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

TR-IIS-09-004

Power-Rate-Distortion Optimized Resource-Scalable Low-Complexity Video Coding in Wireless Multimedia Sensor Networks

Li-Wei Kang and Chun-Shien Lu

April 29, 2009 || Technical Report No. TR-IIS-09-004

http://www.iis.sinica.edu.tw/page/library/LIB/TechReport/tr2009/tr09.html

Power-Rate-Distortion Optimized Resource-Scalable Low-Complexity Video Coding in Wireless Multimedia Sensor Networks

Li-Wei Kang and Chun-Shien Lu

***Abstract*—Wireless multimedia sensor networks (WMSNs) have been potentially applicable for several emerging applications. However, the available resources, *i.e.*, power and rate, of visual sensors in a WMSN are very limited. Hence, it is important but challenging to achieve efficient resource allocation and optimal video compression while maximizing the overall network lifetime. In this paper, a power-rate-distortion (PRD) optimized resource-scalable low-complexity multiview video encoding scheme is proposed. In our video encoder, both the temporal and interview information can be efficiently exploited based on the comparisons of extracted media hashes without performing motion and disparity estimations, which are known to be time-consuming. We present a PRD model to characterize the relationship between the available resources and the RD performance of our encoder. 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 coding mode decision. Analytic results are provided to verify the resource scalability and accuracy of our PRD model, which can provide a theoretical guideline for performance optimization under limited resource constraints. Both the analytic and simulation results have shown the applicability of our video coding scheme for WMSNs.**

***Index Terms*—Low-complexity video coding, multiview video coding, resource-scalable video coding, power-rate-distortion optimization, wireless multimedia sensor networks, robust media hash.**

I. Introduction

A. Background

With the availability of low-cost hardware, wireless multimedia sensor networks (WMSNs) have been potentially applicable for several emerging applications, such as security monitoring and environmental tracking [1]-[2]. A WMSN is a network of several wireless embedded devices supporting to retrieve visual, acoustic, and scalar data from a monitored physical environment. Here, a WMSN consisting of several battery-powered visual sensor nodes (VSNs) scattered in several sensor fields is considered, as shown in Fig. 1. Each VSN equipped with a low-cost camera (*e.g.*, complementary metal-oxide-semiconductor, *i.e.*, CMOS 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), usually supporting a powerful decoder for video decoding. Compared with traditional network systems, WMSN operates under several resource constraints (*e.g.*, lower computational capability, limited power supply, and narrow transmission bandwidth). This will pose an important problem of simultaneously minimizing the power consumption and optimizing the video compression performance for each VSN in a WMSN while maximizing the overall network lifetime [3].

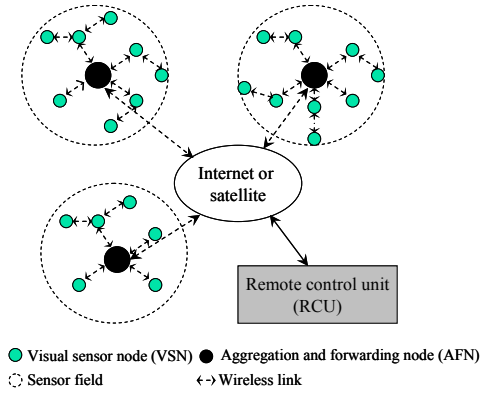

Fig. 1. An illustrative example of a wireless multimedia sensor network (WMSN) architecture.

Based on the experimental analysis presented in [4], in typical scenarios of wireless video communication among visual sensor devices, the video encoding process consumes a significant portion (about 40%~60%) of the total power consumption for a video device. For example, for a well-known WMSN hardware platform, called Crossbow Stargate [2], equipped with a low-power USB video camera, the video encoder consumes about 48% of the total power while the wireless transmission consumes about 11% of the total power [4]. In this paper, we focus on minimization of the power consumptions for the two major components (video encoding and wireless video transmission) to prolong the operational lifetime of each wireless visual sensor in order to maximize the overall network lifetime while optimizing video compression

Li-Wei Kang is with Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC (e-mail: lwkang@iis.sinica.edu.tw).

Chun-Shien Lu is with Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC (phone: +886-2-2788-3799 ext. 1513; fax: +886-2-2782-4814; e-mail: lcs@iis.sinica.edu.tw).

performance. The power consumption of the remaining components depends on specific system design and cannot be easily controlled from a video encoding perspective [4].

To reduce the power consumption of both the video encoder and wireless video transmission for a VSN, low-complexity and high-efficiency video encoding is critically desired. If the video encoding complexity can be minimized while certain compression efficiency can be kept, the power consumption for both video encoding and video data transmission can be simultaneously decreased. However, current video coding standards (*e.g.*, H.264/AVC [5]) usually perform complex interframe encoding (*e.g.*, motion estimation with high computational complexity for exploiting temporal correlation of successive frames). On the other hand, to sufficiently exploit correlations among adjacent VSNs in a WMSN, current multiview video coding (MVC) schemes (*e.g.*, H.264/AVC-based MVC, *i.e.*, joint multiview video model) [6]-[7] usually perform both interview (*e.g.*, disparity estimation) and temporal (*e.g.*, motion estimation) predictions at the encoder with high complexity. In addition, to perform interview prediction, uncompressed frames must be exchanged among VSNs, which is prohibitive in a WMSN.

B. Related Works

To reduce the power consumption of a video encoder, several approaches have been proposed, including motion estimation-based low-complexity video encoder designs [8]-[9], hardware-based video encoder designs [10], and joint video encoder/decoder and hardware adaptation schemes [11]-[12]. It is claimed that the existing approaches, so far, focused on reduction of encoder complexity and power consumption through heuristic adaption, instead of systematic power optimization, due to their lack of an analytic model to characterize the optimal trade-off between power consumption and video encoding efficiency [4]. Hence, in [4], [13]-[14], a power-rate-distortion (PRD) optimization framework is proposed for optimal resource allocation and performance optimization of wireless video communication among video sensor devices. Based on the PRD model, the minimum video distortion that a video encoder can achieve under current power and rate constraints can be derived.

However, the above-mentioned approaches [4], [8]-[14] are all based on a standard or motion estimation-based video encoder, which is essentially with high encoding complexity. Recently, several "motion estimation-free" low-complexity video encoding schemes have been proposed, which can be roughly classified into three categories shown as follows.

(i) Still image-based or standard codec-based low-complexity video coding [15]-[18]: without performing motion estimation, the most straightforward video encoder is to apply still-image/intraframe encoding to each frame individually. In [18], the H.264/AVC intraframe encoder was used as baseline benchmarking to evaluate their low-complexity encoding scheme. To further exploit temporal correlation, H.264/AVC interframe coding with no motion, where all the motion vectors are set to zeros, has been shown to be very efficient and difficult to be defeated [18].

(ii) Collaborative video coding and transmission [19]-[20]: To further increase coding efficiency, interview correlation among VSNs can be exploited via collaborative video coding and transmission. While transmitting the intra-encoded frames from adjacent VSNs toward the AFN through the same intermediate node, this node can perform an image matching procedure to detect the similar regions, which can be encoded once only for these frames. However, image matching is usually a complex task, but current researches [19]-[20] usually assume that they can be easily detected or already known as prior knowledge.

(iii) Distributed video coding (DVC) [18], [21]-[24]: The major characteristic of DVC is that individual frames are encoded independently, but decoded jointly. The major encoding burden, *i.e.*, motion estimation, can be shifted to the decoder while preserving a certain coding efficiency. More specifically, DVC models lossy video coding as a channel coding problem based on Wyner-Ziv information theory. The statistical dependency between two correlated sources, a frame W (called Wyner-Ziv frame) and its side information Y, is modeled as a virtual correlation channel. At the encoder, the compression of W can be achieved by transmitting only part of the parity bits (called Wyner-Ziv bits) derived from the channel-encoded version of W. The decoder uses the received Wyner-Ziv bits and the side information Y derived from previous decoded video signals to perform channel decoding for the reconstruction of W. The side information Y can be generated by decoder-side motion estimation exploiting the temporal and/or interview correlations, respectively, from current VSN and adjacent VSNs.

C. Overview of the Proposed Scheme

So far, the existing "motion estimation-free" low-complexity video encoding schemes lack an analytic model to characterize the optimal trade-off between power consumption and video encoding efficiency. Hence, in this paper, we focus on designing a resource-scalable "motion estimation-free" low-complexity multiview video encoder (preliminarily presented in [25]-[27]), and deriving a PRD model for optimal resource allocation and performance optimization of our video encoder (preliminarily presented in [28]).

