ABSTRACT

To meet the requirements of resource-limited video sensors, low-complexity video encoding technique is highly desired. In this paper, we propose a low-complexity power-scalable multi-view distributed video encoding scheme by using the correlations among video frames from adjacent video sensor nodes via robust media hashing extracted at encoder and using the global motion parameters estimated and fed back from the decoder. In addition, the proposed video encoding scheme is power-scalable, which is adaptive based on the available power supply of the video sensor. The power-rate-distortion behavior of the proposed video encoding scheme is also analyzed in order to maximize the video quality under limited video sensor resource allocation. The theoretical achievable minimum distortion (AMD) of reconstructed video under a total power supply constraint for a video sensor is also derived. Based on the AMD estimation, a guideline is provided to decide the power supply for each video sensor based on desired video quality or acceptable distortion before deploying the wireless video sensor network.

Index Terms—Low-complexity video coding, multi-view distributed video coding, power-scalable video coding, wireless video sensor networks, power-rate-distortion analysis.

1. INTRODUCTION

With the availability of low-cost hardware, such as CMOS cameras, wireless video sensor networks (WVSNs) are potential to promote several emerging applications, such as security monitoring, emergency response, and environmental tracking [1]-[2]. In a WVSN shown in Fig. 1, some video sensor nodes (VSNs) are usually scattered in a sensor field. Each

VSN equipped with a camera can capture and encode visual information, and deliver 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 decoding and further processing [2]. However, each VSN operates under several resource constraints including lower computational capability, limited power supply, and narrow transmission bandwidth. Hence, in a WVSN, video compression and transmission are the two major concerns for a VSN. However, to achieve higher coding efficiency, current video compression approaches usually perform complex encoding operations (e.g., motion estimation for exploiting temporal correlation of successive frames in the same view or disparity estimation for exploiting inter-view correlation) [3]-[6], which cannot be applicable for a VSN.
the decoder while preserving a certain coding efficiency. DVC can consider either a single VSN (single view) [7]-[15] or adjacent VSNs together (multi view) [16]-[21]. The major characteristic of multi-view DVC is that inter-VSN communications can be avoided during encoding to save energy.

Another recent popular approach is to collaborate video coding and transmission [22]-[23]. First, each frame from a VSN is intra-encoded using a still image encoder or intra-frame video encoder (e.g., JPEG-2000 or H.264/AVC intra-frame encoder [3]-[4]). While transmitting the encoded frames from adjacent VSNs through the same intermediate node, the node will perform image matching procedure to detect the similar/overlapping regions for the frames. Then, the overlapping regions will be encoded only once to further compress the frames. The major challenge here is how to efficiently and accurately identify the overlapping regions between two frames in an intermediate VSN under resource-limited constraints.

On the other hand, for a resource-limited VSN, optimal resource allocation for maximizing decoded video quality is highly desirable [2], [24]-[25]. First, it is needed to design a power-scalable video encoder for a VSN, which can adjust its computational complexity and power consumption under power constraint. In [2], [24], a parametric complexity-scalable video encoder is developed by adjusting the three encoding parameters based on encoding complexity constraint. In addition, by using a popular CMOS circuits design technology, called dynamic voltage scaling (DVS), the power scalability is equivalent to the complexity scalability. Then, an analytic power-rate-distortion (PRD) model is developed to characterize the inherent relationship between the power consumption of their power-scalable video encoder and its rate-distortion performance. Based on the PRD analysis, the optimal power allocation between video encoding and wireless data transmission can be achieved.

In this paper, a low-complexity power-scalable multi-view distributed video encoding scheme, extended from our previous work [21], is proposed by exploiting the characteristics of the aforementioned two kinds of video encoding approaches with additional but limited inter-VSN communications. The PRD behavior of the proposed encoder is also analyzed in order to maximize the video quality under limited VSN resource allocation.

The remainder of this paper is organized as follows. Our robust media hashing technique [26] for inter-VSN communication during the encoding process is described in Sec. 2. The proposed low-complexity power-scalable multi-view video encoding scheme is described in Sec. 3. The PRD analysis for the proposed video encoding scheme is addressed in Sec. 4. Simulation results are presented in Sec. 5, followed by conclusions and future works in Sec. 6.

2. ROBUST MEDIA HASHING FOR INTER-VSN (VIDEO SENSOR NODE) COMMUNICATION

In the proposed video encoding scheme, to further reduce encoding bit-rate and reduce transmission power, limited media hash bits are allowed to be exchanged among adjacent video sensor nodes. Our robust media hashing scheme, called structural digital signature (SDS) [26], which can extract the most significant components and provide a compact representation of a frame (or an image block) efficiently, meets the requirement.

