-scale DWT (discrete wavelet transform) is performed. Let $w_s(x, y)$ represent a wavelet coefficient at scale s and position (x, y), $0 \leq s < J$, $1 \leq x \leq n$, and $1 \leq y \leq n$. For each pair consisting of a parent node, $w_{s+1}(x, y)$, and its four child nodes, $w_s(2x + i, 2y + j)$, $0 \leq i, j \leq 1$, the maximum magnitude difference (*max_mag_diff*) value is calculated as

$$max\_mag\_diff_{s+1}(x,y) = \max_{0 \leq i,j \leq 1} \left\| w_{s+1}(x,y) \right| - \left| w_s(2x+i, 2y+j) \right\|. \tag{1}$$

Then, all the parent-4 children pairs are arranged in decreasing order based on their *max_mag_diff* values. The first L (L is denoted as the hash length) pairs in the decreasing order are selected to be significant and are selected for constructing the SDS of the block.

Once the significant parent-4 children pairs are selected, each pair is assigned a symbol representing what kind of relationship this pair carries. According to the interscale relationship existing among wavelet coefficients, there are four possible relationship types. Assume the magnitude of a parent node p is larger than that of its child node c with *max_mag_diff* value. When $|p| \geq |c|$, the four possible relationships of the pair are (a) $p \geq 0$, $c \geq 0$; (b) $p \geq 0$, $c < 0$; (c) $p < 0$, $c \geq 0$; and (d) $p < 0$, $c < 0$. To make the above-mentioned relationships compact, relations (a) and (b) can be merged to form a signature symbol "+1" when $p \geq 0$ and c is ignored. On the other hand, relations (c) and (d) can be merged into another signature symbol "-1" when $p < 0$ and c is ignored. That is, the sign of the larger node is kept unchanged while ignoring the smaller one under the constraint that their original interscale relationship is still preserved. Similarly, the signature symbols "+2" and "-2" can be defined under the constraint $|p| < |c|$. In summary, for each selected pair of a parent node p and its child node c with *max_mag_diff* value in an image block, B, the signature symbol $Sym(B, p, c)$ can be defined as:

$$Sym(B,p,c) = \begin{cases} +1 & if \quad (|p| \geq |c|) \quad and \quad (p \geq 0), \\ -1 & if \quad (|p| \geq |c|) \quad and \quad (p < 0), \\ +2 & if \quad (|p| < |c|) \quad and \quad (c \geq 0), \\ -2 & if \quad (|p| < |c|) \quad and \quad (c < 0). \end{cases} \tag{2}$$

That is, an image block can be translated into a symbol sequence. Those pairs not included in the SDS (outside the first L pairs in the decreasing order) are labeled by "0." That is, for an $n \times n$ image block, there are at most $(n/2) \times (n/2)$ parent-4 children pairs, and hence, the SDS for an $n \times n$ image block can be a symbol sequence in raster scan order, consisting of L significant symbols (each belongs to +1, -1, +2, or -2) and $[(n/2) \times (n/2) - L]$ "0" symbols, which can be efficiently compressed via run-length coding and entropy coding techniques. Due to the fact that the position of a parent node can indicate the positions of its child nodes, for simplicity, by considering the horizontal coordinate p_x and the vertical coordinate p_y for a parent node p in an $n \times n$ block B, the SDS of B can be expressed by

$$SDS(B) = \left\{ S\left(B, p_x, p_y\right) | \, S(B, p_x, p_y) = 0, \pm 1, \pm 2, \;\; 0 \le p_x < n/2, \;\; 0 \le p_y < n/2 \right\}, \tag{3}$$

where $S(B, p_x, p_y) = \pm 1$ or ± 2 means that the SDS symbol $S(B, p_x, p_y)$ is in the selected L symbols with maximum *max_mag_diff* values while $S(B, p_x, p_y) = 0$ means that the SDS symbol $S(B, p_x, p_y)$ is outside the selected L symbols. Usually, the hash length L is selected to be relatively small, *i.e.*, $L << (n/2) \times (n/2) - L$, *i.e.*, $L << n^2/8$.

B. Hash-based Video Coding

In this paper, the major purpose is to efficiently extract the most significant components of an image block for compressing the block without performing motion estimation. Based on the characterization of signal reconstruction in [31], image signals can be approximately reconstructed from their multiscale information derived from the DWT domain. In this paper, the multiscale information of an image block is derived from its SDS. To compress a block, its most significant components can be extracted by comparing its SDS and that of its reference block (the co-located block in its reference frame). For each symbol $S(B, p_x, p_y) \neq 0$ (in the selected L SDS symbols) of the block B, if $S(B, p_x, p_y) \neq S(B', p_x, p_y)$, then $S(B, p_x, p_y)$ is determined to be significant, where B' is the reference block of B, and $S(B, p_x, p_y)$ and $S(B', p_x, p_y)$ have the same parent node position (p_x, p_y); otherwise, $S(B, p_x, p_y)$ is determined to be insignificant. For each significant SDS symbol, its corresponding 5 wavelet coefficients (denoted by significant coefficients) will be quantized and encoded. For each insignificant SDS symbol, its corresponding 5 wavelet coefficients (denoted by insignificant coefficients) are replaced by zeros. Then, for the block B, all the coefficients (significant and insignificant coefficients) arranged in the raster scan order can be efficiently compressed via the run-length coding and entropy coding techniques to form the bitstream. To reconstruct the block B, based on the reconstructed reference block B', the decoded coefficients for B are used to modify B' to obtain $ß$, which will have SDS similar to that of B. Hence, $ß$ can be regarded as the reconstructed version of B. More specifically, the problem of exploiting the SDS for image block encoding and reconstruction can be formulated as follows. For an image block B to be encoded, its most significant components, extracted by comparing the SDS of B and that of its reference block B', should be properly selected such that

$$PSNR(B, ß) \geq \text{desired PSNR value, and} \tag{4}$$

$$PSNR(B, ß) >> PSNR(B', ß), \tag{5}$$

where *PSNR* denotes the PSNR (peak signal to noise ratio), and $ß$ is an estimate of B obtained by modifying B' using the SDS of B such that B and $ß$ have the same SDS. The detailed encoding/decoding processes of the proposed video coding scheme are described in Sec. IV.

IV. Proposed Low-complexity Resource-scalable Multiview Distributed Video Coding Scheme

Assume that there are N_{VSN} adjacent VSNs observing the same target scene in a WVSN. For each VSN, V_s, $s = 0, 1, 2, \ldots, N_{VSN} - 1$, a captured video sequence is divided into several GOPs (group of pictures) with GOP size, $GOPS_s$, in which a GOP consists of a key frame, $K_{s,t}$, where t mod $GOPS_s = 0$, and some non-key frames, $W_{s,t}$, where t mod $GOPS_s \neq 0$. An example of the GOP structure with $N_{VSN} = 3$ is shown in TABLE I. In the proposed video encoding scheme, each key frame is encoded using the H.264/AVC intraframe encoder [3]. In the following, the proposed low-complexity single-view video coding scheme for non-key frames will be described in Sec. IV-A. Then, based on the concepts of the proposed single-view video coding, DVC, and the collaborative video coding and transmission approach [21]-[22], the proposed low-complexity multiview video coding scheme will be described in Secs. IV-B and IV-C for non-key frames and key frames, respectively.

TABLE I
A Simple Example of the GOP Structure for a WVSN with $N_{VSN} = 3$, where $GOPS_0 = 1$, $GOPS_1 = 4$, and $GOPS_2 = 2$.

VSN / Time instant	t	$t+1$	$t+2$	$t+3$	$t+4$	•••
V_0	$K_{0,t}$	$K_{0,t+1}$	$K_{0,t+2}$	$K_{0,t+3}$	$K_{0,t+4}$	•••
V_1	$K_{1,t}$	$W_{1,t+1}$	$W_{1,t+2}$	$W_{1,t+3}$	$K_{1,t+4}$	•••
V_2	$K_{2,t}$	$W_{2,t+1}$	$K_{2,t+2}$	$W_{2,t+3}$	$K_{2,t+4}$	•••

A. Proposed Low-complexity Single-view Video Coding Scheme for Non-key Frames

A.1. Hash-assisted video coding

The block diagram of the proposed low-complexity single-view video coding scheme is shown in Fig. 2. At the encoder, for each non-key frame $W_{s,t}$ captured by VSN V_s at time instant t, its nearest key frame $R_{s,t}$ is determined to be its reference frame. For example, if the previous frame of $W_{s,t}$ is a key frame, the reference frame of $W_{s,t}$ is $R_{s,t} = K_{s,t-1}$. Each non-key frame is decomposed into several non-overlapping blocks of size $n \times n$. Let $B_{s,t,b}$ denote a block in $W_{s,t}$, where b is the block index. The coding mode of $B_{s,t,b}$ will be determined by comparing $B_{s,t,b}$ and the co-located block $B'_{s,t,b}$ (called reference block) in $R_{s,t}$ to be one of the three possible coding modes: intra, inter, or skip. The coding mode decision based on current available resources will be later described in Sec. V. For each block with skip mode, only the coding mode information is encoded. Each block with intra mode is encoded using the H.264/AVC intraframe encoder [3]. Each block with inter mode will be encoded using block hash, described as follows.

Without allowing motion estimation at the encoder, the temporal correlation of successive frames is exploited by performing efficient hash extraction and comparison. For a pair of a block $B_{s,t,b}$ with inter mode and its reference block $B'_{s,t,b}$, their respective hashes, $S(B_{s,t,b}, p_x, p_y)$ and $S(B'_{s,t,b}, p_x, p_y)$, will be extracted and compared based on Eqs. (2)-(3), where (p_x, p_y) denotes a parent node position. For each non-zero SDS symbol of $B_{s,t,b}$ and its co-located symbol of $B'_{s,t,b}$, if $S(B_{s,t,b}, p_x, p_y) \neq S(B'_{s,t,b}, p_x, p_y)$, then the corresponding 5 wavelet coefficients for $S(B_{s,t,b}, p_x, p_y)$ are determined to be significant; otherwise, they are insignificant and skipped without compression. Finally, as mentioned in Sec. III, all the significant coefficients are quantized and entropy-encoded to form the bitstream for the block $B_{s,t,b}$ with inter mode.

At the decoder, each key frame is decoded using the H.264/AVC intraframe decoder. For a non-key frame, each block with skip mode is decoded by copying from the reconstructed reference block. Each block with intra mode is decoded using the H.264/AVC intraframe decoder. For each block $B_{s,t,b}$ with inter mode, all the significant coefficients are decoded and used to modify the reconstructed reference block $B'_{s,t,b}$ to obtain $ß_{s,t,b}$ (reconstructed $B_{s,t,b}$) by filling the decoded coefficients into the corresponding positions in the DWT version of the reconstructed $B'_{s,t,b}$, followed by the inverse DWT. That is, block $B_{s,t,b}$ is reconstructed from its multiscale information derived from the DWT domain, and incorporated with the modification of its reference block $B'_{s,t,b}$ to obtain its reconstructed version $ß_{s,t,b}$, such that $PSNR(B_{s,t,b}, ß_{s,t,b}) >> PSNR(B'_{s,t,b}, ß_{s,t,b})$.