In our multiview video encoder, both the temporal and interview predictive coding can be efficiently achieved by extracting the significant differences between a frame and its reference frames, respectively, from the same VSN and the adjacent VSNs based on comparing the extracted media hashes [29] without performing motion and disparity estimations. To exploit interview correlation, limited inter-VSN communications during the encoding process are allowed to exchange hash information of relatively small size.

In particular, the unique characteristic of our scheme is that a PRD model is proposed to characterize the relationship between the available resources (*e.g.*, power supply and target bit rate) and the RD performance of our video encoder. Based

on this model, the 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, resource allocation can be efficiently performed at the encoder by adjusting the encoding parameters according to the available resources while optimizing the reconstructed video quality. It should be noted that the major goal of this paper is to propose a resource-scalable low-complexity video encoder for a WMSN and a PRD model for resource allocation and performance optimization of our encoder, instead of competing the coding performance against the existing standard, motion estimation-based, or non-resource-scalable video encoding schemes.

The remainder of this paper is organized as follows. The problem formulation of this paper is described in Sec. II. The proposed hash-based video coding scheme and the proposed low-complexity video coding scheme are described in Sec. III and Sec. IV, respectively. The proposed PRD model for resource allocation and performance optimization of our low-complexity video encoder together with analytical results is described in Sec. V. Simulation results are presented in Sec. VI. Finally, conclusions and future works are given in Sec. VII.

II. Problem Formulation

In this section, we will formulate the problems we want to solve, including (1) low-complexity video encoding without performing motion estimation and (2) PRD optimized resource allocation and performance optimization for our low-complexity video encoder, as follows. Our low-complexity video encoder is based on the proposed hash-based video coding scheme described in Sec. III, which can be formulated as follows.

Problem 1 (media hash-based video block coding): Given two video blocks, B and B', where B is the block to be encoded and B' is the reference block of B. The significant component of B, significantly different from B', can be extracted by comparing the respective media hashes of B and B'. The media hash should be properly extracted such that β is more similar to B than B', where β is an estimate of B and obtained by modifying B' using the significant component of B.

That is, the compression of a block can be achieved via encoding its significant component derived from hash comparison while the reconstruction of the block can be achieved via integrating its significant component and its reconstructed reference block. The problem 1 will be realized and solved in Sec. III, which will contribute the basis of the proposed low-complexity video encoding scheme described in Sec. IV.

On the other hand, the resource-scalability and PRD optimization of our low-complexity video encoder can be formulated as an optimal block coding mode decision problem under current resource constraints. For a frame consisting of several non-overlapping blocks, the three possible coding modes for each block are intra, inter, and skip modes with different computational complexity. First, the encoding procedure can be roughly divided into the combination of the three "atom operations," including the intra-mode block encoding, the inter-mode block encoding, and the entropy encoding operations, whose computational complexities are C_1, C_2, and C_3, respectively. Second, let X, Y, and Z, respectively, denote N_{Intra}/N_b, N_{Inter}/N_b, and N_{Skip}/N_b, where N_{Intra}, N_{Inter}, and N_{Skip} are the number of blocks with intra, inter, and skip modes, respectively, of a frame consisting of N_b ($= N_{Intra} + N_{Inter} + N_{Skip}$) blocks, and $X + Y + Z = 1$. Third, the blocks in a frame are sorted based on their motion activities estimated from the SAD (sum of absolute difference) between each block and its reference block. Overall, the optimal coding mode decision problem can be formulated as follows.

Problem 2 (optimal coding mode decision): Given the available resources, including the encoding power P, target bit rate R, and a decreasingly sorted list of motion activities of blocks, B_i, $i = 1, 2, \ldots, N_b$, in a frame, the optimization problem can be expressed as:

$$\min_{\{X,Y\}} D(X,Y,R)$$
$$\text{s.t. } F(C_1X + C_2Y + C_3R) \le \Phi(P), \quad (1)$$

where D is the distortion function, F denotes the frame rate, and Φ is the power function described in Sec. V-B.

Once the optimal values of X, Y, and Z ($= 1 - X - Y$) are derived, the optimal coding mode for each block in the decreasingly sorted list can be decided by sequentially assigning N_{Intra} ($= X \times N_b$) intra mode, N_{Inter} ($= Y \times N_b$) inter mode, and N_{Skip} ($= Z \times N_b$) skip mode blocks. The problem 2 will be realized and solved in Sec. V.

III. Proposed Hash-based Video Coding

In this section, our robust media hashing scheme [29] is first described in Sec. III-A. Then, based on this hashing scheme, a hash-based video block coding scheme is proposed in Sec. III-B, which contributes the basis of the proposed low-complexity video coding scheme described in Sec. IV.

A. Robust Media Hashing

At the encoder of our multiview video codec, temporal correlation is exploited by efficiently comparing the block-based media hash information among successive frames, while interview correlation is exploited by exchanging limited block-based hash information among adjacent VSNs. Our robust media hashing scheme, called structural digital signature (SDS) [29], which can extract the most significant components and provide a compact representation for a video block efficiently, meets the aforementioned requirements.

To extract the SDS for a video block of size $n \times n$, a J-scale DWT (discrete wavelet transform) is performed. Here, to make the SDS for a block be representative, the block size should be large enough. Let $w_s(x, y)$ represent a wavelet coefficient at scale s and position (x, y), $0 \le s < J$, $1 \le x, y \le n$. For each pair consisting of a parent node, $w_{s+1}(x, y)$, and its four child nodes,

$w_s(2x + i, 2y + j)$, $0 \le i, j \le 1$, the maximum magnitude difference (*max_mag_diff*) value is calculated as

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

All the parent-4 children pairs are then 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 determined to be significant. Each significant pair is assigned a symbol representing what kind of relationship it carries. According to the interscale relationship existing among wavelet coefficients, there are four possible types. When the magnitude of a parent node p is larger than that of its child node c with *max_mag_diff* value, *i.e.*, $|p| \ge |c|$, the four possible relationships are (a) $p \ge 0, c \ge 0$; (b) $p \ge 0, c < 0$; (c) $p < 0, c \ge 0$; and (d) $p < 0, c < 0$. To make the above relationships compact, the relations (a) and (b) can be merged to form a signature symbol "+1" when $p \ge 0$ and c is ignored. In addition, the relations (c) and (d) can be merged into another signature symbol "-1" when $p < 0$ and c is ignored. Similarly, the signature symbols "+2" and "-2" can be defined under the constraint $|p| < |c|$. Those pairs not included in the selected L pairs are labeled by "0." For an $n \times n$ block, there are at most $(n/2)\times(n/2)$ parent-4 children pairs, and hence, the SDS for the block is a symbol sequence in raster scan order, consisting of L significant symbols and $[(n/2)\times(n/2) - L]$ "0" symbols, which can be efficiently compressed via run-length and entropy coding techniques. Since the position of a parent node can indicate the positions of its child nodes, by using the coordinates, p_x and p_y, for a parent node p in an $n \times n$ block B, the SDS of B can be expressed by

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

where $S(B, p_x, p_y)$ denotes the SDS symbol of the pair with parent node position (p_x, p_y) in B. 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 Block Coding

Here, the major goal is to efficiently extract the most significant components of a block for compressing it without performing motion estimation. Based on the fact that image signals can be approximately reconstructed from their multiscale information derived from the DWT domain [30], in our video codec, the multiscale information of a 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 $S(B, p_x, p_y) \ne 0$ of the block B, if $S(B, p_x, p_y) \ne S(B', p_x, p_y)$, 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 insignificant. For each significant SDS symbol, its corresponding five wavelet coefficients will be reserved. For each insignificant SDS symbol, its corresponding five wavelet coefficients will be replaced by zeros. Then, for the block B, all the reserved coefficients can be efficiently compressed via the quantization and entropy encoding techniques for DWT-based images (*e.g.*, JPEG-2000 [31]). To reconstruct the block B, based on the reconstructed reference block B', the decoded DWT coefficients for B are used to modify B' (fill the decoded coefficients of B into the corresponding positions in the DWT representation of B', followed by performing inverse DWT) to obtain β, which will have SDS similar to that of B, and can be regarded as the reconstruction of B. Our hash-based video block coding scheme will be used to develop our low-complexity multiview video coding scheme, described in Sec. IV.