To exploit the SDS for an image block encoding and reconstruction, the problem can be formulated as follows. For an image block, $B$, its most significant components, extracted by comparing the SDS of $B$ and that of its reference block $B'$, should be properly selected such that

$$\begin{align*}
\text{PSNR}(B, \beta) & \geq \text{desired PSNR value}, \\
\text{PSNR}(B, \beta) & >> \text{PSNR}(B', \beta),
\end{align*}$$

where $\text{PSNR}$ denotes the peak signal to noise ratio (PSNR), and $\beta$ is an estimate of $B$, obtained by modifying $B'$ using the SDS of $B$.

To extract the SDS for an image block of size $n \times n$, a J-scale discrete wavelet transform (DWT) is performed. Let $w_{s,o}(x, y)$ represent a wavelet coefficient at scale $s$, orientation $o$, 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, $w_{s+1, o}(x, y)$, and its four child nodes, $w_{s+1}(2x+i, 2y+j)$, the maximum magnitude difference ($\text{max}_\text{mag_diff}$) is calculated as

$$\text{max}_\text{mag_diff}(x, y) = \max_{s, o, i, j} |w_{s+1, o}(x, y)| - |w_{s}(2x+i, 2y+j)|$$

Then, all the parent-4 children pairs will be arranged in the decreasing order based on their $\text{max}_\text{mag_diffs}$. The first $L$ ($L$ is denoted as hash length) pairs in the decreasing order are selected for constructing the SDS of the block.

Once the significant parent-4 children pairs are selected, each pair will be 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 $c$. 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, the 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, the relations (c) and (d) can be merged into another signature symbol “-1” when $p < 0$ and $c$ is ignored. That is, one should keep the sign of the larger node unchanged while ignoring the smaller one under the constraint that their original interscale relationship is still preserved. Similarly, the signature symbols “+2” and “-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 \(\text{max} \_ \text{mag} \_ \text{diff}\) in an image block, \(B\), the signature symbol \(\text{Sym}(B, p, c)\) can be defined as

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

(4)

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.” For an \(n \times n\) image block translated into a symbol sequence \((+1, −1, +2, −2, \text{ or } 0)\), 2 bits and \(\log_2(n \times n)\) bits are required to indicate each symbol and its parent node position. The other symbols without indication are “0” symbols. Totally, \(L \times (2 + \log_2(n \times n))\) bits are required to indicate the SDS with length \(L\) for an image block.

3. PROPOSED LOW-COMPLEXITY POWER-SCALABLE MULTI-VIEW VIDEO ENCODING SCHEME

Assume that there are \(N_{\text{VSN}} \geq 3\) adjacent video sensor nodes (VSNs) observing the same target scene in a wireless video sensor network (WVSN). For each VSN, \(V_s, s = 0, 1, 2, \ldots, N_{\text{VSN}} − 1\), a captured video sequence is divided into several groups of pictures ( GOPs) with \(\text{GOP}\) size, \(\text{GOP}_s\), in which a \(\text{GOP}\) consists of a key frame, \(K_s\), where \(t \mod \text{GOP}_s = 0\), followed by some non-key frames, \(W_s\), where \(t \mod \text{GOP}_s \neq 0\). An example of the \(\text{GOP}\) structure with \(N_{\text{VSN}} = 3\) is shown in Table 1. In the proposed video encoding scheme, shown in Fig. 2 (for two adjacent VSNs, \(V_i\) and \(V_j\), observing the same target scene), each key frame is encoded using the H.264/AVC intraframe encoder [4] or the proposed key frame re-encoding scheme while each non-key frame is encoded using the proposed non-key frame encoding scheme. In Fig. 2, the key frame, \(K_{ij}\), from \(V_j\) is re-encoded by treating the warped \(K_{ij}\) at the same time instant \(t\) from \(V_i\) as its reference frame while the non-key frame, \(W_{j,t}\) from \(V_j\) is encoded via inter-VSN communication by treating the warped \(K_{ij}\) at the same time instant \(t\) from \(V_i\) as its second reference frame. The detailed proposed video encoding scheme will be addressed in the following subsections.

3.1. Proposed Multi-View Key Frame Encoding Scheme

For each video sequence captured in a VSN, the first key frame is encoded using the H.264/AVC intraframe encoder [4] and transmitted to the decoder (RCU). For a pair of key frames captured at the same time instant and coming from adjacent VSNs, the difference (derived from different viewing angles) between them can be estimated via a global motion model [16], [18]-[19], [21]. However, the global motion estimation process is very complex and cannot be performed in a VSN. Hence, the global motion estimation task is performed for each pair of intra-decoded key frames from adjacent VSNs at the decoder. The estimated motion parameters will be transmitted back to the corresponding VSNs via a feedback channel for warping and encoding subsequent frames, as an example shown in Fig. 3.