A.2. Computational complexity

The computational complexity for encoding a block with inter mode includes those of the SDS extraction, SDS comparison, quantization, and entropy-encoding. The computational complexity of the SDS extraction and comparison is dominated by that of the DWT. Hence, without performing motion estimation, the computational complexity for encoding a block with inter mode should be very similar to that for encoding a block using a traditional intraframe encoder, consisting of transformation (DCT or DWT), quantization, and entropy-encoding.

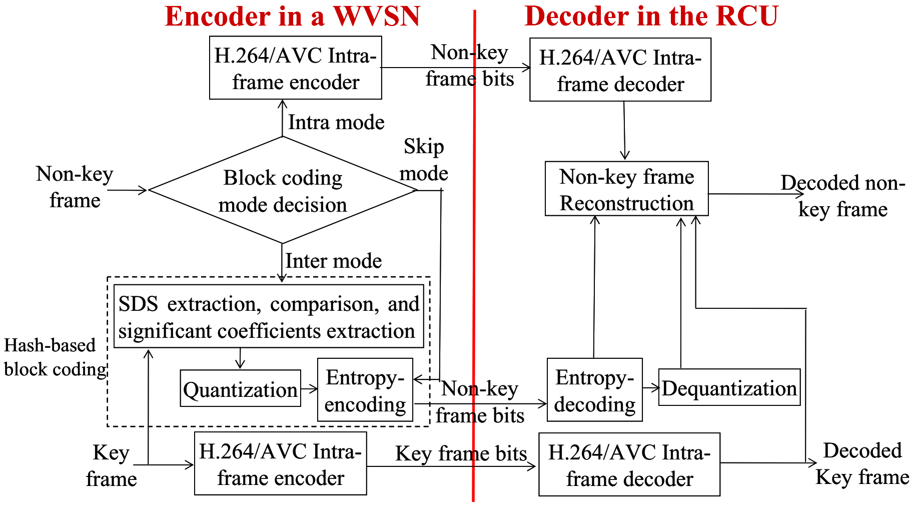


Fig. 2. Block diagrams of the proposed low-complexity single-view video coding scheme.

B. Proposed Low-complexity Multiview Video Coding Scheme for Non-key Frames

B.1. Multiple-frame referencing

To achieve better coding efficiency by extending the proposed single-view video coding scheme to multiview video coding, for each non-key frame from a VSN, the multi-reference frames from the same VSN and the adjacent VSNs are jointly exploited. However, as mentioned before, the frames from adjacent VSNs may be captured from different viewpoints. Hence, before exploiting the interview correlation among the frames from adjacent VSNs, these frames should be transformed to the same viewpoint. The global disparities among these frames from adjacent VSNs can be represented by global motion models [8], [15]-[19]. Here, a well-known global motion model, called affine transformation, is exploited, which has been successfully employed in traditional multiview video coding at the encoder [8] and multiview DVC at the decoder [15]-[19] to exploit interview correlation. Consider a frame *W* captured by a VSN V_s at time instant *t*, and one of its reference frames, *K*, from a VSN adjacent to V_s at the same time instant *t*. In the affine transformation model [8], [15]-[19], each pixel location (x_i, y_i) in *K* can be mapped to the pixel location (x'_i, y'_i) in *W* at the time instant *t* via the transformation

$$\begin{bmatrix} x_i' \\ y_i' \end{bmatrix} = A_t \begin{bmatrix} x_i \\ y_i \end{bmatrix} + C_t, \tag{6}$$

where

$$A_t = \begin{bmatrix} a_{1,t} & a_{2,t} \\ b_{1,t} & b_{2,t} \end{bmatrix}, \quad C_t = \begin{bmatrix} c_{1,t} \\ c_{2,t} \end{bmatrix}, \tag{7}$$

and $a_{1,t}$, $a_{2,t}$, $b_{1,t}$, $b_{2,t}$, $c_{1,t}$, and $c_{2,t}$ are the six transform parameters which can be estimated by minimizing the sum of squared errors over all *N* corresponding pairs of pixels inside the frames *W* and *K* as follows [8]:

$$E = \sum_{i=1}^{N} \left[W(x_i', y_i') - K(x_i, y_i) \right]^2, \tag{8}$$

where $W(x'_i, y'_i)$ and $K(x_i, y_i)$ are the pixel values in the frame *W* and *K*, respectively.

As a simple example, consider the two adjacent VSNs, V_0 and V_1, shown in TABLE I. For V_1 to encode a non-key frame $W_{1,t}$ captured at time instant *t*, its nearest key frame $R_{1,t}$ (*e.g.*, if the immediately previous frame $K_{1,t-1}$ of $W_{1,t}$ is a key frame, $R_{1,t} = K_{1,t-1}$) from the same VSN and the key frame $K_{0,t}$ captured by V_0 can be jointly considered to be its reference frames. Note that, to encode a non-key frame, only a key frame in the same VSN or the key frame(s) captured by adjacent VSN(s) at the same time instant can be considered as the reference frames. First, the reference frame $K_{0,t}$ from V_0 is warped to the same viewpoint of V_1 to get $K'_{0,t}$ based on the affine transform parameters between V_0 and V_1. Then, for encoding $W_{1,t}$, $R_{1,t}$ from the same VSN V_1 and $K'_{0,t}$ from the adjacent VSN, V_0 can serve as its first and second reference frames, respectively.

B.2. Hash-assisted video coding

As described in Sec. IV-B.1, the global motion estimation process (minimization of Eq. (8)) is required for multiple-frame referencing, but is too complex to be performed in a VSN. Hence, similar to the DVC approach, this complex task should be shifted to the decoder at RCU, shown in Fig. 3. Due to the fact that the key frame can be independently intra-encoded and intra-decoded and the fact that RCU can usually support powerful computational capability, the global motion estimation between each pair of intra-decoded key frames captured at the same time instant from adjacent VSNs is performed at the decoder. Then, the estimated motion parameters are transmitted back to their corresponding pair of VSNs via a feedback channel for warping and encoding subsequent frames in the current GOP. Note that it is assumed that after deploying a WVSN, changing the location and viewpoint of each VSN is not allowed. Hence, the estimated global motion parameters between a pair of key frames should be approximately similar to those for the subsequent frame pairs in the same GOP with smaller GOP size.

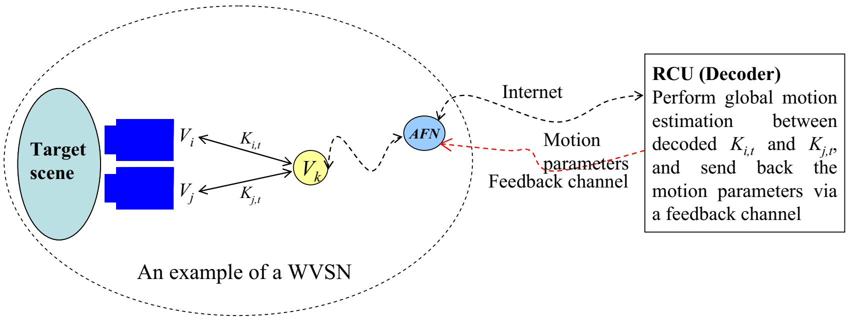

Target scene
V_i
$K_{i,t}$
V_j
$K_{j,t}$
V_k
AFN
Internet
Motion parameters
Feedback channel
RCU (Decoder)
Perform global motion estimation between decoded $K_{i,t}$ and $K_{j,t}$, and send back the motion parameters via a feedback channel
An example of a WVSN

Fig. 3. An example for global motion estimation performed at the decoder.

The proposed multiview video coding scheme for non-key frame can be illustrated by the example shown in Fig. 4. For encoding a non-key frame $W_{j,t}$ captured from VSN V_j (j = 1 and t = 45 in Fig. 4), its nearest key frame, $R_{j,t}$ ($R_{j,t} = R_{1,45} = K_{1,44}$ in Fig. 4), captured from the same VSN V_j is determined to be its "first" reference frame. Similar to the proposed single-view video coding, the coding mode for each block $B_{j,t,b}$ in $W_{j,t}$ is determined by comparing $B_{j,t,b}$ and the co-located block $B'_{j,t,b}$, in $R_{j,t}$ (step (a) in Fig. 4). Here, $B'_{j,t,b}$ is the "first" reference block for $B_{j,t,b}$. The coding mode decision based on current available resources will be described in Sec. V. For each block $B_{j,t,b}$ with inter mode, without performing motion estimation, the respective hashes for $B_{j,t,b}$ and its reference block $B'_{j,t,b}$, $S(B_{j,t,b}, p_x, p_y)$ and $S(B'_{j,t,b}, p_x, p_y)$, are extracted and compared (step (b) in Fig. 4) to extract the "initial" significant SDS symbols (step (c) in Fig. 4), where (p_x, p_y) denotes a parent node position. For each pair of non-zero SDS symbols of $B_{j,t,b}$ and $B'_{j,t,b}$, if $S(B_{j,t,b}, p_x, p_y) \neq S(B'_{j,t,b}, p_x, p_y)$, then $S(B_{j,t,b}, p_x, p_y)$ is determined to be an "initial" significant symbol; otherwise, it is determined to be insignificant and can be predicted by the "first" reference block. To further reduce the number of initial significant symbols, the initial significant symbols will be compared with the co-located symbols in the "second" reference frame from an adjacent VSN as follows.

Without allowing uncompressed frame exchanges between VSNs during the encoding process, V_j will send a message containing each initial significant SDS symbol in $W_{j,t}$ to its adjacent VSN V_i to announce it needs the "second" reference frame. As mentioned in Sec. III, for each block with inter mode in a non-key frame, the significant SDS symbols and insignificant SDS symbols (replaced by "0") can be arranged to a raster scan order and efficiently compressed via run-length coding and entropy coding techniques. Hence, it will not consume too much transmission power in sending the message from V_j to V_i.

After receiving the message from V_j, V_i will warp $K_{i,t}$ to the viewpoint of $W_{j,t}$ (i = 0, j = 1, and t = 45 in Fig. 4) using the global motion parameters estimated by their previous key frame pair to get $K'_{i,t}$ (the affine-transformed $K_{i,t}$). Then, each "initial" significant SDS symbol $S(B_{j,t,b}, p_x, p_y)$ of block $B_{j,t,b}$ in $W_{j,t}$ will be compared with the co-located SDS symbol $S(B''_{i,t,b}, p_x, p_y)$ of block $B''_{i,t,b}$ in $K'_{i,t}$ (step (d) in Fig. 4). Here, $K'_{i,t}$ is the "second" reference frame of $W_{j,t}$ while $B''_{i,t,b}$ is the second reference block of $B_{j,t,b}$. If $S(B_{j,t,b}, p_x, p_y) \neq S(B''_{i,t,b}, p_x, p_y)$, then $S(B_{j,t,b}, p_x, p_y)$ is determined to be "true" significant symbol (step (e) in Fig. 4); otherwise, it is determined to be insignificant and can be predicted by the "second" reference block. After that, V_i will send a message containing the parent node position for each "true" significant symbol for $W_{j,t}$ to V_j. The parent node position can be expressed by a bitmap and efficiently compressed via run-length coding. Hence, it will not consume too much transmission power in sending the message from V_i to V_j.