Someone may wonder why not directly compare the co-located DWT coefficients between two blocks to decide the significant coefficients. The reason is that it is hard to determine a threshold to judge the similarity between two co-located coefficients, and determine the significance of a coefficient in the block to be encoded. In our hash-based video block encoder, we just use a predefined hash length L to determine the significant DWT coefficients for a block, which can be decided empirically based on current available bit rates.

IV. Proposed Low-complexity Hash-based Video Codec

In this section, based on the hash-based video block coding scheme described in Sec. III-B, a hash-based low-complexity single-view video coding scheme is proposed in Sec. IV-A, and then extended to the low-complexity multiview video coding scheme described in Sec. IV-B. Note that our multiview video encoder exploits the two kinds of correlations via hash comparison among successive frames in a VSN and limited inter-VSN hash information exchange without performing motion estimation.

Assume that there are several adjacent VSNs observing the same target scene in a WMSN. Similar to the video structure used in [18], [22]-[24], a video sequence consists of several group of pictures (GOP), where each GOP consists of one key frame and one non-key frame, *i.e.*, GOP size is 2, which has been shown to provide the best performance in most evaluated sequences when motion estimation cannot be performed at the encoder [18]. A key frame similar to an I frame (intraframe) in traditional video coding serves as an anchor frame, which can be independently encoded and decoded. Here, each key frame is encoded using the H.264/AVC intraframe encoder [5], while each non-key frame is encoded using our low-complexity video encoder, described in this section. To consider adjacent VSNs at the same time instant, similar to [22]-[24], the adjacent frames for each key frame are non-key frames, while the adjacent frames for each non-key frame are key frames. An example of the structure with three adjacent VSNs is shown in Table 1, where $K_{s,t}$ and $W_{s,t}$, respectively, denote the key and non-key frames captured by VSN V_s at time instant t. The first frame for each VSN is forced to be a key frame, and the remaining frames are interlaced by key and non-key frames in both temporal and interview directions.

A. Block-based Single-view Video Codec

Our block-based single-view video codec is shown in Fig. 2. At the encoder, for each non-key frame $W_{s,t}$ captured by V_s at

time t, its nearest key frame $R_{s,t}$ (*e.g.*, $R_{s,t} = K_{s,t-1}$) is determined to be its reference frame. Each $W_{s,t}$ is decomposed into several non-overlapping $n \times n$ (in this paper, n is set to 128) blocks $B_{s,t,b}$, where b is the block index. The coding mode of $B_{s,t,b}$ is 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 modes, based on the proposed PRD model according to the available resources, described in Sec. V. Each block with intra mode is encoded using the H.264/AVC intraframe encoder [5]. Each block with inter mode will be encoded using our hash-based low-complexity video block encoder described in Sec. III-B. For each block with skip mode, only the coding mode information is encoded. At the decoder, each block with inter mode is decoded via our hash-based video block decoder described in Sec. III-B. Each block with intra mode is decoded using the H.264/AVC intraframe decoder while each block with skip mode is decoded by copying and pasting the co-located block from the previous decoded frame.

TABLE I
AN EXAMPLE OF THE STRUCTURE OF OUR LOW-COMPLEXITY MULTIVIEW VIDEO CODEC WITH THREE ADJACENT VSNS.

VSN/ Time instant	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$	•••
V_1	$K_{1,t}$	$K_{1,t+1}$	$W_{1,t+2}$	$K_{1,t+3}$	$W_{1,t+4}$	$K_{1,t+5}$	$W_{1,t+6}$	$K_{1,t+7}$	•••
V_2	$K_{2,t}$	$W_{2,t+1}$	$K_{2,t+2}$	$W_{2,t+3}$	$K_{2,t+4}$	$W_{2,t+5}$	$K_{2,t+6}$	$W_{2,t+7}$	•••
V_3	$K_{3,t}$	$K_{3,t+1}$	$W_{3,t+2}$	$K_{3,t+3}$	$W_{3,t+4}$	$K_{3,t+5}$	$W_{3,t+6}$	$K_{3,t+7}$	•••

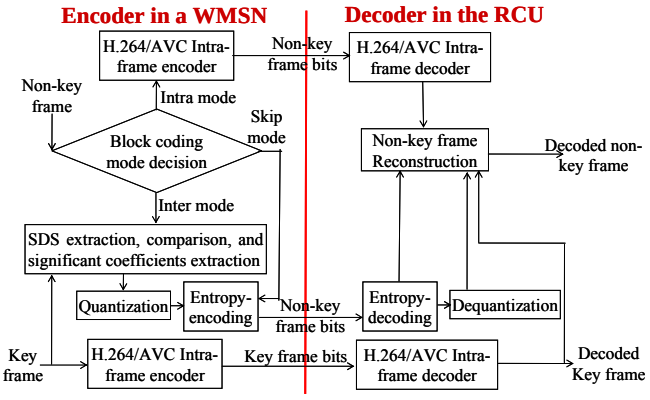

Fig. 2. Block diagrams of our low-complexity single-view video coding scheme.

B. Block-based Multiview Video Codec

To achieve better coding efficiency by extending our video codec from single-view to multiview, for each non-key frame, the multi-reference frames from the same VSN and the adjacent VSNs are jointly exploited. However, the frames from adjacent VSNs may be captured from different viewpoints, and hence, they should be transformed to the same viewpoint. The global disparities among them can be represented by global motion models. Here, the affine transform model is exploited, which has been successfully employed in traditional multiview video coding at the encoder [7] and recent low-complexity multiview video coding at the decoder [22]-[24] to exploit interview correlation. Consider a frame $W_{j,t}$ captured by V_j at time t and one of its reference frames $K_{i,t}$ from V_i adjacent to V_j. In the affine transformation model, each pixel (x, y) in $K_{i,t}$ can be mapped to the pixel (x', y') in $W_{j,t}$ via the estimated global motion parameters. However, global motion estimation (GME) is too complex to be performed in a VSN, and should be shifted to the decoder at the RCU. It is recalled that the RCU at receiver side can usually support powerful video decoding capability. The GME between the pair of the first intra-decoded key frames captured at the same time instant from adjacent VSNs is performed at the decoder. That is, the GME is performed only once for each pair of adjacent VSNs at the decoder. Then, the estimated global motion parameters are transmitted back to the corresponding VSNs via a feedback channel for processing subsequent frames until the configuration of the WMSN is significantly changed or the significant target scene change occurs. Note that the availability of a feedback channel is a common assumption of most recent low-complexity video encoding approaches [18], [21]-[23].

Our multiview video codec can be illustrated by an example shown in Fig. 3. For encoding $W_{j,t}$ ($j = 1$ and $t = 45$), its nearest key frame, $R_{j,t}$ ($R_{j,t} = R_{1,45} = K_{1,44}$) is determined to be its first reference frame. Similar to our single-view video encoder, 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 $B'_{j,t,b}$ (the first reference block in the first reference frame of $W_{j,t}$) in $R_{j,t}$ (step (a)). The coding mode decision and the resource scalability of our video encoder will be described in Sec. V. For each block $B_{j,t,b}$ with inter mode, the respective media hashes (described in Sec. III-A) for $B_{j,t,b}$ and $B'_{j,t,b}$ are extracted and compared (step (b)) to extract the initial significant SDS symbols for $B_{j,t,b}$ (step (c)), which will be compared with the co-located symbols in its second reference block as follows. Without allowing uncompressed frame exchanged between VSNs, V_j will send a message containing each initial significant SDS symbol of $W_{j,t}$ to its adjacent VSN V_i to announce it needs the second reference block for each of its blocks with inter mode. V_i will warp $K_{i,t}$ to the same viewpoint of $W_{j,t}$ ($i = 0, j = 1$, and $t = 45$) to get $K'_{i,t}$ (the second reference frame of $W_{j,t}$). Then, each initial significant SDS symbol of $W_{j,t}$ will be compared with the co-located symbol of $K'_{i,t}$ (step (d)) to determine the true significant SDS symbols (step (e)), whose parent node positions will be sent back to V_j. Finally, the DWT coefficients corresponding to each true significant SDS symbol are encoded via our hash-based video block encoder described in Sec. III-B.

Additional auxiliary information, including the coding mode information and the bitmap for each block with inter mode indicating which reference block (the first or second) should be referred for each significant DWT coefficient, can be efficiently encoded via the run-length and entropy encoding techniques. In this paper, the block size (*e.g.*, 128×128) is relatively large compared to the frame size, *i.e.*, the number of blocks in a frame is relatively small. Hence, the auxiliary information will be not a significant overhead. On the other hand, each block with intra mode is encoded using the H.264/AVC intraframe encoder while for each block with skip

mode only the coding mode information is encoded. At the decoder, each block is decoded according to its coding mode via the procedure similar to our single-view video decoder described in Sec. IV-A.