Table 1. A simple example of the \(\text{GOP}\) structure for a \(\text{WVSN}\) with \(N_{\text{VSN}} = 3\), where \(\text{GOP}_1 = 1\), \(\text{GOP}_2 = 4\), and \(\text{GOP}_3 = 2\).

<table>
<thead>
<tr>
<th>VSN / Time instant</th>
<th>(t)</th>
<th>(t + 1)</th>
<th>(t + 2)</th>
<th>(t + 3)</th>
<th>(t + 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(V_0)</td>
<td>(K_0)</td>
<td>(W_0)</td>
<td>(K_0)</td>
<td>(W_0)</td>
<td>(K_0)</td>
</tr>
<tr>
<td>(V_1)</td>
<td>(K_1)</td>
<td>(W_1)</td>
<td>(K_0)</td>
<td>(W_0)</td>
<td>(K_0)</td>
</tr>
<tr>
<td>(V_2)</td>
<td>(K_2)</td>
<td>(W_2)</td>
<td>(K_1)</td>
<td>(W_1)</td>
<td>(K_0)</td>
</tr>
</tbody>
</table>

Fig. 2. A diagram for the proposed low-complexity multi-view distributed video encoding scheme.

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

For each VSN, its non-first key frame is also encoded using the H.264/AVC intraframe encoder first and then transmitted toward the RCU. For a pair of intra-encoded key frames, \(K_{ij}\) and \(K_{jk}\) (with size \(M \times N\)), from the adjacent VSNs, \(V_i\) and \(V_j\), at time instant \(t\), transmitted to the intermediate node \(V_k\), \(V_k\) will perform the proposed key frame re-encoding scheme (Fig. 4) to compress the key frames further. First, \(V_k\) will perform intra-decoding to obtain \(K_{ij}\) and \(K_{jk}\) respectively. For re-encoding \(K_{ij}',\) by treating \(K_{ij}\) as its reference frame, \(K_{ij}'\) will be warped to the viewing angle of \(V_f\) via the global motion parameters estimated by their previous key frame pair and fed back from the RCU to get \(K_{ij}\). Here, it is assumed that after a WVSN is completely deployed, each VSN is not allowed to change its location and viewing angle, and the \(\text{GOP}\) size should be not too large. Hence, the latest estimated global motion parameters preserve certain accuracy.
3.2. Proposed Multi-View Non-Key Frame Encoding Scheme

For non-key frame encoding (Fig. 5), a hash-based multi-reference encoding scheme is proposed. For encoding a non-key frame \( W_{j,t} \), its nearest key frame, \( R_{j,t} \), (e.g., \( R_{j,t} = K_{j,t} \) or \( K_{j,t+1} \)) from the same VSN, \( V_j \), is determined to be its “first” reference frame. Similar to the key frame re-encoding, each block in \( W_{j,t} \) is compared with the co-located block in \( R_{j,t} \) by calculating their MSE (step (a) in Fig. 5). If the MSE is smaller than the threshold, \( T_{N} \), the block will be skipped. For each non-skipped block, \( B_{j,t,b} \), in \( W_{j,t} \), its SDS with length, \( L_{j,t,b} \), is compared with that of the co-located block in \( R_{j,t} \) (step (b) in Fig. 5) to extract the “initial” significant symbols (step (c) in Fig. 5). Let the number of the initial significant symbols for \( B_{j,t,b} \) be \( L_{j,t,b,init} \), \( L_{j,t,b,init} < L_{j,t,b} \). Then, the initial significant symbols will be compared with the co-located symbols in the co-located block in the “second” reference frame determined as follows.