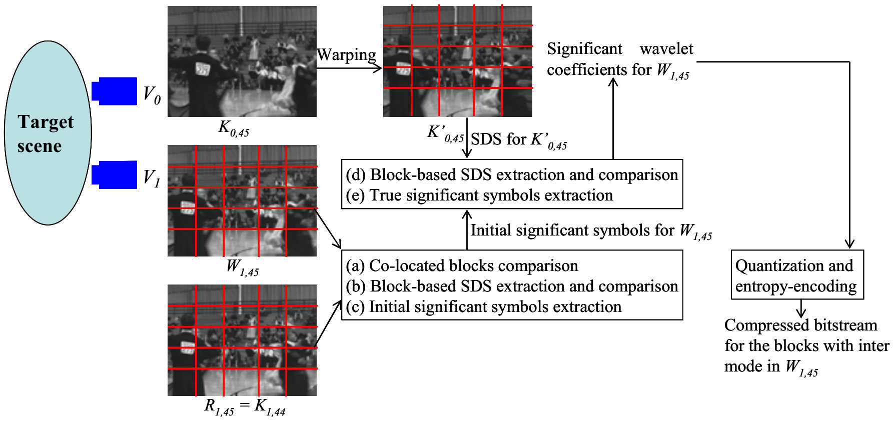

Target scene
V_0
V_1
$K_{0,45}$
Warping
$K'_{0,45}$
Significant wavelet coefficients for $W_{1,45}$
SDS for $K'_{0,45}$
(d) Block-based SDS extraction and comparison
(e) True significant symbols extraction
Initial significant symbols for $W_{1,45}$
$W_{1,45}$
(a) Co-located blocks comparison
(b) Block-based SDS extraction and comparison
(c) Initial significant symbols extraction
Quantization and entropy-encoding
Compressed bitstream for the blocks with inter mode in $W_{1,45}$
$R_{1,45} = K_{1,44}$

Fig. 4. An example for the proposed low-complexity multiview video coding scheme for non-key frames.

After V_j receives the message sent from V_i, for each block with inter mode in $W_{j,t}$, the wavelet coefficients corresponding to each "true" significant SDS symbol are determined to be significant and quantized. The wavelet coefficients corresponding to each insignificant symbol, which can be predicted by the "first" reference block, are replaced with "0" while those corresponding to each insignificant symbol, which can be predicted by the "second" reference block, are replaced with "1." Finally, all the coefficients for the block can be efficiently compressed via run-length coding and entropy coding to form the bitstream of this block. Usually, most SDS symbols corresponding to the background region can be well-predicted by their first reference block while some of the symbols corresponding to the foreground (moving objects) can be well-predicted by their second reference block. To encode a non-key frame, a two-way data exchange between adjacent VSNs is required. However, for small-motion sequences, if the first reference frame can predict a non-key frame well, only the single-view video coding scheme will be performed. The original non-key frame $W_{1,45}$ to be encoded in Fig. 4 and the spatial display of its true significant wavelet coefficients (no blocks with intra mode in this example) are shown in Fig. 5. It can be observed from Fig. 5 that the proposed scheme can indeed extract the most significant components (*e.g.*, important edges or object shapes) for a non-key frame.

At the decoder, for each block with inter mode, all the true significant coefficients are entropy-decoded and dequantized. All the insignificant coefficients are recovered by copying the corresponding coefficients from the decoded first and second reference blocks. Finally, the inverse DWT is performed to reconstruct this block.

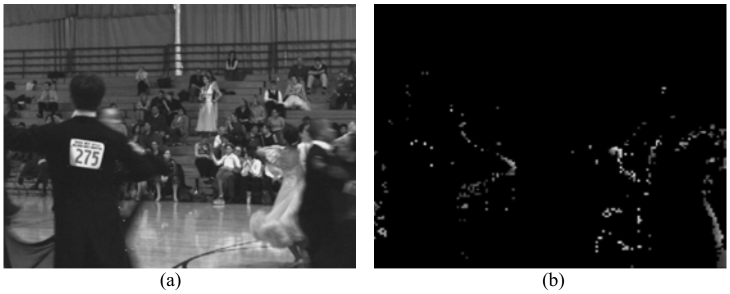

(a) (b)

Fig. 5. (a) The original non-key frame $W_{1,45}$ to be encoded in Fig. 4 and (b) the spatial display of its true significant wavelet coefficients.

Currently, all key frames are intra-encoded and intra-decoded using the H.264/AVC intraframe codec. Each key frame usually serves as the reference frame for non-key frame, and should have higher quality. However, the bit rate for a compressed key frame is relatively high. In fact, a key frame can be further compressed by referring the other key frame captured at the same time instant by adjacent VSNs while they are transmitted toward AFN through the same intermediate node, which will be addressed in Sec. IV-C.

B.3. Computational complexity

The complexity of the proposed multiview video encoder approximately consists of that of the proposed single-view video encoder and that for exchanging hash data between two VSNs. Due to the sizes of the exchanged hash data being relatively small, the complexity of the proposed multiview video encoder should be approximately in the order of that of traditional intraframe encoder.

C. Proposed Low-complexity Multiview Video Coding Scheme for Key Frames

Similar to the concept of the collaborative video coding and transmission approaches [21]-[22], the key frames of adjacent VSNs can be further compressed while they are transmitted through the same intermediate node. For a pair of key frames captured at the same time instant from adjacent VSNs, the global disparity between them can be modeled via a global motion model (Eq. (6)). As an example, in Fig. 3, for a pair of adjacent VSNs, V_i and V_j, their first key frame pair will be intra-encoded individually and transmitted to the decoder for estimation of global motion parameters, which will be transmitted back to the corresponding VSNs via a feedback channel. The non-first key frame pair captured by V_i and V_j at the same time instant will be also intra-encoded individually first, and then transmitted toward AFN through the same intermediate node V_k. The node V_k will perform key frame re-encoding, as described in the following.

As an example, in Fig. 6, V_k will first perform intra-decoding to decode intra-encoded $K_{i,t}$ and $K_{j,t}$ to obtain $K'_{i,t}$ and $K'_{j,t}$, respectively ($i = 0, j = 1$, and $t = 48$ in Fig. 6). For re-encoding $K'_{j,t}$ by treating $K'_{i,t}$ as its reference frame, $K'_{i,t}$ will be warped to the viewpoint of V_j via the global motion parameters between them to get $\acute{K}_{i,t}$ (the affine-transformed $K'_{i,t}$). This is reasonable because the estimated global motion parameters from the previous key frame pair, as mentioned in Sec. IV-B.1, can preserve certain accuracy for smaller GOP size. Then, both $K'_{j,t}$ and $\acute{K}_{i,t}$ are partitioned into several non-overlapped blocks, respectively. Due to the fact that the key frame should have higher quality (usually also with higher bit rates) to serve as the reference frames for encoding non-key frames, more blocks should be with intra mode. Hence, without considering the coding mode decision procedure to be described in Sec. V, the coding mode for each block $B'_{j,t,b}$ in $K'_{j,t}$ is only roughly determined based on the PSNR between $B'_{j,t,b}$ and the co-located block $B'_{i,t,b}$, in $\acute{K}_{i,t}$. If $PSNR(B'_{j,t,b}, B'_{i,t,b}) < T_L$, $B'_{j,t,b}$ is determined to be with intra mode. If $PSNR(B'_{j,t,b}, B'_{i,t,b}) > T_S$, $B'_{j,t,b}$ is determined to be with skip mode. Otherwise, $B'_{j,t,b}$ is determined to be with inter mode. The two thresholds, T_L and T_S, are empirically adjusted based on the target bit rates, respectively. Then, the proposed single-view video coding scheme is performed to each kind of coding mode to form the bitstream of the re-encoded $K'_{j,t}$.

After decoding each key frame pair, the global motion parameters between them will be estimated and fed back to the encoder for warping and encoding subsequent frames. Obviously, the computational complexity of the proposed video encoding scheme for key frame re-encoding is very similar to that of the proposed single-view video encoding scheme, and should be approximately in the order of that of traditional intraframe encoder. The original key frame $K_{1,48}$ to be re-encoded in Fig. 6 and the spatial display of its significant wavelet coefficients (with some blocks with intra mode) are shown in Fig. 7. It can be observed from Fig. 7 that the proposed scheme can indeed extract the most significant components for blocks with inter mode and preserve large-motion blocks to be with intra mode.

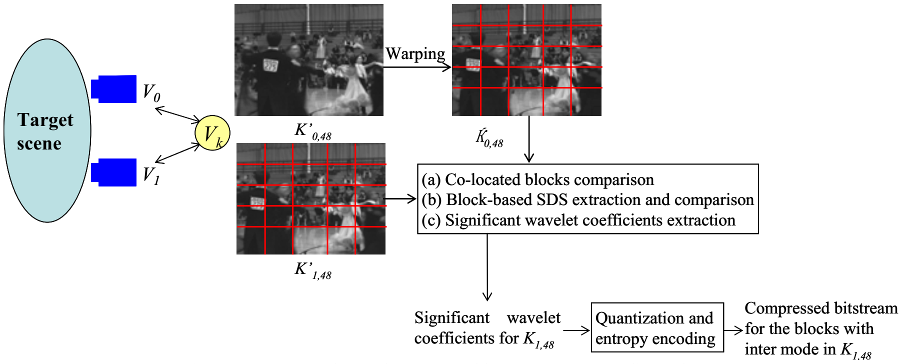

Target scene
V_0
V_1
V_k
Warping
$K'_{0,48}$
$\acute{K}_{0,48}$
$K'_{1,48}$
(a) Co-located blocks comparison
(b) Block-based SDS extraction and comparison
(c) Significant wavelet coefficients extraction
Significant wavelet coefficients for $K_{1,48}$
Quantization and entropy encoding
Compressed bitstream for the blocks with inter mode in $K_{1,48}$

Fig. 6. An example for the proposed key frame re-encoding scheme.

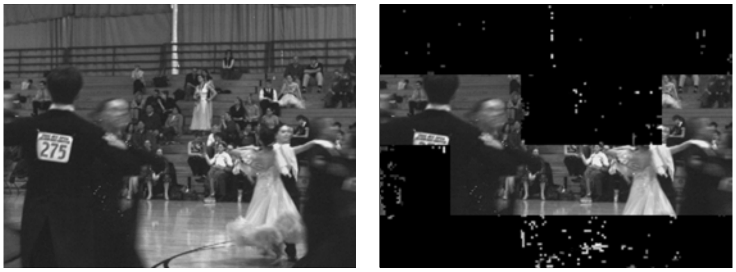

(a) (b)

Fig. 7. (a) The original key frame $K_{1,48}$ to be re-encoded in Fig. 6 and (b) the spatial display of its significant wavelet coefficients with some blocks with intra mode.