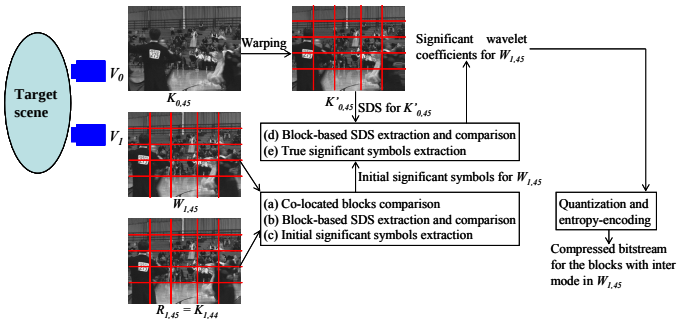

Fig. 3. An example for our low-complexity multiview video coding scheme for non-key frames.

V. Proposed PRD Optimized Multiview Video Coding

In this section, a power-rate-distortion (PRD) model for optimal resource allocation and performance optimization of the proposed low-complexity multiview video encoder described in Sec. IV-B is proposed. As mentioned in Sec. IV, the block coding mode is determined based on PRD optimization according to the available resources. Since block coding mode is related to the RD performance, it is important to characterize the relationship between the available resources and the RD performance. The major objective is to optimize the resource allocation and video coding performance while maximizing the lifetime for a VSN under current resource constraints.

Based on [4], [13]-[14], to analyze and control the power consumption of a VSN, a CMOS circuit design technology, 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.

A. Block Coding Mode Decision

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 decided to be encoded with intra, inter, and skip modes, respectively, in a non-key frame, where $N_{Intra} + N_{Inter} + N_{Skip} = N_b$. Let $\{B_i, i = 1, 2, \ldots, N_{Intra}\}$, $\{B_i, i = N_{Intra} + 1, N_{Intra} + 2, \ldots, N_{Intra} + N_{Inter}\}$, and $\{B_i, i = N_{Intra} + N_{Inter} + 1, N_{Intra} + N_{Inter} + 2, \ldots, N_{Intra} + N_{Inter} + N_{Skip} (= N_b)\}$ denote the sets of blocks with intra, inter, and skip modes, respectively, where the sorted block list, B_i, $i = 1, 2, \ldots, N_b$, is sorted by the SAD between each B_i and its reference block. 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 according to the current resources is equivalent to the coding mode decision for each block, which can be achieved via our PRD optimized resource allocation described in Secs. V-B and V-C.

B. Power-Rate-Distortion Model

In the proposed PRD model, our non-key frame 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 is 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, $0 \le C_1, C_2, C_3 \le 1$, respectively, where C_1, C_2, and C_3 can be estimated via averaging respective execution time obtained from several simulations, followed by being normalized to [0, 1]. 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:

$$F \times (C_1 \times X + C_2 \times Y + C_3 \times R) \le \Phi(P), \tag{4}$$

where F is the normalized frame rate, $0 \le F \le 1$, and $\Phi(P)$, $0 \le \Phi(P) \le 1$, is the normalized power consumption for the encoding power P transformed by the power function $\Phi(\bullet)$ under the assumption that DVS is employed [4], [13]-[14]. To optimally decide the coding mode for each block according to the current available resources (P and R), an RD function for non-key frame encoding should be derived and minimized, which can be derived as follows.

The classic RD function can be expressed as [13]-[14]:

$$D = \min_{R_i} \frac{1}{N_b} \sum_{i=1}^{N_b} \left(\sigma_i^2 \cdot 2^{-2\gamma R_i} \right), \text{ s.t. } \frac{1}{N_b} \sum_{i=1}^{N_b} R_i = R, \tag{5}$$

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. Based on the Lagrangian multiplier technique, the minimum distortion obtained by the optimal bit allocation can be expressed as:

$$D = \left(\prod_{i=1}^{N_b} \sigma_i^2 \right)^{\frac{1}{N_b}} \cdot 2^{-2\gamma R}. \tag{6}$$

Based on Eq. (6), obviously, the RD function for a block with intra mode can be expressed as:

$$D_{Intra} = \left(\prod_{i=1}^{N_{Intra}} \sigma_{i,Intra}^2 \right)^{\frac{1}{N_{Intra}}} \cdot 2^{-2\gamma \frac{N_b}{N_{Intra} + N_{Inter}} R}, \tag{7}$$

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

On the other hand, a block with inter mode includes some significant DWT coefficients (corresponding to the significant SDS symbols) being entropy-encoded, and the remaining insignificant DWT 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:

$$D_{Inter} = \left(\prod_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}} \sigma^2_{i,Inter} \right)^{\frac{1}{N_{Inter}}} \cdot 2^{-2\gamma \frac{N_b}{N_{Intra}+N_{Inter}} R} + \frac{1}{N_{Inter}} \sum_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}} \delta^2_{i,Inter}, \quad (8)$$

where $\sigma^2_{i,Inter}$ is the variance of the pixels corresponding to the significant DWT coefficients in the ith block with inter mode, while $\delta^2_{i,Inter}$ is the mean squared error (MSE) between the pixels corresponding to the insignificant DWT coefficients and the corresponding pixels in its reference block. Note that in the block coding mode decision process, for a block 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 media hash data exchanges between VSNs.

In addition, for a block with skip mode, the RD function is simply the MSE (denoted by $\delta^2_{i,Skip}$) between the block and its reference block as:

$$D_{Skip} = \frac{1}{N_{Skip}} \sum_{i=N_{Intra}+N_{Inter}+1}^{N_b} \delta^2_{i,Skip}. \quad (9)$$

C. Power-Rate-Distortion Optimization

Based on Eqs. (7)-(9), the overall RD function of a block in our multiview video codec can be expressed as:

$$\begin{aligned} D_{Overall} &= (1/N_b)\times(N_{Intra}\times D_{Intra} + N_{Inter}\times D_{Inter} + N_{Skip}\times D_{Skip}) \\ &= X\times D_{Intra} + Y\times D_{Inter} + Z\times D_{Skip}. \end{aligned} \quad (10)$$

To minimize $D_{Overall}$ based on optimally selected X, Y, and Z, where $Z = 1 - X - Y$, under the constraint shown in Eq. (4), $D_{Overall}$ should be translated into a function of X and Y, described as follows.

C.1. Model parameter γ estimation

First, based on Eqs. (7)-(10), 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:

$$\gamma = \frac{-\left(N_{Intra\_t} + N_{Inter\_t}\right)}{2N_b R} \log_2 \left[\frac{D_t - \frac{Y_t}{N_{Inter\_t}} \sum_{i=N_{Intra\_t}+1}^{N_{Intra\_t}+N_{Inter\_t}} \delta^2_{i,Inter} - \frac{Z_t}{N_{Skip\_t}} \sum_{i=N_{Intra\_t}+N_{Inter\_t}+1}^{N_b} \delta^2_{i,Skip}}{X_t \left(\prod_{i=1}^{N_{Intra\_t}} \sigma^2_{i,Intra} \right)^{\frac{1}{N_{Intra\_t}}} + Y_t \left(\prod_{i=N_{Intra\_t}+1}^{N_{Intra\_t}+N_{Inter\_t}} \sigma^2_{i,Inter} \right)^{\frac{1}{N_{Inter\_t}}}} \right] \quad (11)$$

Then, the updated γ can be used for the PRD optimization of the next non-key frame, and γ can be similarly updated iteratively. Several offline estimated γ's can be averaged to get the parameter γ for a certain scene in a period. The parameter γ related to encoding efficiency should be adaptively updated according to the available resources, past frame complexities, and past resource allocation configurations for current frame.

C.2. RD function for blocks with intra mode

Second, the function D_{Intra} defined in Eq. (7) can be converted to a continuous-time function. Usually, only a small number of blocks in a non-key frame are with intra mode (N_{Intra} should be small). It can be observed from the curve "Actual" in Fig. 4(a) that, in a non-key frame, the first few blocks in the decreasingly sorted list of motion activities of blocks usually have larger variances, and these variances will decrease as the motion activities decrease. Hence, it is reasonable to model $\sigma^2_{i,Intra}$ as a decreasing linear function:

$$G(t) = A\cdot(1-t),\ A>0,\ 0\le t\le 1,\ t = i/N_b,\ 1\le i\le N_{Intra}. \quad (12)$$

From Fig. 4(a), when the block index $i < 5$ ($X < 0.25$) in the total 20 blocks, the function $G(t)$ is accurate enough to model $\sigma^2_{i,Intra}$.