While encoding \( W_{j,t}, V_j \) will send a message containing the parent node position for each initial significant symbol in all the non-skip blocks in \( W_{j,t} \) to its adjacent VSN, \( V_i \) (encoding the key frame, \( K_i \), at the same time instant \( t \)) to announce it needs the second reference frame. Then, \( V_i \) will warp \( K_i \), to the viewing angle of \( V_j \) via the global motion parameters estimated by their previous key frame pair to get \( K_i \). The SDS symbols for each corresponding block, \( B_{i,t,b} \), in \( K_i \), serving for the second reference frame for \( W_{j,t} \), will be transmitted to \( V_j \). Here, each SDS symbol for \( B_{i,t,b} \) corresponding to \( B_{j,t,b} \) with the parent node position of the corresponding initial significant symbol. Hence, the hash length for \( B_{i,t,b} \) is \( L_{j,t,b,init} \), and the transmitted SDS data size is \( 3 \times L_{j,t,b,init} \) bits (five possible symbols, +1, -1, +2, -2, or 0 for each). Note that the parent node position for each symbol in \( B_{i,t,b} \) is decided by the corresponding initial significant symbol in \( B_{j,t,b} \) and needn’t be transmitted. Usually, the hash length, \( L_{j,t,b,init} \), for block, \( B_{j,t,b} \), in the second reference frame is relatively short to eliminate the inter-VSN communication overhead. However, the amounts of hash data that can be transmitted from \( V_i \) to \( V_j \) highly depend on the available data transmission power for \( V_j \), and data reception power for \( V_i \), parts of which will be addressed in Sec. 4.

After receiving the SDS for \( K_i \), from \( V_i \), for each non-skip block in \( W_{j,t} \), the initial significant symbols will be compared with the co-located symbols in \( K_i \), (step (d) in Fig. 5) to extract the “true” significant symbols for \( W_{j,t} \) (step (e) in Fig. 5). That is, some initial significant symbols can be filtered out by being compared with the corresponding symbols in the second reference frame with a little auxiliary information to indicate that the symbols are predicted by the second reference frame. Usually, most symbols corresponding to the background region in \( W_{j,t} \) can be filtered out by comparing with the first reference frame (\( R_{j,t} \) from the same VSN) while part of the symbols corresponding to the foreground (moving objects) can be filtered out by comparing with the second reference frame (\( K_i \) from the adjacent VSN). Finally, all the coefficients corresponding to the true significant symbols are quantized and entropy-encoded to form the bitstream for \( W_{j,t} \). The corresponding decoding process can be easily done at the decoder.
3.3. Power-Scalability of the Proposed Video Encoding Scheme

For a battery-powered VSN, it is essential to adjust the encoding operations based on the available power supply to maximize the power efficiency and video quality. Similar to [2], [24], to analyze and control the power consumption of a VSN, a CMOS circuits design technology for a mobile device, called dynamic voltage scaling (DVS) is assumed to design the VSNs in this study. It is claimed [2], [24] 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.

Similar to the concept of the parametric power-scalable video encoder developed in [2], [24], the proposed video encoder is scalable in computational complexity and power consumption, achieved by adjusting the three parameters, $T_{\text{K}}$, $T_{\text{K2}}$, and $T_{\text{N}}$, which control the numbers of skipped blocks in the key frame re-encoding process and non-key frame encoding process, respectively. The larger the three parameters, the larger the numbers of skipped blocks are, i.e., the lower the computational complexity and power consumption are. To maximize the rate-distortion performance under the power constraint, the encoder should perform an optimal computational power allocation based on the power-rate-distortion model analyzed in the following section.

4. POWER-RATE-DISTORTION ANALYSIS

The performance of the proposed video encoder can be evaluated by the quality of video transmitted from a video sensor node (VSN) to the AFN. The video quality evaluated by the end-to-end distortion can be defined as:

$$ D = D_{\text{c}} + D_{\text{t}}, $$

where $D_{\text{c}}$ and $D_{\text{t}}$ are the distortions caused by lossy video compression and transmission errors, respectively. $D_{\text{c}}$ and $D_{\text{t}}$ can be given by

$$ D_{\text{c}} = \text{MSE}(I, \hat{I}) \quad \text{and} \quad D_{\text{t}} = \text{MSE}(\hat{I}, \tilde{I}), $$

where $I$, $\hat{I}$, and $\tilde{I}$ are the original frame, error-free decoded version of $I$, and reconstructed version at the AFN of $I$, respectively.

For a VSN, let $P_s$, $P_r$, $P_r$, and $P_0$ denote the power used for video encoding, data transmission, data reception, and the power budget, respectively, where $P_s + P_r + P_r \leq P_0$. By integrating the characteristic of the proposed video encoder and [2], the $D_{\text{c}}$ for non-key frame is a function of $P_s$, $P_r$, and coding bitrate $R_s$, while $D_{\text{t}}$ is a function of $P_r$ and transmission distance $d$. Hence, the objective function for maximizing the video quality for non-key frame under the power constraint can be formulated as

$$ \min_{\{P_r, P_s, P_r\}} D = D_{\text{c}}(R_s, P_s, P_r) + D_{\text{t}}(P_r, d) $$

s.t. $P_s + P_r + P_r \leq P_0$.