V. Power-rate-distortion Optimized Resource Allocation

As mentioned in Sec. IV-A.1, the block coding mode will be determined based on current available resources (*e.g.*, encoding power and target bit rate). Since block coding mode is related to the RD performance, it is, thus, important to characterize the relationship between the available resources and the RD performance. The major objective is to optimize the reconstructed video quality and maximize the lifetime for a VSN under current resource constraints.

Based on [25]-[27], to analyze and control the power consumption of a VSN, a CMOS circuit design technology for a mobile device, called dynamic voltage scaling (DVS), is assumed to design the VSNs employed in this paper. It is claimed that the power consumption of a video encoder can be controlled by adjusting its computational complexity. That is, for a video encoder, its computational complexity can be translated into its power consumption. Hence, based on DVS, the power scalability is equivalent to the complexity scalability. In this section, the strategy of power-rate-distortion (PRD) optimized resource allocation for the proposed video coding scheme will be described.

A. Block Coding Mode Determination

First, without performing motion estimation, for a non-key frame consisting of N_b blocks, the motion activity for each block is estimated by the SAD between itself and its reference block. A block with larger motion activity has a larger probability of being decided to be with intra mode whereas a block with smaller motion activity has a larger probability of being decided to be with skip mode. Then, all the blocks in a non-key frame are sorted in a decreasing order based on their motion activities. Assume that there are N_{Intra}, N_{Inter}, and N_{Skip} blocks determined to be coded with intra mode, inter mode, and skip mode, respectively, in a non-key frame, where $N_{Intra} + N_{Inter} + N_{Skip} = N_b$. Let $\{B_i, i = 1, 2, \ldots, N_{Intra}\}$ denote the set of blocks with intra mode, let $\{B_i, i = N_{Intra} + 1, N_{Intra} + 2, \ldots, N_{Intra} + N_{Inter}\}$ denote the set of blocks with inter mode, and let $\{B_i, i = N_{Intra} + N_{Inter} + 1, N_{Intra} + N_{Inter} + 2, \ldots, N_{Intra} + N_{Inter} + N_{Skip} (= N_b)\}$ denote the set of blocks with skip mode. Let X, Y, and Z, respectively, denote N_{Intra}/N_b, N_{Inter}/N_b, and N_{Skip}/N_b, where $X + Y + Z = 1$. The optimal determination of X, Y, and Z for a non-key frame according to the current resources is equivalent to the determination of the coding mode for each block, which can be achieved based on PRD optimized resource allocation described in the next subsections.

B. Power-Rate-Distortion (PRD) Model

Similar to the power-scalable video encoder design criteria [26], the proposed non-key frame video encoding procedure can be roughly viewed as the combination of several "atom operations," including the intra-mode block encoding (DCT and quantization), the inter-mode block encoding (DWT, hash extraction, hash exchange, hash comparison, and quantization), and the entropy encoding operations. The encoding operation for a block with skip mode is ignored due to only the coding mode information being encoded, which will be included in the entropy encoding operation. Let the normalized computational complexity for the intra encoding, inter encoding, and entropy encoding operations be C_1, C_2, and C_3, $0 < C_1, C_2, C_3 < 1$, respectively. For the available resources consisting of the encoding power P (watt = Joule per second) and target bit rate R (bits per pixel, *i.e.*, bpp), the computational complexity for non-key frame encoding per second can be expressed as:

$$F \times (C_1 \times X + C_2 \times Y + C_3 \times R) \leq \Phi(P), \tag{9}$$

where F is the normalized frame rate, $0 < F < 1$, and $\Phi(P)$, $0 < \Phi(P) < 1$, is the normalized power consumption for the encoding power P transformed by the power function $\Phi(\bullet)$ under the assumption that DVS energy consumption management technology is employed [25]-[27]. For example, when the battery power of a VSN for encoding is full, $\Phi(P) = 1$. When 20% of the power for encoding is consumed, $\Phi(P) = 0.8$. To optimally decide the coding mode for each block according to the current available resources (*i.e.*, P and R), a RD function for non-key frame encoding should be derived and minimized.

The classic RD function can be expressed as [26]:

$$D = \min_{R_i} \frac{1}{N_b} \sum_{i=1}^{N_b} \left(\sigma_i^2 \cdot 2^{-2\gamma R_i} \right), \quad \text{s.t.} \quad \frac{1}{N_b} \sum_{i=1}^{N_b} R_i = R, \tag{10}$$

where R_i is the bit rate of the ith block, σ_i^2 is the variance of the ith block, and γ is a model parameter related to encoding efficiency. Here, the variance means the mean of the squared pixel values in a block. That is, the variance σ_i^2 means the maximum possible distortion for the ith block, which can be calculated as the MSE (mean squared error) between the ith block and a zero block. Based on the Lagrangian multiplier technique, the minimum distortion obtained by the optimal bit allocation can be expressed as:

$$D = \left(\prod_{i=1}^{N_b} \sigma_i^2 \right)^{\frac{1}{N_b}} \cdot 2^{-2\gamma R}. \tag{11}$$

Based on Eq. (11), obviously, the RD function for a block with intra mode can be expressed as:

$$D_{Intra} = \left(\prod_{i=1}^{N_{Intra}} \sigma_{i,Intra}^2 \right)^{\frac{1}{N_{Intra}}} \cdot 2^{-2\gamma R}, \tag{12}$$

where $\sigma_{i,Intra}^2$ is the variance of the ith block with intra mode.

On the other hand, a block with inter mode includes some significant coefficients (corresponding to the significant SDS symbols) being entropy-encoded, and the other insignificant coefficients being skipped and predicted by the corresponding coefficients in its reference block. Hence, the RD function for a block with inter mode can be expressed as:

$$D_{Inter} = \left(\prod_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}} \sigma_{i,Inter}^2 \right)^{\frac{1}{N_{Inter}}} \cdot 2^{-2\gamma R} + \frac{1}{N_{Inter}} \sum_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}} \delta_{i,Inter}^2, \tag{13}$$

where $\sigma_{i,Inter}^2$ is the variance of the pixels corresponding to the significant coefficients (denoted by "significant pixels") in the ith block with inter mode. To calculate the variance of the significant pixels, the pixels corresponding to the insignificant coefficients (denoted by "insignificant pixels") in the current block and the corresponding pixels in its reference block are replaced by zeros. $\delta_{i,Inter}^2$ is the MSE between the insignificant pixels and the corresponding pixels in its reference block. To calculate the MSE of the insignificant pixels, the significant pixels in the current block and the corresponding pixels in its reference block are replaced by zeros. Note that in the block coding mode decision process, for a block to be encoded with inter mode, only the first reference block from the same VSN is considered. This is because prior to actual video encoding, it is unworthy to waste power to perform data exchanges between VSNs. In addition, for a block with skip mode, the RD function is simply the MSE (denoted by $\delta_{i,Skip}^2$) between the block and its reference block as:

$$D_{Skip} = \frac{1}{N_{Skip}} \sum_{i=N_{Intra}+N_{Inter}+1}^{N_b} \delta_{i,Skip}^2. \tag{14}$$

C. Power-Rate-Distortion (PRD) Optimization

Based on the above derivations shown in Eqs. (12)-(14), the overall RD function of a block in the proposed video coding scheme can be expressed as:

$$D_{Overall} = \frac{1}{N_b}(N_{Intra} \times D_{Intra} + N_{Inter} \times D_{Inter} + N_{Skip} \times D_{Skip})$$

$$= X \times D_{Intra} + Y \times D_{Inter} + Z \times D_{Skip}. \tag{15}$$

To minimize $D_{Overall}$, we need to formulate D_{Intra}, D_{Inter}, and D_{Skip}. First, based on Eqs. (12)-(15), the parameter γ can be estimated as follows. For a scene to be observed, several sets of estimated encoding parameters (X, Y, Z, N_{Intra}, N_{Inter}, and N_{Skip}) and the corresponding actual distortions, respectively, obtained from the PRD optimization processes and the actual video encoding/decoding processes are collected offline. Consider the parameters, X_t, Y_t, Z_t, N_{Intra_t}, N_{Inter_t}, and N_{Skip_t}, obtained from the PRD optimization process with a given initial parameter $\gamma = \gamma_{Init}$ and the actual distortion D_t, for a non-key frame W_t, the parameter γ can be updated as:

$$\gamma = \frac{-1}{2R}\log_2\left[\frac{D_t - \dfrac{Y_t}{N_{Inter\_t}}\displaystyle\sum_{i=N_{Intra\_t}+1}^{N_{Intra\_t}+N_{Inter\_t}}\delta_{i,Inter}^2 - \dfrac{Z_t}{N_{Skip\_t}}\displaystyle\sum_{i=N_{Intra\_t}+N_{Inter\_t}+1}^{N_b}\delta_{i,Skip}^2}{X_t\left(\displaystyle\prod_{i=1}^{N_{Intra\_t}}\sigma_{i,Intra}^2\right)^{\frac{1}{N_{Intra\_t}}} + Y_t\left(\displaystyle\prod_{i=N_{Intra\_t}+1}^{N_{Intra\_t}+N_{Inter\_t}}\sigma_{i,Inter}^2\right)^{\frac{1}{N_{Inter\_t}}}}\right]. \tag{16}$$

Then, the updated γ can be used to perform the PRD optimization for the next frame, and γ can be similarly updated iteratively. Several offline estimated parameters γ can be averaged to be the parameter γ for a certain scene in a period. To minimize $D_{Overall}$ based on optimally selected X, Y, and Z, where $Z = 1 - X - Y$, under the constraint shown in Eq. (9), the function $D_{Overall}$ should be translated into a function of X and Y, which can be achieved by means of an approximation strategy, as described in the following.