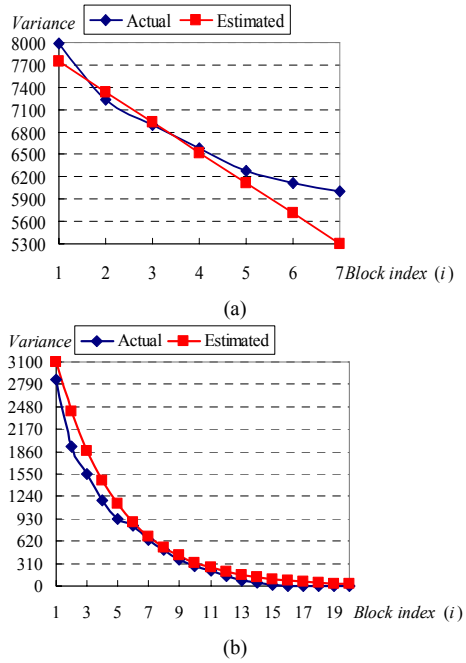

Fig. 4. (a) The curve "Actual" shows the variances of the first few blocks in the decreasingly sorted list of motion activities of blocks for the *Ballroom* and *Exit* sequences. All the variances for the same block index in each non-key frame are averaged. The curve "Estimated" shows the linear function $G(t)$; (b) the curve "Actual" shows the variances of the significant pixels in the blocks in the decreasingly sorted list of motion activities of blocks for the two sequences. The curve "Estimated" shows the function $H(t)$.

Due to X is usually much smaller than 0.25, using $G(t)$ to model $\sigma^2_{i,Intra}$ is reasonable. The parameter A in Eq. (12) can be derived from the previous PRD optimization result as follows. Assume N_{Intra_pre} denotes the number of blocks with intra mode obtained

from the previous coding mode decision. Hence, in the current non-key frame, A can be estimated from

$$(1/N_b)\sum_{i=1}^{N_{Intra\_pre}}\sigma_{i,Intra}^2 = \sum_{t=1/N_b}^{N_{Intra\_pre}/N_b}G(t) = \int_0^{N_{Intra\_pre}/N_b}A(1-t)dt, \quad (13)$$

as: $$A = 2N_b\sum_{i=1}^{N_{Intra\_pre}}\sigma_{i,Intra}^2 \Big/ \left(2N_bN_{Intra\_pre} - N_{Intra\_pre}^2\right). \quad (14)$$

To get the continuous-time version of D_{Intra}, we let

$$S = \left(\prod_{i=1}^{N_{Intra}}\sigma_{i,Intra}^2\right)^{\frac{1}{N_{Intra}}}, \quad (15)$$

and obtain

$$\ln S = \frac{1}{N_{Intra}}\sum_{i=1}^{N_{Intra}}\left(\ln\sigma_{i,Intra}^2\right). \quad (16)$$

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

$$\ln S = \left(N_b/N_{Intra}\right)\sum_{t=1/N_b}^{N_{Intra}/N_b}\ln G(t) = \frac{1}{X}\int_0^X \ln\left[A(1-t)\right]dt. \quad (17)$$

By applying the Taylor expansion to Eq. (17), S can be derived as:

$$S = A\cdot e^{-1-\frac{1}{X}(1-X)\ln(1-X)} \approx A\times(1-0.5\times X),\ 0\le X\le 1. \quad (18)$$

Hence, based on Eqs. (7) and (12)-(18), D_{Intra} can be derived as:

$$D_{Intra}(X,Y,R) = A(1-0.5X)\cdot 2^{-2\gamma\frac{R}{X+Y}}. \quad (19)$$

C.3. RD function for blocks with inter mode

Third, D_{Inter} in Eq. (8) can be expressed as $D_{Inter}(X, Y, R_{Inter})$ described as follows. Usually, the variance of the significant pixels corresponding to the significant DWT coefficients for a block with inter mode will decrease as the motion activity decreases. Based on Fig. 4(b), it is reasonable to model $\sigma_{i,Inter}^2$ as a decreasing exponential function as:

$$H(t) = B_1e^{-B_2t},\ B_1>0,\ B_2>0,\ 0\le t\le 1,\ t = i/N_b,$$
$$N_{Intra}+1\le i\le N_{Intra}+N_{Inter}. \quad (20)$$

The parameter B_1 in Eq. (20) can be derived from the previous PRD optimization result as follows. Assume N_{Intra_pre} and N_{Inter_pre} denote the numbers of blocks with intra mode and inter mode, respectively, obtained from the previous coding mode decision. Hence, in the current non-key frame, B_1 can be estimated from

$$(1/N_b)\sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}}\sigma_{i,Inter}^2 = \sum_{t=(N_{Intra\_pre}+1)/N_b}^{(N_{Intra\_pre}+N_{Inter\_pre})/N_b}H(t) = \int_{(N_{Intra\_pre}+1)/N_b}^{(N_{Intra\_pre}+N_{Inter\_pre})/N_b}B_1e^{-B_2t}dt \quad (21)$$

as:

$$B_1 = \left(B_2/N_b\right)\sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}}\sigma_{i,Inter}^2 \Big/ \left(e^{-B_2((N_{Intra\_pre}+1)/N_b)} - e^{-B_2((N_{Intra\_pre}+N_{Inter\_pre})/N_b)}\right), \quad (22)$$

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. To get the continuous-time version of the first term of Eq. (8), we let

$$T = \left(\prod_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}}\sigma_{i,Inter}^2\right)^{\frac{1}{N_{Inter}}}. \quad (23)$$

By considering the continuous-time version of ln$T$, we get:

$$\ln T = \left(N_b/N_{Inter}\right)\sum_{t=(N_{Intra}+1)/N_b}^{(N_{Intra}+N_{Inter})/N_b}\ln H(t) = \frac{1}{Y}\int_X^{X+Y}\ln\left(B_1e^{-B_2t}\right)dt. \quad (24)$$

Then, T can be derived as

$$T = B_1\times e^{-B_2(X+Y/2)}. \quad (25)$$

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

$$T\approx B_1\times h(X,Y),\ 0\le X,\ Y\le 1 \text{ and } X+Y\le 1, \quad (26)$$

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

$$h_1(X) = \left(0.5B_2^2e^{-0.3B_2}\right)X^2 - B_2e^{-0.3B_2}\left(1+0.3B_2\right)X + e^{-0.3B_2}\left(0.045B_2^2+0.3B_2+1\right), \quad (27)$$

$$h_2(Y) = \left(0.125B_2^2e^{-0.2B_2}\right)Y^2 - B_2e^{-0.2B_2}\left(0.5+0.1B_2\right)Y + e^{-0.2B_2}\left(0.02B_2^2+0.2B_2+1\right). \quad (28)$$

On the other hand, as the motion activity decreases, the MSE of the pixels corresponding to the insignificant DWT coefficients for a block with inter mode will be also decreased. Based on Fig. 5(a), it is reasonable to model $\delta_{i,Inter}^2$ as a decreasing linear function:

$$I(t) = C\cdot(1-t),\ C>0,\ 0\le t\le 1,\ t = i/N_b,$$
$$N_{Intra}+1\le i\le N_{Intra}+N_{Iner}. \quad (29)$$

It can be observed from Fig. 5(a) that when the block index $i \ge 16$ in the total 20 blocks, the function $I(t)$ is somewhat inaccurate in modeling $\delta_{i,Inter}^2$. However, the latter few blocks in the decreasingly sorted list are usually with skip mode; hence, using $I(t)$ to model $\delta_{i,Inter}^2$ is reasonable. The parameter C in Eq. (29) can be derived from the previous PRD optimization result as follows. Assume N_{Intra_pre} and N_{Inter_pre} denote the numbers of blocks with intra mode and inter mode, respectively, obtained from the previous coding mode decision. Hence, in the current non-key frame, C can be estimated from

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

as:

$$C = \frac{2N_b\sum_{i=N_{Intra\_pre}+1}^{N_{Intra\_pre}+N_{Inter\_pre}}\delta_{i,Inter}^2}{2N_bN_{Inter\_pre} - 2N_b + 2N_{Intra\_pre} - 2N_{Intra\_pre}N_{Inter\_pre} - N_{Inter\_pre}^2 + 1}. \quad (30)$$

