MULTI-VIEW DISTRIBUTED VIDEO CODING WITH LOW-COMPLEXITY INTER-SENSOR COMMUNICATION OVER WIRELESS VIDEO SENSOR NETWORKS+

*Li-Wei Kang and Chun-Shien Lu**

Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC
Email: {lwkang, lcs}@iis.sinica.edu.tw

ABSTRACT

To meet the requirements of resource-limited video sensors, low-complexity video encoding technique is highly desired. In this paper, a low-complexity multi-view distributed video encoding scheme by using the correlations among video frames from adjacent video sensor nodes (VSNs) via robust media hashing at encoder and the global motion parameters estimated and fed back from the decoder is proposed. The frames from adjacent VSNs are warped into the same view-direction based on the global motion parameters. Then, the significant differences between the warped key frame and the non-key frame from adjacent VSNs are efficiently extracted based on robust media hashing for non-key frame compression. The key is that few data (hash information) exchanges among adjacent VSNs are allowed to efficiently exploit the correlations among VSNs. The coding performance and energy consumption of the proposed encoder have been verified through simulations and comparisons with existing low-complexity video encoders.

Index Terms— Low-complexity video coding, multi-view distributed video coding, wireless video sensor networks

1. INTRODUCTION

In a wireless video sensor network (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) via the Internet or satellite for decoding and further processing [1]-[2]. Compared with traditional networks, WVSN must work under several resource constraints, such as each VSN is with lower computational capability, limited battery power supply, and narrow transmission bandwidth. Hence, in a WVSN, video compression and transmission are the major concerns for each VSN. However, to achieve higher coding efficiency, current video compression approaches usually perform complex encoding operations [3], which will consume a significant portion of the battery power in a VSN.

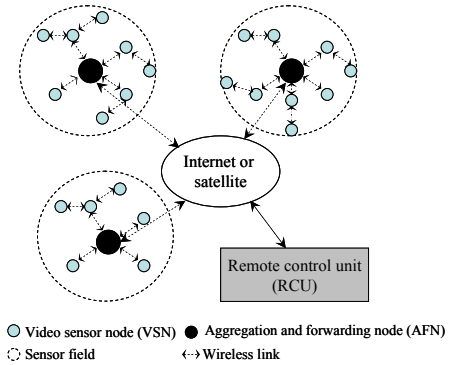

Fig. 1. A wireless video sensor network architecture.

To meet the requirements of resource-limited VSNs, low-complexity video coding techniques have been recently very popular. A famous one is the distributed video coding (DVC) approach, where individual frames are encoded independently, but decoded jointly [4]-[5]. The computational burden at encoder can be shifted to the decoder while preserving a certain coding efficiency. The DVC can consider either a single VSN (single view) [4]-[6] or adjacent VSNs together (multi view) [7]-[9]. The major characteristic of the multi-view DVC is that inter-VSN communications can be avoided during encoding to save energy. However, if additional but limited inter-VSN communications can be allowed during the encoding process, more inter-VSN correlations can be exploited to increase coding efficiency.

In this paper, a low-complexity multi-view distributed video encoding scheme is proposed. Similar to the current multi-view DVC approaches [7], [9], the global motion estimation tasks are shifted to the decoder. However, the proposed scheme will perform low-complexity jointly video encoding by exploiting low-complexity inter-VSN communications.

2. MEDIA HASH FOR INTER-VSN COMMUNICTION

+This work was supported in part by National Science Council, Taiwan, ROC, under Grants NSC95-2422-H-001-031 and NSC 95-2221-E-001-022.
*Corresponding author (lcs@iis.sinica.edu.tw)

To further reduce encoding bit-rate and transmission energy, limited media hash bits are allowed to be exchanged among adjacent VSNs. Our robust media hashing scheme, called structural digital signature (SDS) [10], which can extract the most significant components and provide a compact representation of a frame (or an image block) efficiently, meets the requirement.