Second, the function D_{Intra} defined in Eq. (12) can be converted to a continuous-time function. Usually, only a small number of blocks in a non-key frame are with intra mode, *i.e.*, N_{Intra} or X should be small. It can be observed from the curve "Actual" in Fig. 8 that, in a non-key frame, the first few blocks in the decreasingly sorted list of motion activity blocks usually have larger variances, and these variances will decrease as the motion activities decrease. Hence, it is reasonable to model $\sigma_{i,Intra}^2$ as a decreasing linear function:

$$G(t) = A \cdot (1 - t), A > 0, 0 \le t \le 1, t = i / N_b, 1 \le i \le N_{Intra}. \tag{17}$$

It can be observed from Fig. 8 that when the block index $i < 5$ ($X < 25\%$) in the total 20 blocks, the function $G(t)$ (the "Estimated" curve of Fig. 8) is accurate enough to model $\sigma_{i,Intra}^2$. Due to X being usually much smaller than 25%, using $G(t)$ to model $\sigma_{i,Intra}^2$ is promising. The parameter A in Eq. (17) can be derived from the previous PRD optimization result. Assume N_{Intra_pre} denotes the number of blocks with intra mode in a non-key frame obtained from the previous coding mode decision. Hence, in the current non-key frame, A can be estimated from

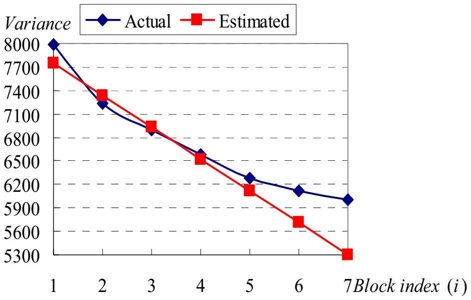


Fig. 8. The curve "Actual" shows the variances of the first few blocks in the decreasingly sorted list of motion activity blocks in a non-key frame. In this figure, the variances of the first 7 blocks in the decreasing block list consisting of the total 20 blocks in each non-key frame in the *Ballroom* and *Exit* sequences are evaluated. All the variances for the same block index are averaged. The curve "Estimated" shows the linear function $G(t) = A\cdot(1 - t)$, $t = i / N_b$, used to model these actual variances.

$$\frac{1}{N_b}\sum_{i=1}^{N_{Intra\_pre}}\sigma^2_{i,Intra} = \sum_{t=\frac{1}{N_b}}^{\frac{N_{Intra\_pre}}{N_b}} G(t) = \int_0^{\frac{N_{Intra\_pre}}{N_b}} A(1-t)dt$$

as:

$$A = \frac{2N_b\sum_{i=1}^{N_{Intra\_pre}}\sigma^2_{i,Intra}}{2N_bN_{Intra\_pre} - N^2_{Intra\_pre}}. \tag{18}$$

To get the continuous-time version of D_{Intra} in Eq. (12), we let $S = \left(\prod_{i=1}^{N_{Intra}}\sigma^2_{i,Intra}\right)^{\frac{1}{N_{Intra}}}$ and obtain

$$\ln S = \frac{1}{N_{Intra}}\sum_{i=1}^{N_{Intra}}\left(\ln\sigma^2_{i,Intra}\right). \tag{19}$$

The continuous-time version of ln*S* can be written as:

$$\ln S = \frac{N_b}{N_{Intra}}\sum_{t=\frac{1}{N_b}}^{\frac{N_{Intra}}{N_b}}\ln G(t) = \frac{1}{X}\int_0^X \ln[A(1-t)]dt\,. \tag{20}$$

By applying the Taylor expansion to Eq. (20), *S* can be derived as:

$$S = A\cdot e^{-1-\frac{1}{X}(1-X)\ln(1-X)} \approx A\times(1-0.5\times X),\ 0\le X\le 1. \tag{21}$$

The accuracy of the approximation for *S* is shown in Fig. 9. Hence, based on Eqs. (12), (19), and (21), D_{Intra} can be derived as:

$$D_{Intra}(X,R_{Intra}) = A(1-0.5X)\cdot 2^{-2\gamma R}\,. \tag{22}$$

Third, D_{Inter} in Eq. (13) can be expressed as a more complex form denoted by $D_{Inter}(X, Y, R_{Inter})$ as follows. Usually, the variance of the significant pixels for a block with inter mode will decrease as the motion activity decreases. Based on Fig. 10, it is reasonable to model $\sigma^2_{i,Inter}$ as a decreasing exponential function as:

$$H(t) = B_1e^{-B_2t},\ B_1 > 0,\ B_2 > 0,\ 0 \le t \le 1,\ t = i / N_b,\ N_{Intra} + 1 \le i \le N_{Intra} + N_{Inter}. \tag{23}$$

The parameter B_1 in Eq. (23) can be derived from the previous PRD optimization result. Assume N_{Intra_pre} and N_{Inter_pre} denote the numbers of blocks with intra mode and inter mode, respectively, in a non-key frame, obtained from the previous coding mode decision. Hence, in the current non-key frame, B_1 can be estimated from

$$\frac{1}{N_b}\sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}} \sigma_{i,Inter}^2 = \sum_{t=\frac{N_{Intra\_pre}+1}{N_b}}^{\frac{N_{Intra\_pre}+N_{Inter\_pre}}{N_b}} H(t) = \int_{\frac{N_{Intra\_pre}+1}{N_b}}^{\frac{N_{Intra\_pre}+N_{Inter\_pre}}{N_b}} B_1 e^{-B_2 t}\,dt$$

as:

$$B_1 = \frac{\dfrac{B_2}{N_b}\displaystyle\sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}} \sigma_{i,Inter}^2}{e^{-B_2\left(\frac{N_{Intra\_pre}+1}{N_b}\right)} - e^{-B_2\left(\frac{N_{Intra\_pre}+N_{Inter\_pre}}{N_b}\right)}}, \tag{24}$$

where B_2 controls the degradation speed of the exponential function *H(t)*, which can be obtained by some pre-training for each sequence. Usually, B_2 is a constant for the same scene.

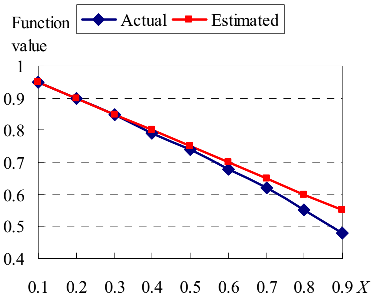


Fig. 9. The curve "Actual" shows the function $e^{-1-\frac{1}{X}(1-X)\ln(1-X)}$ in Eq. (21), and the curve "Estimated" shows its linear approximation function (1 – 0.5*X*).

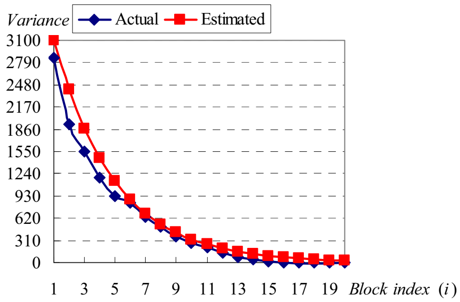


Fig. 10. The curve "Actual" shows the variances of the significant pixels in the blocks in the decreasingly sorted list of motion activity blocks in each non-key frame in the *Ballroom* and *Exit* sequences. All the variances for the same block index are averaged. The curve "Estimated" shows the exponential function $H(t) = B_1 e^{-B_2 t}$, $t = i / N_b$, used to model these actual variances.

To get the continuous-time version of the first term of Eq. (13), we let $T = \left(\prod_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}} \sigma_{i,Inter}^2\right)^{\frac{1}{N_{Inter}}}$, and obtain

$$\ln T = \frac{1}{N_{Inter}} \sum_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}} \left(\ln \sigma_{i,Inter}^2\right). \tag{25}$$

By considering the continuous-time version of ln*T* in Eq. (25), we get:

$$\ln T = \frac{N_b}{N_{Inter}} \sum_{t=\frac{N_{Intra}+1}{N_b}}^{\frac{N_{Intra}+N_{Inter}}{N_b}} \ln H(t) = \frac{1}{Y}\int_{X}^{X+Y} \ln\left(B_1 e^{-B_2 t}\right) dt \cdot \tag{26}$$

Then, *T* can be derived as

$$T = B_1 \times e^{-B_2\left(X+\frac{Y}{2}\right)}. \tag{27}$$

By applying the Taylor expansion technique, *T* can be approximated as:

$$T \approx B_1 \times h(X,Y),\ 0 \le X,\ Y \le 1 \text{ and } X + Y \le 1, \tag{28}$$

where $h(X, Y) = h_1(X) \times h_2(Y)$, and

$$h_1(X) = \left(0.5B_2^2 e^{-0.3B_2}\right)X^2 - B_2 e^{-0.3B_2}\left(1+0.3B_2\right)X + e^{-0.3B_2}\left(0.045B_2^2 + 0.3B_2 + 1\right), \tag{29}$$

$$h_2(Y) = \left(0.125B_2^2 e^{-0.2B_2}\right)Y^2 - B_2 e^{-0.2B_2}\left(0.5+0.1B_2\right)Y + e^{-0.2B_2}\left(0.02B_2^2 + 0.2B_2 + 1\right). \tag{30}$$

The accuracy of the approximation for *T* is shown in Fig. 11.

On the other hand, as the motion activity decreases, the MSE of the insignificant pixels for a block with inter mode will be also decreased. Based on Fig. 12, it is reasonable to model $\delta_{i,Inter}^2$ as a decreasing linear function:

$$I(t) = C\cdot(1 - t),\ C > 0,\ 0 \le t \le 1,\ t = i / N_b,\ N_{Intra} + 1 \le i \le N_{Intra} + N_{Iner}. \tag{31}$$

It can be observed from Fig. 12 that when the block index $i \ge 16$ in the total 20 blocks, the function $I(t)$ (the "Estimated" curve of Fig. 12) is somewhat inaccurate in modeling $\delta_{i,Inter}^2$. However, the latter few blocks in the decreasingly sorted list of motion activity blocks for a non-key frame are usually with skip mode; hence, using $I(t)$ to model $\delta_{i,Inter}^2$ is promising. The parameter *C* in Eq. (31) can be derived from the previous PRD optimization result. Assume N_{Intra_pre} and N_{Inter_pre} denote the numbers of blocks with intra mode and inter mode, respectively, in a non-key frame, obtained from the previous coding mode decision. Hence, in the current non-key frame, *C* can be estimated as follows:

$$\frac{1}{N_b} \sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}} \delta_{i,Inter}^2 = \sum_{t=\frac{N_{Intra\_pre}+1}{N_b}}^{\frac{N_{Intra\_pre}+N_{Inter\_pre}}{N_b}} I(t) = \int_{\frac{N_{Intra\_pre}+1}{N_b}}^{\frac{N_{Intra\_pre}+N_{Inter\_pre}}{N_b}} C(1-t)dt, \tag{32}$$

and

$$C = \frac{2N_b \sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}} \delta_{i,Inter}^2}{2N_b N_{Inter\_pre} - 2N_b + 2N_{Intra\_pre} - 2N_{Intra\_pre}N_{Inter\_pre} - N_{Inter\_pre}^2 + 1}. \tag{33}$$

By considering the continuous-time version of the second term of Eq. (13), we can get:

$$\frac{1}{N_{Inter}}\sum_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}}\delta_{i,Inter}^{2}=\frac{N_b}{N_{Inter}}\sum_{t=\frac{N_{Intra}+1}{N_b}}^{\frac{N_{Intra}+N_{Inter}}{N_b}} I(t)=\frac{1}{Y}\int_{X}^{X+Y}C(1-t)dt=C(1-X-0.5Y)\cdot \tag{34}$$

Hence, based on Eqs. (13), (25), (28), and (34), D_{Inter} can be derived as:

$$D_{Inter}(X,Y,R_{Inter})=B_1h(X,Y)\cdot 2^{-2\gamma R}+C(1-X-0.5Y). \tag{35}$$

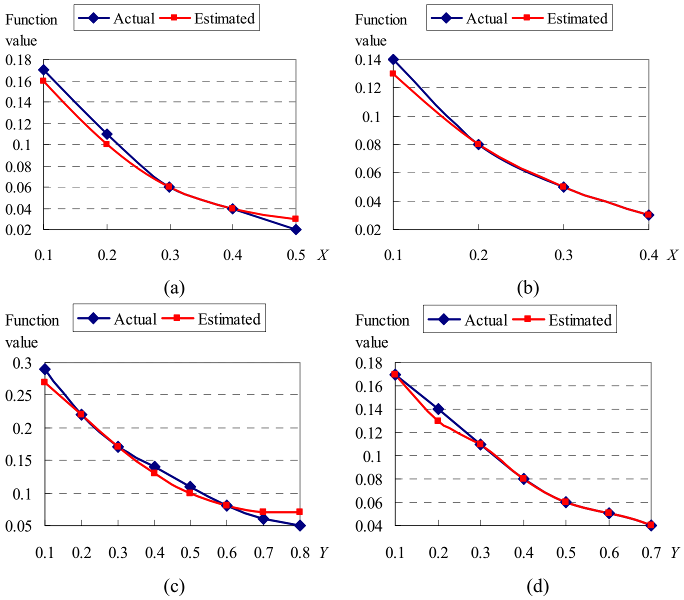


Fig. 11. The curve "Actual" shows the function $e^{-B_2\left(X+\frac{Y}{2}\right)}$ in Eq. (27), and the curve "Estimated" shows its approximation function $h(x, y)$: (a) Y is fixed to be 0.5; (b) Y is fixed to be 0.6; (c) X is fixed to be 0.2; and (d) X is fixed to be 0.3.