In addition, $D_{\text{c}}$ for non-key frame, can be derived as

$$ D_{\text{c}}(R_s, P_s, P_r) = \sigma^2 e^{-\gamma(R_h)} g(P_r) g(P_s) R_h, $$

where

$$ g(P_r) = P_r^{1/3}. $$

$P_r(R_h)$ is the power function used for receiving the hash (SDS) data with bit rate $R_h$ from the adjacent VSN for encoding a non-key frame, where

$$ P_r(R_h) = \alpha R_h, $$

and $\alpha = 135$ nJ/bit [27]. The RD curves for different encoding power consumption levels and power for receiving hash levels derived from Eq. (8) and the actual measurements for the Ballroom sequence [28] are shown in Fig. 6. It should be noted that Fig. 6 only represents the power-rate-distortion (PRD) behavior for a number of non-key frames, exclusive of key frames, because the encoding behavior for key frame is different from that of non-key frame. From Fig. 6, it can be seen that while the encoding power and power for receiving hash decrease, $D_{\text{c}}$ becomes flatter, i.e., the video coding efficiency is reduced. That is, while the encoder doesn’t have enough encoding power and enough hash data from adjacent VSN, more available bitrates cannot be efficiently exploited. Based on the actual measurements in Fig. 6, the $D_{\text{c}}$ function in Eq. (8) is fairly accurate. It is noted that in Fig. 6, the $P_r$ and $P_r$ for each curve are approximately adjusted to be the same levels, instead of the same values. The cases that $P_r$ and $P_r$ with very different levels simultaneously are not considered. It is reasonable to assume that if a VSN can receive large hash data from its adjacent VSN, it should consume higher encoding power to sufficiently exploit these hash data and perform the encoding operations.

It should be noted that in the current PRD analysis, only the proposed non-key frame video encoding scheme is considered. That is, only the three tasks, namely, (i) video encoding, (ii) reception of the hash data from adjacent VSN, and (iii) transmission of the compressed video bitstream outside the VSN, are
Finally, we can obtain
\[
P_s = \frac{2}{5} \left(P_0 - \alpha \cdot R_H \right).
\]
interframe coding with default setting (full search motion estimation with 5 reference frames, and only one intraframe (I frame), followed by several interframes (P frames)) [4], the H.264/AVC intraframe coding [4], our previous single-view low-complexity video codec (denoted by Single) [15], and our previous single-reference frame multi-view low-complexity video codec without key frame re-encoding (denoted by Multi) [21] were employed for comparisons with the proposed video coding scheme. The latter three approaches used for comparisons and the proposed scheme here are all with low complexity. The rate-distortion (RD) performances for \( V_1 \) (the second view) are shown in Figs. 9-10, respectively, for the Ballroom and Exit video sequences. Note that Figs. 9-10 only show the RD performances for the proposed video encoding scheme without considering power consumption issue.

It can be observed from Figs. 9-10 that the PSNR performance gains of the proposed scheme outperform those of the three low-complexity approaches, especially in larger motion sequence (e.g., Ballroom). For the Ballroom sequence with large motions, the performance gaps among these approaches are more significant. That is, in the proposed video encoding scheme, some moving regions in a non-key frame can be efficiently encoded by the assistance from its second reference frame from adjacent VSN. However, the single-view [15] and single-reference frame [21] approaches for comparisons cannot well utilize the advantages from multi-view and multi-reference frames. On the other hand, for the Exit sequence with relatively small motions, the performance gaps among these approaches are somewhat small. The performance gap between the proposed scheme and the H.264/AVC interframe coding is about 1 dB while the computational complexity of the proposed scheme is much lower than that of the H.264/AVC interframe coding.

6. CONCLUSIONS AND FUTURE WORKS

In this paper, a low-complexity power-scalable multi-view distributed video encoding scheme for wireless video sensor networks is proposed. Based on the global motion parameters estimated at the decoder and fed back to the encoder, the key frames from different video sensor nodes (VSNs) can be re-encoded further. Based on the feed-backed global motion parameters and low-complexity hash-based inter-VSN communications, a non-key frame can be efficiently encoded based on two reference frames. In addition, the power-rate-distortion (PRD) behavior of the proposed video encoder is analyzed to maximize video quality under the power constraint. The achievable minimum distortion indicating the theoretically achievable minimum distortion or optimal video quality for a given power supply for a VSN is also derived. Based on the AMD estimation, a guideline is provided to decide the power supply for each VSN based on desired video quality or acceptable distortion before deploying the wireless video sensor network.

For the future researches, transmission errors should be considered so that more accurate but complex PRD behavior can be analyzed. In addition, error resilience encoding, error concealment, and secure transmission techniques for the proposed video encoding scheme should be also developed.

REFERENCES