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), $0 \le s < J$, $1 \le x \le n$, and $1 \le y \le n$. For each pair consisting of a parent node, $w_{s+1,o}(x, y)$, and its four child nodes, $w_{s,o}(2x + i, 2y + j)$, the maximum magnitude difference (*max_mag_diff*) is calculated as

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

Then, all the parent-4 children pairs will be arranged in the decreasing order based on their *max_mag_diff*s. The first L pairs in the decreasing order are selected for constructing the SDS of the block. For each selected pair of a parent node p and its child node c with *max_mag_diff*, the signature symbol $sym(p, c)$ can be defined as

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

That is, a block is 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."

3. PROPOSED LOW-COMPLEXITY MULTI-VIEW DISTRIBUTED VIDEO ENCODING

3.1. System Architecture

Assume that there are N_{VSN} (≥ 3) adjacent VSNs observing the same target scene in a WVSN. For each VSN, V_s, $s = 0, 1, 2, \ldots, N_{VSN} - 1$, a captured video sequence is divided into several group of pictures (GOPs) with GOP size, $GOPS_s$, in which a GOP consists of a key frame, $K_{s,t}$, where t mod $GOPS_s = 0$, followed by some non-key frames, $W_{s,t}$, where t mod $GOPS_s \neq 0$. An example of the GOP structure with $N_{VSN} = 3$ is shown in Table 1.

Table 1. A simple example of the GOP structure for a WVSN with $N_{sensor} = 3$, where $GOPS_0 = 1$, $GOPS_1 = 4$, and $GOPS_2 = 2$.

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

3.2. Proposed Multi-View DVC Encoder

Each key frame is encoded using the H.264/AVC intraframe encoder [3] to form the key frame bits. For each non-key frame partitioned into several non-overlapping blocks, each block can be encoded by referring the corresponding reference block. The determination of a reference block is described as follows.

3.2.1. Intra-VSN reference block determination

For each VSN, V_s, a non-key frame, $W_{s,t}$, will be partitioned into several non-overlapping blocks, $B_{s,t,b}$ (with size $n \times n$, and b is the block index), and the SDS for $B_{s,t,b}$, SDS($B_{s,t,b}$), will be extracted. Then, the reference block, $R_{s,t,b,intra}$, corresponding to $B_{s,t,b}$ from the same VSN, V_s, is determined to be the co-located block in its nearest key frame. For example, if the immediate previous frame of $W_{s,t}$ is a key frame, $K_{s,t-1}$, $R_{s,t,b,intra}$ is set to the block $B_{s,t-1,b}$ in $K_{s,t-1}$. First, if the mean square error (MSE), between $B_{s,t,b}$ and $R_{s,t,b,intra}$ is small enough, $R_{s,t,b,intra}$ is directly determined to be the final reference block for encoding $B_{s,t,b}$, or $B_{s,t,b}$ can be skipped. If all the blocks in $W_{s,t}$ are with small MSE, no inter-VSN communication is required during the encoding process. Usually, a block in background or static regions will select its reference block from the same VSN, or be skipped.

On the other hand, the similarity, $Sim(B_{s,t,b}, R_{s,t,b,intra})$, between $B_{s,t,b}$ and $R_{s,t,b,intra}$ will be calculated as [10]:

$$Sim(B, B') = Sim(\mathrm{SDS}(B), \mathrm{SDS}(B')) = (N^+ - N^-) / L, \quad (3)$$

where B and B' denote a block and its candidate reference block, respectively, SDS(B) and SDS(B') denote the SDS for B and B', respectively, N^+ denotes the number of signature symbols for B are the same as the corresponding symbols for B', N^- denotes the number of signature symbols for B are different from the corresponding symbols for B', and L denotes the SDS length. Then, another reference frame from adjacent VSN will be checked as follows.

3.2.2. Inter-VSN reference block determination

If inter-VSN communication is required to find possible candidate reference blocks, V_s will transmit the index for each block which may need a reference block from adjacent VSN to its nearest adjacent VSN. This is the first time of the inter-VSN communication during the encoding process with only few index data. It should be noted that the desired adjacent VSN must satisfy that the corresponding reference frame for $W_{s,t}$ is a key frame at the same time instant t.