Finally, D_{Skip} in Eq. (14) can be derived as follows. By considering the inverse order of the decreasing motion activity block list, as the motion activity increases, the MSE of a block with skip mode will be increased as shown in the "Actual" curve of Fig. 13. Based on this, it is reasonable to model $\delta_{i,Skip}^{2}$ as an increasing exponential function as:

$$K(t)=D_1e^{D_2t},\ D_1>0,\ D_2>0,\ 0\le t\le 1,\ t=i\,/\,N_b,\ 1\le i\le N_{Skip}. \tag{36}$$

It can be observed from Fig. 13 that when the inverse block index $i \ge 14$ in all 20 blocks, the function $K(t)$ (the "Estimated" curve of Fig. 13) is somewhat inaccurate in modeling $\delta_{i,Skip}^{2}$. However, the latter few blocks in an inverse decreasing motion activity block list for a non-key frame are usually with intra or inter modes, hence using $K(t)$ to model $\delta_{i,Skip}^{2}$ is promising. The parameter D_1 in Eq. (36) can be derived from the previous PRD optimization result. Assume N_{Skip_pre} denotes the number of blocks with skip mode in a non-key frame, obtained from the previous coding mode decision. Hence, in the current non-key frame, D_1 can be estimated from

$$\frac{1}{N_b}\sum_{i=N_b-N_{Skip\_pre}+1}^{N_b}\delta_{i,Skip}^{2}=\int_{0}^{\frac{N_{Skip\_pre}}{N_b}}K(t)dt=\int_{0}^{\frac{N_{Skip\_pre}}{N_b}}D_1e^{D_2t}dt$$

as:

$$D_1 = \frac{D_2 \sum_{i=N_b-N_{Skip\_pre}+1}^{N_b} \delta_{i,Skip}^2}{N_b\left(e^{D_2\left(\frac{N_{Skip\_pre}}{N_b}\right)}-1\right)}, \tag{37}$$

where D_2 controls the increment speed of the exponential function $K(t)$, which can be obtained by some pre-training for each sequence. Usually, D_2 is a constant for the same scene.

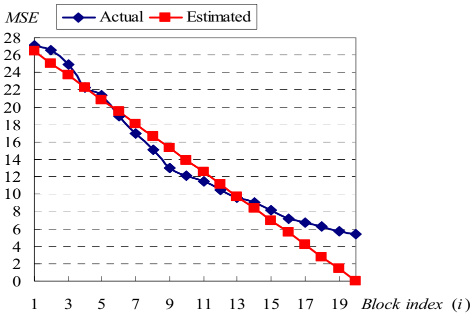


Fig. 12. The curve "Actual" shows the MSEs of the insignificant pixels in the decreasingly sorted list of motion activity blocks in each non-key frame in the *Ballroom* and *Exit* sequences. All the MSEs for the same block index are averaged. The curve "Estimated" shows the decreasing linear function $I(t) = C(1 - t)$, $t = i / N_b$, used to model these actual MSEs.

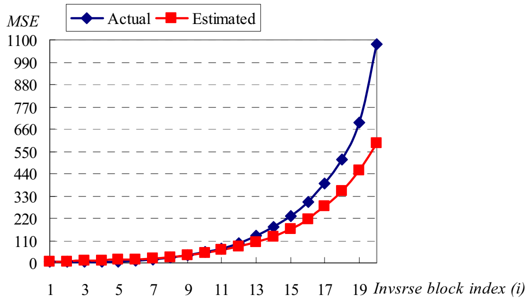


Fig. 13. The curve "Actual" shows the MSEs of the blocks in the inversed order of the decreasingly sorted list of motion activity blocks in each non-key frame in the *Ballroom* and *Exit* sequences. All the MSEs for the same block index are averaged. The curve "Estimated" shows the exponential function $K(t) = D_1 e^{D_2 t}$, $t = i / N_b$, used to model these actual MSEs.

By approximating $\delta_{i,Skip}^2$ in Eq. (14) using Eq. (37) and transferring Eq. (14) into a continuous form, we have

$$D_{Skip} = \frac{1}{N_{Skip}} \sum_{i=N_{Intra}+N_{Inter}+1}^{N_b} \delta_{i,Skip}^2 = \frac{N_b}{N_{Skip}} \sum_{t=\frac{1}{N_b}}^{\frac{N_{Skip}}{N_b}} K(t) = \frac{1}{Z}\int_0^Z D_1 e^{D_2 t}\,dt = \frac{D_1}{ZD_2}\left(e^{D_2 Z}-1\right). \tag{38}$$

By applying the Taylor expansion technique and based on Fig. 14, Eq. (38) can be approximated as:

$$D_{Skip} = \frac{D_1}{ZD_2}\left(k(Z)-1\right), \text{ where } k(Z) = \left(0.5D_2^2 e^{0.3D_2}\right)Z^2 + D_2 e^{0.3D_2}\left(1-0.3D_2\right)Z + e^{0.3D_2}\left(1-0.3D_2+0.045D_2^2\right). \tag{39}$$

It can be observed from Fig. 14 that the function $e^{D_2 Z}$ can be accurately estimated by *k*(*Z*) when $0 \le Z \le 0.5$. For video sequences with medium or larger motion, the percentage of blocks with skip mode is usually smaller or slightly larger than 50%; thus, using *k*(*Z*) to estimate $e^{D_2 Z}$ is promising. Then, based on Eq. (39), D_{Skip} can be derived as:

$$D_{Skip}(X,Y) = \frac{D_1}{(1-X-Y)D_2}\left[k(1-X-Y)-1\right]\cdot \tag{40}$$

In summary, the overall distortion function can be derived based on Eqs. (15), (22), (35), and (40) as:

$$\begin{aligned} D_{Overall}(X,Y,R) &= X \times D_{Intra}(X,R) + Y \times D_{Inter}(X,Y,R) + (1-X-Y) \times D_{Skip}(X,Y) \\ &= AX(1-0.5X)\cdot 2^{-2\gamma R} + B_1 Y h(X,Y)\cdot 2^{-2\gamma R} + CY(1-X-0.5Y) + \frac{D_1}{D_2}\left[k(1-X-Y)-1\right] \\ &= \left(AX-0.5AX^2\right)\cdot 2^{-2\gamma R} + P(X,Y)\cdot 2^{-2\gamma R} + Q(X,Y), \end{aligned} \tag{41}$$

where

$$\begin{aligned} P(X,Y) = {} & \left(0.0625 B_1 B_2^4 e^{-0.5B_2}\right) X^2 Y^3 - B_1 B_2^3 e^{-0.5B_2}\left(0.25 + 0.05B_2\right) X^2 Y^2 + B_1 B_2^2 e^{-0.5B_2}\left(0.01B_2^2 + 0.1B_2 + 0.5\right) X^2 Y - \\ & B_1 B_2^3 e^{-0.5B_2}\left(0.125 + 0.0375B_2\right) XY^3 + B_1 B_2^2 e^{-0.5B_2}\left(0.03B_2^2 + 0.25B_2 + 0.5\right) XY^2 - \\ & B_1 B_2 e^{-0.5B_2}\left(0.006B_2^3 + 0.08B_2^2 + 0.5B_2 + 1\right) XY + B_1 B_2^2 e^{-0.5B_2}\left(0.005625B_2^2 + 0.0375B_2 + 0.125\right) Y^3 - \\ & B_1 B_2 e^{-0.5B_2}\left(0.0045B_2^3 + 0.0525B_2^2 + 0.25B_2 + 0.5\right) Y^2 + B_1 e^{-0.5B_2}\left(0.0009B_2^4 + 0.015B_2^3 + 0.125B_2^2 + 0.5B_2 + 1\right) Y\,, \end{aligned} \tag{42}$$

$$\begin{aligned} Q(X,Y) = {} & \left(0.5 D_1 D_2 e^{0.3D_2}\right) X^2 + 0.5\left(D_1 D_2 e^{0.3D_2} - C\right) Y^2 - D_1 e^{0.3D_2}\left(0.7D_2 + 1\right) X - \left(0.7 D_1 D_2 e^{0.3D_2} + D_1 e^{0.3D_2} - C\right) Y + \\ & \left(D_1 D_2 e^{0.3D_2} - C\right) XY + \left(0.245 D_1 D_2 e^{0.3D_2} + 0.7 D_1 e^{0.3D_2} + \frac{D_1}{D_2} e^{0.3D_2} - \frac{D_1}{D_2}\right), \end{aligned} \tag{43}$$

A, *B₁*, *C*, and *D₁* are defined in Eqs. (18), (24), (33), and (37), respectively, and B_2 and D_2 are derived from pre-training, which are all derived from video contents.

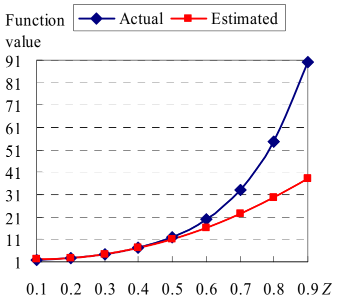


Fig. 14. The curve "Actual" shows the function $e^{D_2 Z}$ in Eq. (38), and the curve "Estimated" shows its approximation function *k*(*Z*).