By considering the continuous-time version of the second term of Eq. (8), we can get:

$$\frac{1}{N_{Inter}}\sum_{i=N_{Intra}+1}^{N_{Intra}+N_{Inter}}\delta_{i,Inter}^2 = \frac{N_b}{N_{Inter}}\sum_{t=\frac{N_{Intra}+1}{N_b}}^{\frac{N_{Intra}+N_{Inter}}{N_b}}I(t) = \frac{1}{Y}\int_X^{X+Y}C(1-t)dt = C(1-X-0.5Y). \quad (31)$$

Hence, based on Eqs. (8), and (20)-(31), D_{Inter} can be derived as:

$$D_{Inter}(X,Y,R_{Inter}) = B_1h(X,Y)\cdot 2^{-2\gamma\frac{R}{X+Y}} + C(1-X-0.5Y). \quad (32)$$

C.4. RD function for blocks with skip mode

Finally, D_{Skip} in Eq. (9) can be derived as follows. By considering the inverse order of the decreasingly sorted list of motion activities of blocks, as the motion activity increases, the MSE of a block with skip mode will be increased as shown in the "Actual" curve of Fig. 5(b). Hence, it is reasonable to model $\delta_{i,Skip}^2$ as an increasing exponential function as:

$$K(t) = D_1e^{D_2t},\ D_1>0,\ D_2>0,\ 0\le t\le 1,$$
$$t = i/N_b,\ 1\le i\le N_{Skip}. \quad (33)$$

It can be observed from Fig. 5(b) that when the inverse block index $i \geq 14$ in the total 20 blocks, the function $K(t)$ is somewhat inaccurate in modeling $\delta^2_{i,Skip}$. However, the latter few blocks in an inverse decreasingly sorted list of motion activities of blocks are usually with intra or inter modes, hence using $K(t)$ to model $\delta^2_{i,Skip}$ is reasonable. The parameter D_1 in Eq. (33) can be derived from the previous PRD optimization result as follows. Assume N_{Skip_pre} denotes the number of blocks with skip mode obtained from the previous coding mode decision. Hence, in the current non-key frame, D_1 can be estimated from

$$(1/N_b)\sum_{i=N_b-N_{Skip\_pre}+1}^{N_b}\delta^2_{i,Skip} = \int_0^{N_{Skip\_pre}/N_b} K(t)dt = \int_0^{N_{Skip\_pre}/N_b} D_1 e^{D_2 t}dt, \quad (34)$$

as:

$$D_1 = D_2 \sum_{i=N_b-N_{Skip\_pre}+1}^{N_b}\delta^2_{i,Skip} \Big/ N_b\left(e^{D_2(N_{Skip\_pre}/N_b)}-1\right), \quad (35)$$

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.

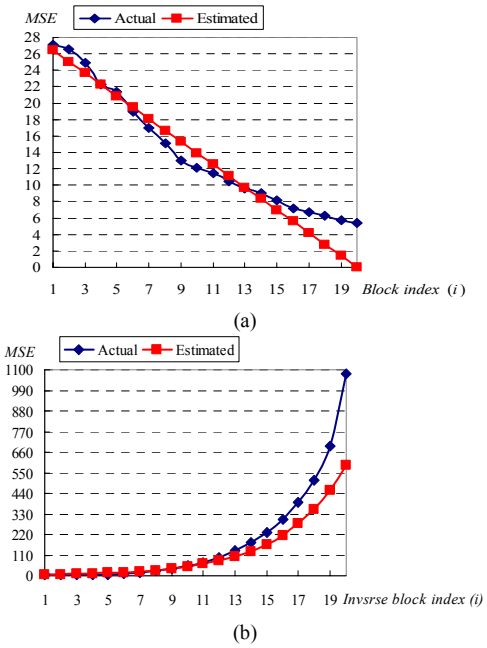

Fig. 5. (a) The curve "Actual" shows the MSEs of the pixels corresponding to the insignificant DWT coefficients in the decreasingly sorted list of motion activities of blocks for the *Ballroom* and *Exit* sequences. All the MSEs for the same block index in each non-key frame are averaged. The curve "Estimated" shows the function $I(t)$; (b) the curve "Actual" shows the MSEs of the blocks in the inversed order of the decreasingly sorted list of motion activities of blocks for the two sequences. The curve "Estimated" shows the function $K(t)$.

By approximating $\delta^2_{i,Skip}$ in Eq. (9) using Eqs. (33)-(35) and transferring Eq. (9) into a continuous form, we have

$$D_{Skip} = \frac{1}{N_{Skip}}\sum_{i=N_{Intra}+N_{Inter}+1}^{N_b}\delta^2_{i,Skip} = \frac{N_b}{N_{Skip}}\sum_{t=\frac{1}{N_b}}^{\frac{N_{Skip}}{N_b}}K(t) = \frac{1}{Z}\int_0^Z D_1 e^{D_2 t}dt = \frac{D_1}{ZD_2}\left(e^{D_2 Z}-1\right) \quad (36)$$

By applying the Taylor expansion, Eq. (36) can be approximated as:

$$D_{Skip} = (D_1/ZD_2)(k(Z)-1), \quad (37)$$

where

$$k(Z) = \left(0.5D_2^2 e^{0.3D_2}\right)Z^2 + D_2 e^{0.3D_2}(1-0.3D_2)Z + e^{0.3D_2}\left(1-0.3D_2+0.045D_2^2\right). \quad (38)$$

Hence, D_{Skip} can be derived as:

$$D_{Skip}(X,Y) = \frac{D_1}{(1-X-Y)D_2}\left[k(1-X-Y)-1\right]. \quad (39)$$

C.5. Optimization of the overall RD function

In summary, the overall distortion function can be derived based on Eqs. (10), (19), (32), and (39) as:

$$\begin{aligned} D_{Overall}(X,Y,R) &= X\times D_{Intra}(X,Y,R) + Y\times D_{Inter}(X,Y,R) + (1-X-Y)\times D_{Skip}(X,Y) \\ &= AX(1-0.5X)\cdot 2^{-2\gamma\frac{R}{X+Y}} + B_1 Yh(X,Y)\cdot 2^{-2\gamma\frac{R}{X+Y}} + CY(1-X-0.5Y) + \frac{D_1}{D_2}\left[k(1-X-Y)-1\right] \\ &= \left(AX-0.5AX^2\right)\cdot 2^{-2\gamma\frac{R}{X+Y}} + P(X,Y)\cdot 2^{-2\gamma\frac{R}{X+Y}} + Q(X,Y), \end{aligned} \quad (40)$$

where

$$P(X,Y) = B_1 Yh(X,Y), \quad (41)$$

$$Q(X,Y) = CY(1-X-0.5Y) + \frac{D_1}{D_2}\left[k(1-X-Y)-1\right] \quad (42)$$

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

$$\begin{aligned} &\min_{\{X,Y\}} D_{Overall}(X,Y,R) \\ &= \min_{\{X,Y\}}\left\{A\left(X-0.5X^2\right)\cdot 2^{-2\gamma\frac{R}{X+Y}} + P(X,Y)\cdot 2^{-2\gamma\frac{R}{X+Y}} + Q(X,Y)\right\}, \\ &\text{s.t. } F(C_1X + C_2Y + C_3R) \leq \Phi(P). \end{aligned} \quad (43)$$

Based on the proposed PRD model, before encoding a non-key frame, the parameters $\{X, Y, Z\}$ can be efficiently solved based on the current available power P and the target bit rate R to minimize the overall distortion $D_{Overall}(X, Y, R)$ defined in Eq. (43). 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. (43), 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. Although there are still some exponential terms of X and Y in Eq. (43), further simplification of Eq. (43) will be investigated in our future work.

Here, discrete sampling on X and Y is used to achieve efficient implementation for solving Eq. (43). Specifically, only a few points, $(X, Y) = (0.05x, 0.05y)$, $x, y = 0, 1, 2, \ldots, 19$, under the constraints, $0 \leq X, Y, X+Y \leq 1$, and $F(C_1X + C_2Y + C_3R) \leq \Phi(P)$, are evaluated to find the optimal point (X, Y) in minimizing Eq. (43). The average optimal parameter sets, $\{X, Y, Z\}$, for the *Ballroom* sequence, minimizing Eq. (43) with different combinations of the available encoding power P and the bit rates R, are shown in Fig. 6, where the parameters, X, Y,

and Z, respectively, of all the non-key frames are averaged. The analytic and actual PRD performances for the *Ballroom* sequence are shown in Fig. 7, where the MSEs of all the non-key frames are averaged.