For the nearest adjacent VSN, V_i, for V_s, its equipped camera may monitor the same target area via different view-direction. The difference induced by different view-direction between the frames from V_s and V_i can be represented by global motion models [7], [9]. In the perspective global motion model, each point (x_1, y_1) in the frame of V_i is mapped to the point (x_2, y_2) in the frame of V_s via the transformation:

$$x_2 = (m_0 x_1 + m_1 y_1 + m_2) / (m_6 x_1 + m_7 y_1 + 1), \quad (4)$$

$$y_2 = (m_3 x_1 + m_4 y_1 + m_5) / (m_6 x_1 + m_7 y_1 + 1), \quad (5)$$

where $m_0, m_1, \ldots, m_7$ are the global motion parameters needed to be estimated. However, the global motion estimation process is very complex and cannot be performed

in a resource-limited VSN. Here, the motion estimation task is shifted to the decoder as follows. For each key frame pair, $K_{s,t}$ and $K_{i,t}$, at the same time instant t from V_s and V_i, respectively, the global motion parameters between them can be estimated from the two corresponding decoded frames at the decoder, and transmitted back to the encoder side (V_s and V_i) via a feedback channel. For a non-key frame, $W_{s,t}$, from V_s and its possible reference frame, $K_{i,t}$, from V_i, the motion parameters between them can be approximated by the parameters estimated from their latest key frame pair. Here, it is assumed that after a WVSN is completely deployed, each VSN is not allowed to change its location and view-direction, and the GOP size should be not too large. Hence, the latest estimated global motion parameters preserve certain accuracy.

Then, V_i will warp the current key frame, $K_{i,t}$, to the view-direction of $W_{s,t}$ to get $R_{s,t,inter}$. For each block, $R_{s,t,b,inter}$, in $R_{s,t,inter}$, with the index b received from V_s, the SDS($R_{s,t,b,inter}$) will be transmitted to V_s. This is the second time of the inter-VSN communication during encoding. Then, $Sim(B_{s,t,b}, R_{s,t,b,inter})$ is calculated. From the two candidate reference blocks, $R_{s,t,b,intra}$ and $R_{s,t,b,inter}$, the one with larger similarity (based on Eq. (3)) will be selected as the final reference block, $R_{s,t,b}$ for $B_{s,t,b}$. Usually, a block in foreground or moving regions will select its reference block from adjacent VSN.

3.2.3. Non-key frame encoding and decoding

After determining the reference block, $R_{s,t,b}$ for $B_{s,t,b}$, similar to [6], $B_{s,t,b}$ will be encoded based on comparing SDS($B_{s,t,b}$) and SDS($R_{s,t,b}$) as follows. Each signature symbol in SDS($B_{s,t,b}$) is compared with the corresponding symbol in SDS($R_{s,t,b}$). If they are the same, the symbol in $B_{s,t,b}$ will be skipped. Otherwise, the corresponding parent-4 children pair of the symbol is determined to be significant. That is, we intend to efficiently extract the wavelet coefficients of $B_{s,t,b}$, that are significantly different from the corresponding ones of $R_{s,t,b}$. Finally, each significant parent-4 children pair will be quantized and entropy-encoded to form the non-key frame bits. The corresponding decoding process can be easily done at the decoder.

3.3. Analysis

In this section, the computational complexity and energy consumption for the proposed encoder can be analyzed as follows. The computational complexity of the proposed encoder mainly comes from the DWT, SDS extraction, SDS comparison, and entropy encoding. The heaviest task in the SDS extraction is the sorting operation, which can be efficiently performed using the quick sort algorithm. Hence, the computational complexity of the proposed encoder is dominated by that of the DWT. Without performing motion estimation, the computational complexity of the proposed encoder should be in the similar order of that of a conventional intraframe encoder, consisting of transformation, quantization, and entropy coding. In addition, the computational complexity of the proposed encoder should be similar to the DWT-based multi-view DVC [7].