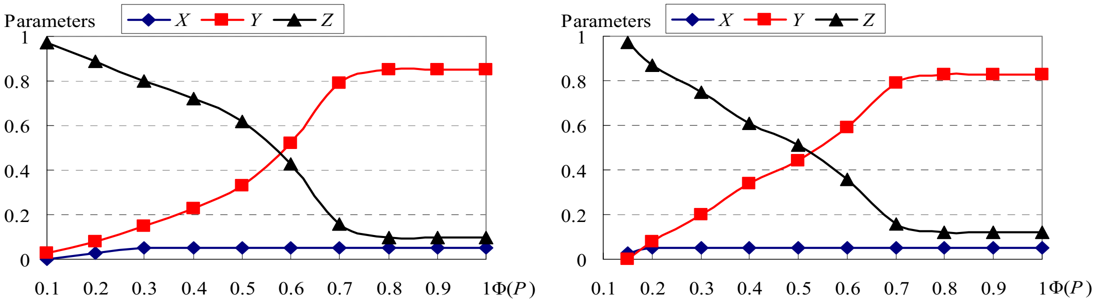
Hence, the overall PRD optimization problem can be formulated as:

$$\begin{aligned} & \min_{\{X,Y\}} D_{Overall}(X,Y,R) = \min_{\{X,Y\}} \left\{ A\left(X - 0.5X^2\right)\cdot 2^{-2\gamma R} + P(X,Y)\cdot 2^{-2\gamma R} + Q(X,Y) \right\}, \\ & \text{s.t. } F(C_1 X + C_2 Y + C_3 R) \le \Phi(P), \end{aligned} \tag{44}$$

where F, C_1, C_2, C_3, R, and $\Phi(P)$ are defined in Eq. (9), $P(X, Y)$ and $Q(X, Y)$ are, respectively, defined in Eqs. (42) and (43). Based on the proposed PRD model, before encoding a non-key frame, the parameters $\{X, Y, Z\}$, where $Z = 1 - X - Y$, can be efficiently solved based on the current available power P and the target bit rate R to minimize the overall distortion $D_{Overall}(X, Y, R)$. That is, the coding mode for each block can be determined based on the available resources while optimizing the reconstructed video quality. When the motion activity of captured video sequence is not too large, the resource allocation procedure can be performed only once every few seconds. The major objective to represent the distortion function in Eq. (44), using the Taylor approximation, in terms of the polynomial of X and Y is that it is expected to more easily find the close form for solving X and Y in minimizing the distortion function. Based on the Lagrangian multiplier technique, two very complex close forms have been, respectively, derived for X and Y. However, it cannot guarantee that the derived X and Y will be always within the constraint of [0, 1]. That is, the minimum of Eq. (44) will not always occur when X and Y are simultaneously within [0, 1]. As a result, discrete sampling on X and Y is used to achieve efficient implementation. Specifically, only a few points, $(X, Y) = (0.05x, 0.05y)$, $x = 0, 1, 2, \ldots, 19$, $y = 0, 1, 2, \ldots, 19$, under the constraints, $0 \le X \le 1$, $0 \le Y \le 1$, $0 \le X + Y \le 1$, and $F(C_1X + C_2Y + C_3R) \le \Phi(P)$, are evaluated to find the optimal point (X, Y) in minimizing Eq. (44).

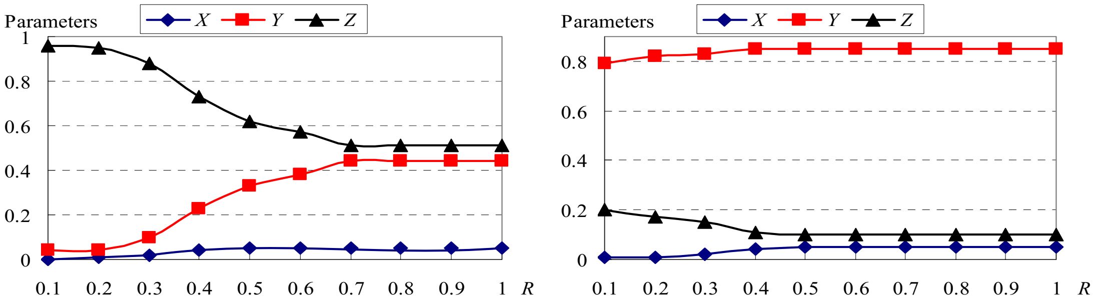
The average optimal parameter sets, $\{X, Y, Z\}$, for the *Ballroom* sequence, minimizing $D_{Overall}(X, Y, R)$ in Eq. (44) with the available encoding power P ranged from 0.1 to 1.0, and the encoding bit rates R fixed to 0.5bpp and 1.0bpp are shown in Fig. 15. The average optimal parameter sets $\{X, Y, Z\}$ for the *Ballroom* sequence, minimizing $D_{Overall}(X, Y, R)$ in Eq. (44) with the available encoding power P fixed to 0.5 and the encoding bit rates R ranging from 0.1 to 1.0 bpp, are shown in Fig. 16. In Figs. 15-16, the parameters, X, Y, and Z, respectively, of all the frames in the whole sequence are averaged. The analytic and actual PRD performances for the *Ballroom* sequence are shown in Fig. 17. In Fig. 17, the MSEs of all the frames in the whole sequence are averaged.

It can be observed from Fig. 15 that, under a fixed target bit rate, when the encoding power increases, Y will increase accordingly because more power is required to encode a block with inter mode. In addition, X is usually small due to the encoding performance for a block with intra mode is usually not good even though the corresponding power consumption is relatively low. Similarly, it can be observed from Fig. 16 that under a fixed medium or high power, when the bit rate increases, Y will increase and X is almost unchanged. When the power is very high (*e.g.*, $\Phi(P) = 1.0$ in Fig. 16(b)), Y will be much larger than X and Z. To evaluate the accuracy of the proposed PRD model for non-key frame coding, all the key frames are encoded with very high quality and only the PRD performance for the luminance component of the non-key frames is shown in Fig. 17. Based on Fig. 17, it can be observed that under a fixed bit rate, when the power increases, the distortion will decrease. On the other hand, under a fixed power, when the bit rate increases, the distortion will also usually decrease. However, when the power is too low, the reduction of MSE will be insignificant even when the bit rate increases. It can also be observed from Fig. 17 that the proposed PRD model is fairly accurate to estimate the actual PRD performance.



(a) (b)

Fig. 15. The average optimal parameter sets, {*X*, *Y*, *Z*}, for the *Ballroom* sequence, minimizing $D_{Overall}(X, Y, R)$ in Eq. (44) with the available encoding power *P* ranged from 0.1 to 1.0, and the encoding bit rates *R* fixed to: (a) 0.5bpp; and (b) 1.0bpp.



(a) (b)

Fig. 16. The average optimal parameter sets, {*X*, *Y*, *Z*}, for the *Ballroom* sequence, minimizing $D_{Overall}(X, Y, R)$ in Eq. (44) with the available encoding power *P* fixed to: (a) 0.5; and (b) 1.0, and the encoding bit rates *R* ranged from 0.1 to 1.0 bpp.

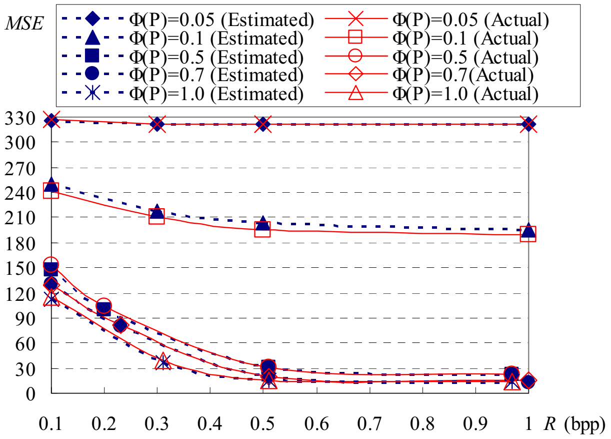


Fig. 17. The analytic and actual PRD performance for the *Ballroom* sequence. The curves "Estimated" show the PRD performance obtained from the proposed PRD model shown in Eq. (44), whereas the curves "Actual" show the actual PRD performance obtained from the proposed video codec.

VI. Simulation Results

Some multiview video sequences [32] consisting of 250 frames, a frame size of 640×480, a GOP size of 4, a block size of 128×128 (n = 128), YUV4:2:0, and a frame rate of 10 frames per second (fps) were used to evaluate the proposed low-complexity multiview video codec under different available resources, *i.e.*, encoding powers and target bit rates. The hash length L is set to 128, 256, or 512 based on the available resources. The more the available resources, the longer the hash length is. The quantization parameter (QP) for each H.264/AVC intra-encoded key frames ranged from 20 to 40. The first two views (VSNs), V_0, and V_1, structured based on TABLE I are considered, where the distance between V_0 and V_1 is 19.5 cm [32]. The proposed low-complexity single-view video codec (denoted by "Proposed Single"), the H.264/AVC intraframe coding (denoted by "H.264 Intra") [3], and the H.264/AVC interframe coding with no motion (denoted by "H.264 No motion") (where all the motion vectors are set to zeros [13]) were employed for comparisons with the proposed multiview video codec. These four approaches are all with low complexity. It should be noted that the studies of power-scalable low-complexity multiview video coding have not appeared in the literature. Hence, only some baseline low-complexity video codecs were selected for comparison with the proposed codec. For V_0 (the first view), all the frames belong to key frames, and are encoded using the H.264/AVC intraframe coding [3] with QP set to 16. The PRD performance for V_1 (the second view) of the proposed multiview video codec and the RD performance for the three approaches used for comparisons are shown in Figs. 18-19, respectively, for the *Ballroom* and *Exit* video sequences. In Figs. 18-19, the PSNR values of all the frames (key frames and non-key frames) in each sequence are averaged.

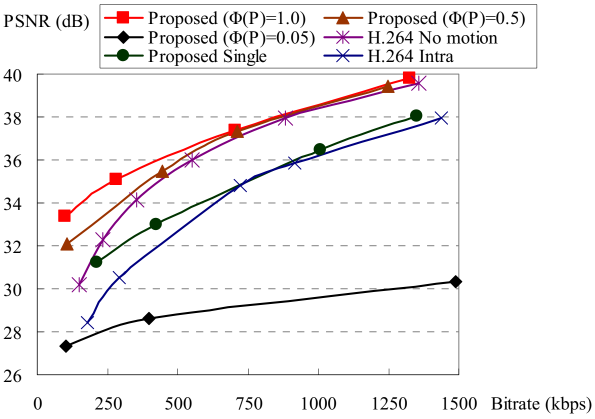


Fig. 18. The RD performance for the *Ballroom* sequence.

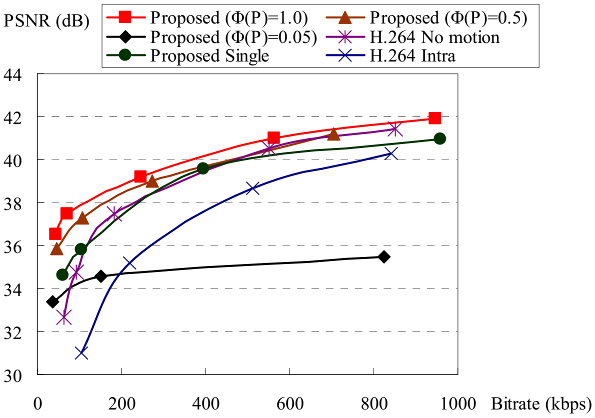


Fig. 19. The RD performance for the *Exit* sequence.

For the *Ballroom* sequence, it can be observed from Fig. 18 that the PSNR performance gains of the proposed multiview video codec at Φ(P) = 1.0 above those of the H.264/AVC interframe coding with no motion are from 0.1 to 4 dB. The PSNR performance gains of the proposed codec at Φ(P) = 1.0 above those of the H.264/AVC intraframe coding are from 2 to 6 dB. The RD performance of the proposed codec at Φ(P) = 0.5 is very close to those of the proposed codec at Φ(P) = 1.0, especially at higher bit rates. The RD performance of the proposed codec at $\Phi(P) = 0.05$ is very poor. The PSNR performance gains of the proposed multiview codec at higher powers can significantly outperform the proposed single-view codec.