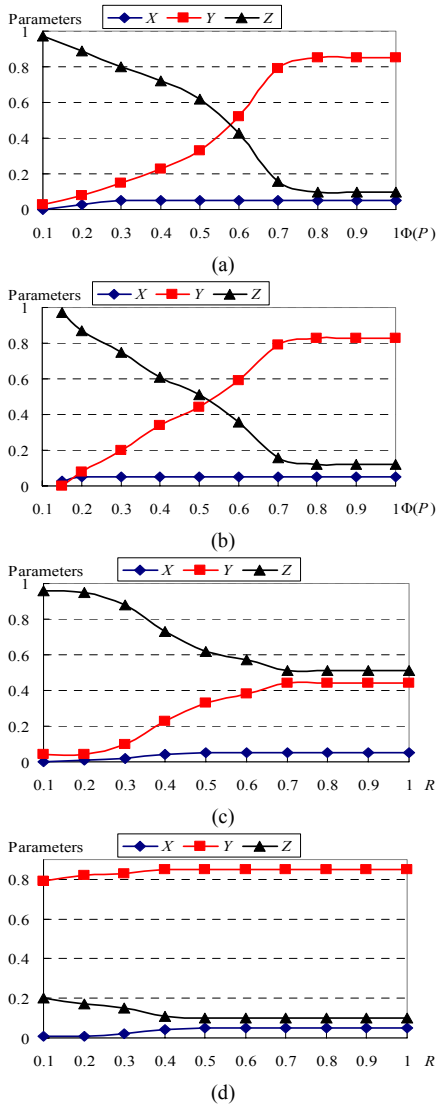

Fig. 6. The optimal parameter sets, $\{X, Y, Z\}$, for the *Ballroom* sequence, minimizing Eq. (43) with $\Phi(P)$ ranged from 0.1 to 1.0, and R fixed to: (a) 0.5bpp; (b) 1.0bpp; and with $\Phi(P)$ fixed to: (c) 0.5; and (d) 1.0, and R ranged from 0.1 to 1.0 bpp.

It can be observed from Fig. 6(a) and (b) that, under a fixed bit rate, when the power increases, Y (the percentage of the blocks with inter mode) will increase accordingly. In addition, X (the percentage of the blocks with intra mode) 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. 6(c) and (d) 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, Y will be much larger than X and Z (the percentage of the blocks with skip mode). To evaluate the accuracy of the proposed PRD model for non-key frame encoding, 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. 7. It can be observed from Fig. 7 that under a fixed bit rate, when the power increases, the distortion will decrease. In addition, 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. 7 that the proposed PRD model is fairly accurate to estimate the actual PRD performance.

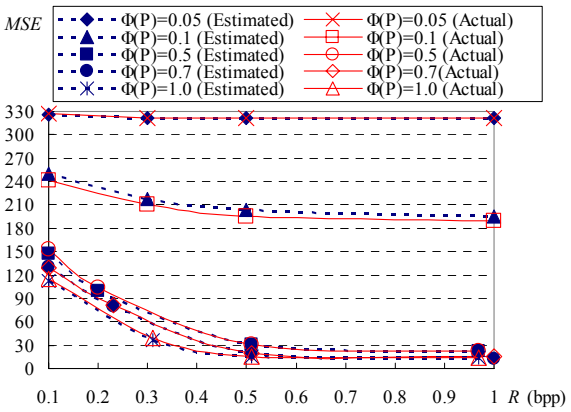

Fig. 7. The analytic and actual PRD performances for the *Ballroom* sequence. The curves "Estimated" show the PRD performance obtained from the proposed PRD model, whereas the curves "Actual" show the actual PRD performance obtained from our multiview video codec.

VI. Simulation Results

The two representative multiview video sequences, *Ballroom* (with large motions) and *Exit* (with medium or low motions) sequences [32] consisting of 250 frames, a frame size of 640×480, a block size of 128×128 (n = 128), YUV4:0:0 (only luminance component was evaluated), and a frame rate of 10 frames per second (fps) were used to evaluate our low-complexity multiview video codec under different available resources (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 are, the longer the hash length is. The quantization parameter (QP) for each H.264/AVC intra-encoded key frames ranged from 28 to 36. The H.264/AVC JM14.2 software is employed.

In the evaluated WMSN, the second, third, and fourth views (VSNs), *i.e.*, V_1, V_2, and V_3, from the total eight views (V_0~V_7), structured based on Table 1, where GOP size is 2, were employed, where the distance between each pair of VSN is 19.5 cm [32]. The structure shown in Table 1 was also similarly employed in [24]. To evaluate only three adjacent VSNs were also conducted in [22]-[24].

The four low-complexity video encoding schemes, including our low-complexity single-view video encoder (Proposed Single) [25], the H.264/AVC intraframe encoder (H.264 Intra) [5], the H.264/AVC interframe encoder with no motion (where all the motion vectors are set to zeros) with GOP size set to 2

(H.264 No motion (GOP = 2)), and the H.264/AVC interframe encoder with no motion with GOP size set to infinity (H.264 No motion (GOP = ∞)) were employed for comparisons with our low-complexity multiview video encoder. The two H.264/AVC-based low-complexity video encoders (H.264 Intra and H.264 No motion (GOP = 2)) were also used for comparisons in [18]. It should be noted that the studies of resource-scalable low-complexity multiview video encoding have rarely appeared in the literature. Hence, only some baseline non-resource-scalable low-complexity video encoders were selected for comparison with our encoding scheme. In this paper, the two metrics, *i.e.*, PRD performance and encoding complexity were used for performance evaluation and comparison.

A. Power-Rate-Distortion Performance

The average PRD performances for the three adjacent VSNs of our multiview video codec and the RD performances for the four schemes used for comparisons are shown in Figs. 8-9, respectively, for the *Ballroom* and *Exit* sequences, where the PSNR (peak signal to noise ratio) values of all the luminance frames from the three VSNs are averaged.

It should be noted that the RD performances of our video encoder shown in Figs. 8-9 don't take the rate used for interview hash data exchange during encoding blocks with inter mode into account. That is, the interview hash data exchange will consume a little power for wireless transmission of a VSN, but will not contribute the rates for final compressed video data.

For the *Ballroom* sequence, we have the following observations from Fig. 8. The PSNR performance gains of our multiview video codec at $\Phi(P) = 1.0$ above those of the H.264 No motion (GOP = 2) are from 2 to 4 dB. The PSNR performance gains of our multiview video codec at $\Phi(P) = 1.0$ above those of the H.264 Intra are from 4 to 5 dB. The RD performance of our multiview video codec at $\Phi(P) = 0.5$ is somewhat close to that at $\Phi(P) = 1.0$. The RD performance of our multiview video codec at $\Phi(P) = 0.1$ is still close to that of the H.264 Intra. The PSNR performance gains of our multiview video codec at higher powers can significantly outperform our single-view video codec. Similar results can also be observed from Fig. 9 for the *Exit* sequence.

More specifically, based on Figs. 8-9, our multiview video codec ($\Phi(P) = 1.0$ or $\Phi(P) = 0.5$) can outperform the three schemes used for comparisons (except H.264 No motion (GOP = ∞)), especially at high power and low bit rates. That is, when the power is high, our multiview video 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 adjacent views, our multiview video 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 performances of our multiview video codec ($\Phi(P) = 0.5$) can still significantly outperform the H.264 Intra, but is very close to those of the H.264 No motion (GOP = 2). 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. 7, 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. 6(c) and (d), where the configurations of *X*, *Y*, and *Z* will be unchanged while the bit rates are greatly increased. On the other hand, when the power is low, the RD performance of our multiview video codec is poor and the RD curves are flatter, which mean the bit rates cannot be efficiently exploited. It can be observed from Figs. 6-9 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 significantly change the block coding mode decision results.