On the other hand, the total energy consumption per bit (in terms of Joule) for a VSN can be defined as [11]:

$$E_{VSN\text{-}B} = E_{ENC\text{-}B} + E_{TX\text{-}B} + E_{RX\text{-}B}, \quad (6)$$

where $E_{ENC\text{-}B}$ denotes the energy consumed in encoding per bit, $E_{TX\text{-}B}$ denotes the energy consumed in transmission per bit, and $E_{RX\text{-}B}$ denotes the energy consumed in reception per bit. Compared with the conventional intraframe encoder, the proposed encoder has similar computational complexity. Hence, only the energy consumed in transmission for the proposed encoder and that for the conventional intraframe encoder are compared as follows. In the proposed encoder, the energy consumed in transmitting an encoded non-key frame can be defined as:

$$E_{Pro\text{-}F} = B_{Pro\text{-}F} \times E_{TX\text{-}B} + B_{index} \times E_{TX\text{-}B} + B_{SDS} \times E_{RX\text{-}B}, \quad (7)$$

where B_F denotes the number of bits required for representing an encoded non-key frame, B_{index} denotes the number of bits required for representing the block index data, transmitted to adjacent VSN, and B_{SDS} denotes the number of bits required for representing the SDS of a reference frame, received from adjacent VSN. For simplicity, B_{index} is relatively small and can be ignored, and $E_{RX\text{-}B}$ is assumed to be similar with $E_{TX\text{-}B}$ (in fact, $E_{RX\text{-}B}$ is smaller than $E_{TX\text{-}B}$ [11]). Hence, the energy consumption for transmitting a non-key frame is approximately $(B_{Pro\text{-}F} + B_{SDS}) \times E_{TX\text{-}B}$. On the other hand, the energy consumed in transmitting an intra-encoded frame is defined as:

$$E_{Intra\text{-}F} = B_{Intra\text{-}F} \times E_{TX\text{-}B}, \quad (8)$$

where $B_{Intra\text{-}F}$ denotes the number of bits for an intra-encoded frame. Finally, compared with the conventional intraframe encoder under the same reconstructed video quality, the energy saving for encoding a non-key frame using the proposed encoder can be approximately defined as:

$$ES_{Pro\text{-}F} = \left(1 - \frac{E_{Pro-F}}{E_{Intra-F}}\right) \times 100\% = \left(1 - \frac{B_{Pro-F} + B_{SDS}}{B_{Intra-F}}\right) \times 100\%. \quad (9)$$

Energy saving of the proposed encoder will be further verified through simulations.

4. SIMULATION RESULTS

Some multi-view video sequences [12] with three views, frame size, 640×480, block size, 128×128, YUV4:2:0, frame rate, 10 frames per second (fps), and different bitrates were used to evaluate the proposed video encoder. The three VSNs, V_0, V_1, and V_2, were structured based on Table 1. The H.264/AVC interframe coding, H.264/AVC intraframe coding [3], and our previous low-complexity single-view video encoder (denoted by Single) [6] were employed for comparisons with the proposed video encoder. The latter two approaches for comparisons and the proposed encoder are all with low complexity. The RD performances for V_1 (the second view) are shown in Figs. 2-3.

It can be observed from Fig. 2 that the PSNR performance gains of the proposed encoder above those of the Single are about *2* dB. The PSNR performance gains of the proposed encoder above those of the H.264/AVC intraframe coding range from *2* to *4* dB. The performance gaps between the proposed encoder and the H.264/AVC interframe coding range from *2* to *3* dB. On the other hand, it can be observed from Fig. 3 that the PSNR performance gains of the proposed encoder above those of the Single range from *1* to *2* dB. The PSNR performance gains of the proposed encoder above those of the H.264/AVC intraframe coding range from *1* to *4* dB. The performance gaps between the proposed encoder and the H.264/AVC interframe coding range from *1* to *2* dB. As a summary, it can be observed that the PSNR performance gains of the proposed encoder outperform those of the two low-complexity approaches, especially in larger motion sequence (*e.g.*, *Ballroom*).