Similarly, for the *Exit* sequence, it can be observed from Fig. 19 that the PSNR performance gains of the proposed multiview video codec set at $\Phi(P) = 1.0$ range from 0.5 to 4 dB above those of the H.264/AVC interframe coding with no motion. The PSNR performance gains of the proposed codec set at $\Phi(P) = 1.0$ range from 2 to 6 dB above those of the H.264/AVC intraframe coding. The RD performance of the proposed codec at $\Phi(P) = 0.5$ is very close to those of the proposed codec at $\Phi(P) = 1.0$, especially at higher bit rates. The RD performance of the proposed codec at $\Phi(P) = 0.05$ is very poor. The PSNR performance gains of the proposed multiview codec at higher powers can outperform the proposed single-view codec, but the performance gains are not large.

More specifically, based on Figs. 18-19, the proposed multiview video codec can outperform the three approaches used for comparisons, especially at high power and low bit rates. That is, when the encoding power is high, the proposed encoder can efficiently exploit the available bit rates to optimize the video quality, even though the bit rate is low. In addition, with the benefits of exploiting the reference frames from the adjacent view, the proposed encoder can have more skipped SDS symbols or skipped blocks, which can save more bit rates. On the other hand, at higher bit rates, the RD performance of the proposed multiview codec can still significantly outperform the H.264/AVC intraframe coding, but is very close to that of the H.264/AVC interframe coding with no motion. That is, for a fixed power, excess bit rates cannot be efficiently exploited, and this is consistent with the analytic PRD results shown in Fig. 17, where the RD curves will be flatter while the bit rates are greatly increased. It is also consistent with the block coding mode decision results shown in Fig. 16, where the configurations of X, Y, and Z will be unchanged while the bit rates are greatly increased, which will result in similar RD performance. On the other hand, when the power is low, the RD performance of the proposed codec is poor and the RD curves are flatter, which mean the bit rates cannot be efficiently exploited. It can be observed from Figs. 15, 17-19 that when the power is low, the block coding modes are almost determined to be the skip mode, which will result in poor RD performance for the video sequences with medium or large motion. Oppositely, when the power is high, the RD performance will be better, but when the power reaches a certain level, the RD performance improvement gaps will be degraded, which means excess power cannot be efficiently exploited, and will not change the block coding mode decision results too much. In addition, it can be observed from Figs. 18-19 that the RD performance gaps between the proposed multiview codec and those of the proposed single-view codec are larger for the video sequence with large motions (*e.g.*, the *Ballroom* sequence), whereas those gaps are smaller for the video sequence with medium or small motions (*e.g.*, the *Exit* sequence).

For the two H.264/AVC codecs used for comparison, namely, the H.264/AVC intraframe coding (H.264 Intra) [3] and the H.264/AVC interframe coding with no motion (H.264 No motion) [13], the proposed multiview codec can significant outperform the H.264 Intra codec for most motion types of video. The H.264 Intra encoder has been shown to be a low-complexity and efficient encoder, which can outperform or be comparable to several current multiview DVC codecs, where no interview communication is performed at the encoder [18]. The proposed multiview codec exploits interview correlation at the encoder via a little interview hash data exchanges; and can, therefore, significantly outperform the H.264 Intra codec. On the other hand, the H.264 No motion encoder has been shown to be a low-complexity and very efficient encoder, which is hard to beat [13]. The proposed multiview codec can significantly outperform the H.264 No motion codec at the low bit rates, which is a benefit for wireless visual sensor network (WVSN) applications with limited bandwidth. For the high bit rate situations, the proposed multiview codec should be further improved to get better RD performance. However, the power-scalability characteristic and the proposed PRD optimization technique are worthy for most low-complexity video coding applications.

VII. Conclusions and Future Works

In this paper, we have proposed a resource-scalable low-complexity multiview distributed video coding scheme. We present a PRD model to characterize the relationship between the available resources (*e.g.*, power supply and target bit rate) and the RD performance of the proposed video codec. More specifically, an RD function in terms of the percentages for different coding modes of blocks and the target bit rate under the available resource constrains is derived for optimal block coding mode decision. Based on this model, the resource allocation can be efficiently performed at the encoder while optimizing the reconstructed video quality. Analytic results have been provided to verify the resource scalability and accuracy of the proposed PRD model. The coding efficiency of the proposed low-complexity video codec has been demonstrated via simulation results to outperform three known low-complexity video codecs, especially at the high powers and low bit rates.

For future work, the distortion induced by wireless video transmission will be integrated into the current distortion function to form a complete end-to-end video distortion function. More precise theoretical analyses, such as the optimal achievable video quality based on available resources and the minimum resource requirements based on acceptable video distortion, can be made to provide a practical guideline in preparation and deployment for a WVSN.

References

[1] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, “Wireless multimedia sensor networks: a survey,” *IEEE Wireless Communications*, vol. 14, no. 6, pp. 32-39, Dec. 2007.

[2] T. Sikora, “Trends and perspectives in image and video coding,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 6-17, Jan. 2005.

[3] T. Wiegand and G. J. Sullivan, “The H.264/AVC video coding standard,” *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 148-153, March 2007.

[4] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, “Multiview imaging and 3DTV,” *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10-21, Nov. 2007.

[5] M. Flierl and B. Girod, “Multiview video compression,” *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 66-76, Nov. 2007.

[6] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis, and A. Koz, “Coding algorithms for 3DTV-a survey,” *IEEE*

Trans. on Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1606-1621, Nov. 2007.

[7] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for multiview video," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1474-1484, Nov. 2007.

[8] X. Guo, Y. Lu, F. Wu, and W. Gao, "Inter-view direct mode for multiview video coding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 12, pp. 1527-1532, Dec. 2006.

[9] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71-83, Jan. 2005.

[10] R. Puri, A. Majumdar, P. Ishwar, and K. Ramchandran, "Distributed video coding in wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 94-106, July 2006.

[11] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi. R. Leonardi, and J. Ostermann, "Distributed monoview and multiview video coding: basics, problems and recent advances," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 67-76, Sept. 2007.

[12] F. Pereira, L. Torres, C. Guillemot, T. Ebrahimi, R. Leonardi, and S. Klomp, "Distributed video coding: selecting the most promising application scenarios," to appear in *Signal Processing: Image Communication*.

[13] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: architecture, techniques and evaluation," in *Proc. of 2007 Picture Coding Symposium*, Lisbon, Portugal, Nov. 2007.

[14] Z. Li, L. Liu, and E. J. Delp, "Rate distortion analysis of motion side estimation in Wyner-Ziv video coding," *IEEE Trans. on Image Processing*, vol. 16, no. 1, pp. 98-113, Jan. 2007.

[15] X. Guo, Y. Lu, F. Wu, D. Zhao, and W. Gao, "Wyner-Ziv-based multi-view video coding," to appear in *IEEE Trans. on Circuits and Systems for Video Technology*.

[16] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Fusion-based multiview distributed video coding," in *Proc. of ACM Int. Workshop on Video Surveillance and Sensor Networks*, Santa Barbara, CA, USA, Oct. 27, 2006.

[17] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Multiview distributed video coding with encoder driven fusion", in *Proc. of European Signal Processing Conf.*, Poznan, Poland, Sept. 2007.

[18] X. Artigas, F. Tarres, and L. Torres, "Comparison of different side information generation methods for multiview distributed video coding," in *Proc. of Int. Conf. on Signal Processing and Multimedia Applications*, Barcelona, Spain, July 2007.

[19] M. Morbee, L. Tessens, H. Q. Luong, J. Prades-Nebot, A. Pizurica, and W. Philips, "A distributed coding-based content-aware multi-view video system," in *Proc. of ACM/IEEE Int. Conf. on Distributed Smart Cameras*, Vienna, Austria, Sept. 2007, pp. 355-362.

[20] C. Yeo and K. Ramchandran, "Robust distributed multi-view video compression for wireless camera networks," in *Proc. of SPIE Visual Communications and Image Processing*, vol. 6508, Jan. 2007.

[21] M. Wu and C. W. Chen, "Collaborative image coding and transmission over wireless sensor networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 70481, 9 pages, 2007, special issue on Visual Sensor Networks.

[22] K. Y. Chow, K. S. Lui, and E. Y. Lam, "Efficient on-demand image transmission in visual sensor networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 95076, 11 pages, 2007, special issue on Visual Sensor Networks.

[23] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, Sept. 2007.

[24] P. V. Pahalawatta and A. K. Katsaggelos, "Review of content-aware resource allocation schemes for video streaming over wireless networks," *Wiley InterScience Wireless Communications and Mobile Computing*, vol. 7, no. 2, pp. 131-142, 2007.

[25] Z. He and S. K. Mitra, "From rate-distortion analysis to resource-distortion analysis," *IEEE Circuits and Systems Magazine*, vol. 5, no. 3, pp. 6-18, 2005.

[26] Z. He, Y. Liang, L. Chen, I. Ahmad, and D. Wu, "Power-rate-distortion analysis for wireless video communication under energy constraints," *IEEE Trans. on*

Circuits and Systems for Video Technology, vol. 15, no. 5, pp. 645-658, May 2005.

[27] Z. He and D. Wu, "Resource allocation and performance analysis of wireless video sensors," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 5, pp. 590-599, May 2006.

[28] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. on Information Theory*, vol. IT-22, no. 1, pp. 1-10, Jan. 1976.

[29] C. S. Lu and H. Y. M. Liao, "Structural digital signature for image authentication: an incidental distortion resistant scheme," *IEEE Trans. on Multimedia*, vol. 5, no. 2, pp. 161-173, June 2003.

[30] C. S. Lu, "On the security of structural information extraction/embedding for images," in *Proc. of IEEE Int. Symposium on Circuits and Systems*, Vancouver, BC, Canada, May 2004, vol. 2, pp. 169-172.

[31] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710-732, July 1992.

[32] Mitsubishi Electric Research Laboratories, "MERL multi-view video sequences," ftp://ftp.merl.com/pub/avetro/mvc-testseq.

[33] L. W. Kang and C. S. Lu, "Low-complexity Wyner-Ziv video coding based on robust media hashing," in *Proc. of IEEE Int. Workshop on Multimedia Signal Processing*, Victoria, BC, Canada, Oct. 2006, pp. 267-272.

[34] L. W. Kang and C. S. Lu, "Multi-view distributed video coding with low-complexity inter-sensor communication over wireless video sensor networks," in *Proc. of IEEE Int. Conf. on Image Processing,* special session on Distributed source coding II: Distributed video and image coding and their applications, San Antonio, TX, USA, Sept. 2007, vol. 3, pp. 13-16 (invited paper).

[35] L. W. Kang and C. S. Lu, "Low-complexity power-scalable multi-view distributed video encoder," in *Proc. of Picture Coding Symposium*, Lisbon, Portugal, Nov. 2007.