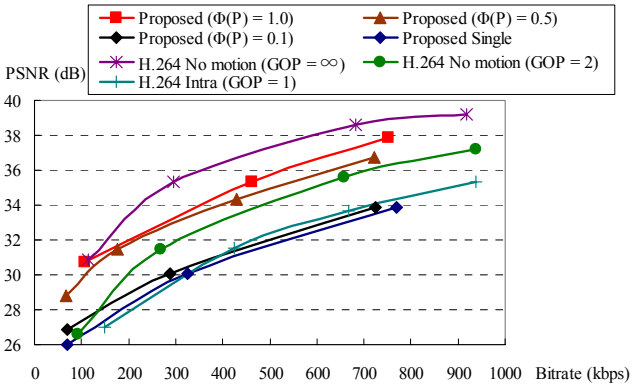

Fig. 8. The PRD performance for the *Ballroom* sequence.

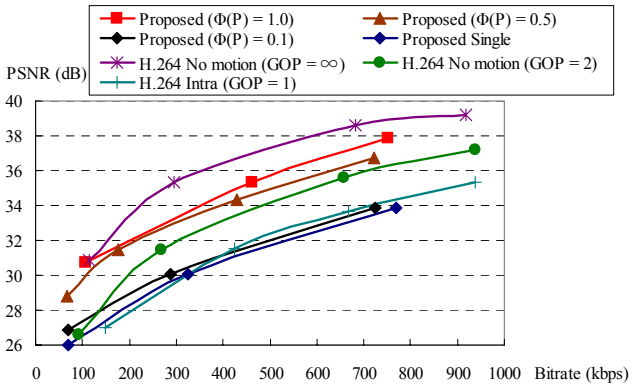

Fig. 9. The PRD performance for the *Exit* sequence.

For the three H.264/AVC-based low-complexity video encoders used for comparison, the H.264 Intra encoder has been shown to be low-complexity and efficient, which can outperform or be comparable to several current single-view or multiview low-complexity video encoders [18], [23]. Our multiview video encoder exploits interview correlation at the encoder via a few interview hash data exchanges and can, therefore, outperform the H.264 Intra encoder. The H.264 No motion (GOP=2) encoder has also been shown to be

low-complexity and very efficient, which is difficult to be defeated [18], [23]. Our multiview video encoder can significantly outperform the H.264 No motion (GOP = 2) at the low bit rates, which is a benefit for WMSN applications with severely limited transmission bit rates. The H.264 No motion (GOP = ∞) encoder has the best RD performance and has not been used as a benchmark in the existing low-complexity video encoding researches [18], [21]-[24]. However, it should be recalled again that the major goal of this paper is to propose a resource-scalable low-complexity video encoder for a WMSN and a PRD model for resource allocation and performance optimization of our encoder, instead of competing the coding performance against existing standard or non-resource-scalable video encoding schemes. In addition, the resource-scalability characteristic and the proposed PRD model are worthy for most low-complexity video encoding applications.

B. *Encoding Complexity*

Although it is claimed that the proposed multiview video encoder and the existing encoders used for comparisons are all with low-complexity, it is still important to compare their encoding complexities. The simplest way to estimate the encoding complexity for a video encoder is to measure its encoding time [18]. The respective average encoding time per frame (of size, 640×480) for the *Ballroom* sequence of our multiview video encoder, the H.264 Intra encoder, the H.264 No motion (GOP = 2), and H.264 No motion (GOP = ∞), measured on a Pentium-4 PC with 3.40GHz CPU and 1.49GB RAM at different bit rates is shown in Fig. 10. The encoding time of our multiview video encoder includes the time consumed in interview hash data exchanges, where the hash data size is relatively small (*e.g.*, average 3.78 kbits, *i.e.*, about 0.16% of the original frame size, per frame), the time consumed in PRD optimization, and the time consumed in the remaining video coding tasks. By considering the typical data transmission rate, 40 kbps, for a common sensor node [1] (actually, the rate may be higher for a VSN, *e.g.*, 250 kbps [1] or 1 Mbps [19]), it takes only 0.09 seconds to achieve interview hash data exchange per frame. Fig. 10 shows that the encoding complexity (in second) of our multiview video encoder is lower than those of the three H.264/AVC-based low-complexity video encoders, even though when the full power, *i.e.*, Φ(*P*) = 1.0, is applied.

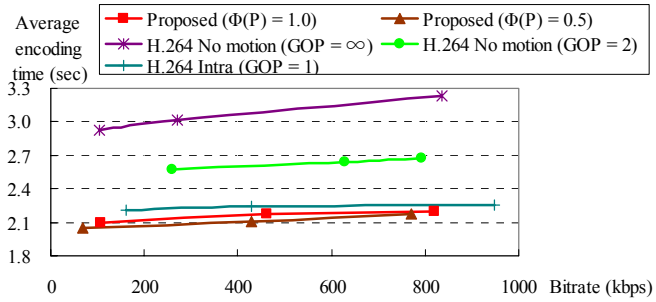

Fig. 10. The average encoding time per frame for the *Ballroom* sequence.

VII. Conclusions and Future Works

In this paper, we have proposed a resource-scalable media hash-based low-complexity multiview video encoder and a PRD model to characterize the relationship between the available resources and the RD performance of our video encoder. 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.

For future work, the distortion induced by wireless video transmission (*e.g.*, packet loss) will be integrated into the current distortion function to form a complete end-to-end 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 derived to provide a practical guideline in preparation and deployment for a WMSN.

VIII. Acknowledgements

This work was supported in part by National Science Council, Taiwan, ROC, under Grants NSC 95-2422-H-001- 031 and NSC 97-2628-E-001-011-MY3.

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] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "Wireless multimedia sensor networks: applications and testbeds," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1588-1605, Oct. 2008.

[3] K. T. Phan, R. Fan, H. Jiang, S. A. Vorobyov, and C. Tellambura, "Network lifetime maximization with node admission in wireless multimedia sensor networks," accepted and to appear in *IEEE Trans. on Vehicular Technology*.

[4] Z. He, W. Cheng, and X. Chen, "Energy minimization of portable video communication devices based on power-rate-distortion optimization," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 596-608, May 2008.

[5] 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.

[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] 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.

[8] H. Kim, N. Kamaci, and Y. Altunbasak, "Low-complexity rate-distortion optimal macroblock mode selection and motion estimation for MPEG-like video coders," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 823–834, July 2005.

[9] D. S. Turaga, M. van der Schaar, and B. Pesquet-Popescu, "Complexity scalable motion compensated wavelet video encoding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 8, pp. 982–993, Aug. 2005.

[10] P. C. Tseng, Y. C. Chang, Y. W. Huang, H. C. Fang, C. T. Huang, and L. G. Chen, "Advances in hardware architectures for image and video coding—A survey," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 184–197, Jan. 2005.

[11] M. van der Schaar and Y. Andreopoulos, "Rate-distortion-complexity modeling for network and receiver aware adaptation," *IEEE Trans. on*

Multimedia, vol. 7, no. 3, pp. 471–479, June 2005.

[12] W. Yuan, K. Nahrstedt, S. V. Adve, D. L. Jones, and R. H. Kravets, "Grace-1: Cross-layer adaptation for multimedia quality and battery energy," *IEEE Trans. on Mobile Computing*, vol. 50, no. 7, pp. 799–815, July 2006.

[13] 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.

[14] 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.

[15] W. Wang, D. Peng, H. Wang, H. Sharif, and H. H. Chen, "Energy-constrained distortion reduction optimization for wavelet-based coded image transmission in wireless sensor networks," *IEEE Trans. on Multimedia*, vol. 10, no. 6, pp. 1169–1180, Oct. 2008.

[16] W. C. Feng, E. Kaiser, W. C. Feng, and M. L. Baillif, "Panoptes: scalable low-power video sensor networking technologies," *ACM Trans. on Multimedia, Computing, Communications and Applications*, vol. 1, no. 2, pp. 151-167, May 2005.

[17] H. Wu and A. A. Abouzeid, "Energy efficient distributed image compression in resource-constrained multihop wireless networks," *Computer Communications*, vol. 28, 1658-1668, 2005.

[18] C. Brites, J. Ascenso, J. Q. Pedro, and F. Pereira, "Evaluating a feedback channel based transform domain Wyner-Ziv video codec," *Signal Processing: Image Communication*, vol. 23, pp. 269-297, 2008.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] X. Guo, Y. Lu, F. Wu, D. Zhao, and W. Gao, "Wyner-Ziv-based multi-view video coding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 6, pp. 713-724, June 2008.

[25] 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.

[26] 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).

[27] 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.

[28] L. W. Kang and C. S. Lu, "Power-rate-distortion model for low-complexity video coding," in *Proc. of Picture Coding Symposium*, Chicago, Illinois, USA, May, 2009.

[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] 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.

[31] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36-58, Sept. 2001.

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