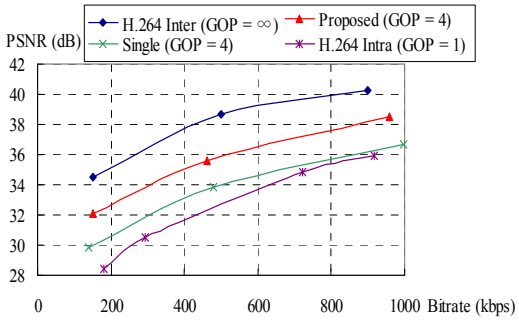

Fig. 2. RD comparison for the *Ballroom* sequence.

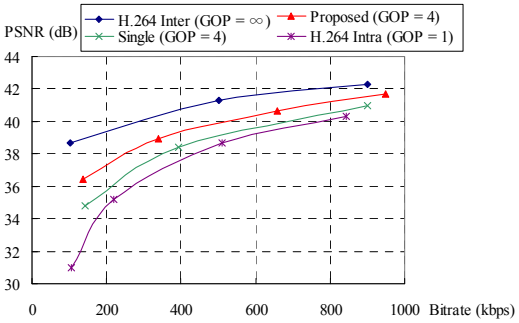

Fig. 3. RD comparison for the *Exit* sequence.

Compared with H.264/AVC intraframe coding having comparable computational complexity with the proposed encoder, the proposed encoder requires fewer bit-rates and less energy than those of the H.264/AVC intraframe encoder under similar video quality. For example, when PSNR is *36* dB for the *Ballroom* sequence shown in Fig. 2, the average bit-rates for each non-key frame obtained using the proposed encoder and each intra-encoded frame obtained using the H.264/AVC intraframe encoder are about *15* Kbits, and *90* Kbits, respectively. Based on Eq. (9), compared with the H.264/AVC intraframe encoder, the energy saving for encoding a non-key frame using the proposed encoder with SDS length, 512 (8 Kbits), is about $(1 - (15 + 8) / 90) \times 100\% \approx 74\%$.

5. CONCLUSIONS

In this paper, a low-complexity multi-view distributed video encoding scheme by using the correlations among video frames from adjacent VSNs via robust media hashing at encoder and the global motion parameters estimated and fed back from the decoder is proposed. While encoding a non-key frame, at most two times of the inter-VSN communications are required to exchange hash information of limited size so that the correlations among adjacent VSNs can be efficiently exploited. Simulations indicate that the proposed encoder consume less energy than the H.264/AVC intraframe encoder. For the future researches, multi-reference frames may be used to encode a non-key frame, simultaneously, and power-rate-distortion performance of the proposed encoder can be analyzed

6. REFERENCES

[1] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Computer Networks*, vol. 51, pp. 921-960, 2007.

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

[3] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Sys. Video Tech.*, vol. 13, pp. 560-576, 2003.

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

[5] R. Puri, A. Majumdar, P. Ishwar, and K. Ramchandran, "Distributed video coding in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, pp. 94-106, 2006.

[6] L. W. Kang and C. S. Lu, "Low-complexity Wyner-Ziv video coding based on robust media hashing," *Proc. of IEEE Int. Workshop on Multimedia Signal Process.*, 2006.

[7] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," *Proc. of SPIE VCIP*, vol. 6077, Jan. 2006.

[8] C. Yeo and K. Ramchandran, "Robust distributed multi-view video compression for wireless camera networks," *Proc. of SPIE VCIP*, vol. 6508, Jan. 2007.

[9] 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*, 2006.

[10] 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, 2003.

[11] H. Wu and A. A. Abouzeid, "Energy efficient distributed JPEG2000 image compression in multihop wireless networks," in *Proc. of IEEE Workshop on Applications and Services in Wireless Networks*, MA, USA, 2004, pp. 152-160.

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