Region Adaptive Measurements for Distributed Compressive Video Sensing

Hongyan Zhai

Guangdong Industry Polytechnic,Guangzhou 510300,China

teacherzhai@126.com

Keywords: Distributed Video Coding, Compressive Sensing, Standard Video Codec, Intra-coding

Abstract. Due to limited resources, wireless video sensor networks (WVSN) needs low complexity methods to realize video capture and compression. Recently, distributed video coding (DVC) has emerged to reduce the video encoding complexity with a certain coding efficiency. Furthermore, compressive sensing (CS) has been proposed to directly capture the compressed data efficiently. In this paper, combining DVC and CS, a novel video coding framework has been designed for WVSN. In order to preserve low complexity at the encoder, odd frames of the original video can be compressed as key frames by standard intra-coding while even frames can be processed as CS frames by CS encoder. Here, flexible mode selection of the encoder is applied to improve coding efficiency. Furthermore, adaptive measurements are allocated for different blocks in view of the object and background regions in the video frames. At the decoder, the bi-directional dictionary is proposed as the sparse basis to improve the recovery quality of CS. The experimental results validate the effectiveness of the proposed scheme with better performance than other compared schemes.

1. Introduction

Nowadays, wireless video sensor networks (WVSN) has emerged in many fields of video capture and processing with mobile devices [1]. However, the processing of huge video data has posed great challenges on the nodes of WVSN. To address the problems in WVSN, such as limited processing capacity and power, a novel framework of video compression should be designed with low computational complexity and high compression efficiency.

Many current video coding standards, such as MPEG-x and H.26x series, mainly adopt the hybrid coding frameworks combing block transform with motion estimation and compensation [2]. Although these video coding standards have high compression efficiency, the standard encoder has high computational complexity, especially in part of motion estimation, which will impact the execution speed of the whole system [3]. Therefore, these video coding standards are difficult to be applied in the sensor nodes with limited resources.

In the past years, distributed video coding (DVC) [4] based on the principle of distributed source coding (DSC) has been proposed to shift the major video encoding computation burden to the decoder. This is helpful for the WVSN. If DVC is introduced in WVSN, the sensor nodes can operate the video encoder with low complexity while the servers can perform the video decoder with high complexity [5]. Furthermore, in view of fast data acquisition, compressive sensing (CS) has drawn significant attention among industry and academic researchers [6]. In the conventional Shannon/Nyquist sampling theorem, it is claimed that when capturing a signal, one must sample at least two times faster than the signal bandwidth in order to avoid losing information. However, CS is a new method to capture and represent compressible signals at a rate significantly below the Nyquist rate. Due to low sampling rate, CS can avoid the big burden of data storage and processing at the sensor nodes of WVSN. Recently, some methods combining DVC and CS have been proposed. In [7], a distributed compressive video sensing (DCVS) framework is proposed to simultaneously capture and compress video data, where almost all computation burdens can be shifted to the decoder, resulting in a very low-complexity encoder. In [8], a

novel framework called distributed compressed video sensing (DISCOS) is proposed as a solution for distributed video coding based on the recently emerging compressed sensing theory. In [9], a new DVC based on CS principles is proposed. At the encoder, each CS frame (non-key frame) is divided into non-overlapping blocks and then sampled, quantized and transmitted. At the decoder, an approximation of the block is obtained as a linear combination of blocks from previously transmitted frames, using the received block measurements. Furthermore, in [10], human visual characteristics can be introduced into the video coding for better visual reconstructed quality.

Inspired by [10], a novel framework of video coding has been designed by combining DVC with CS according to the human visual characteristics. Firstly the original video sequence has been split into two video sub-sequences by odd and even means. Then the odd frames are regarded as key frames, which are compressed by standard intra-coding. And the even frames are considered as CS frames, which are compressed by CS encoder with very low complexity. Furthermore, adaptive measurements are allocated for different blocks according to the object and background regions of the video frames and flexible mode selection of the encoder is also applied to improve coding efficiency. At the decoder, in order to make good use of temporal correlation, the bi-directional dictionary is developed as the sparse basis to improve the recovery quality of CS.

The rest of this paper is organized as follows. In Section 2, the proposed scheme is presented step by step. In Section 3, the performance of the proposed scheme is examined. We conclude the paper in Section 4.

2. Proposed Scheme

2.1 Overview

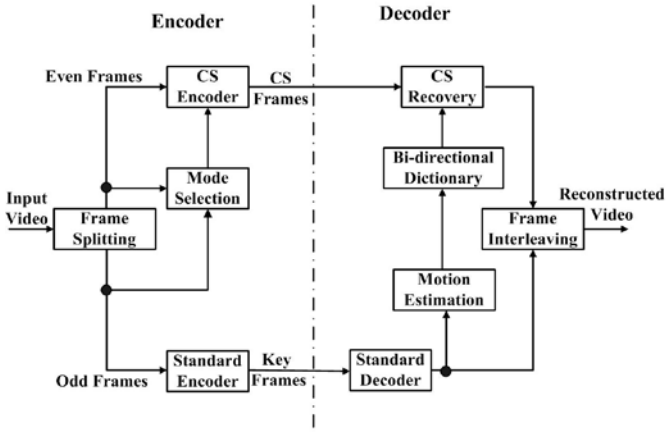


Fig. 1 Block diagram of the proposed scheme

Figure 1 illustrates the block diagram of the proposed scheme. At the encoder, the original video sequence will be split into odd and even frames firstly. The even frames will be processed using CS principles to produce CS frames while the odd frames will be standard encoded as key frames. It is noted that the CS encoder is performed at the block level. If $x \in R^n$ is a column vector organized from the current block and u is its coefficients in some orthonormal basis Ψ, then $x = \Psi^T u$. Using the CS encoder we can obtain $y = \Phi x$, where Φ is $m \times n$ matrix and $y \in R^m$. Since $m < n$, the original signal x can be compressed. And the generated y is called the measurement. At the CS decoder, u can be reconstructed by solving the following optimization problem.

$$\min \|u\|_1, \qquad \text{subject to } y = \Phi\Psi^T u \tag{1}$$

Then according to $x = \Psi^T u$, the original signal x can be obtained. From the process of CS, it is can be found that the CS encoder has low computational complexity.

Here, three coding modes of the block are designed, that is, SKIP mode, SINGLE mode and L1 mode, which have been improved compared with [9]. For the odd frames, the standard encoder can be used without any modifications. In view of low computational complexity of the encoder, the standard intra-coding may be good choice for the odd frames to generate the key frames. At the decoder, the key frames can be uncompressed by the standard decoder, which can be applied to construct the bi-directional dictionary for CS recovery. After CS recovery and standard decoding the video sub-sequences will be interleaved frame by frame for final reconstruction. Next, the important modules are explained as follows.

2.2 Mode Selection

As shown in Figure 1, the input video sequences will be split by odd and even frames firstly. And then in order to improve the rate-distortion performance, we design three coding modes for CS encoder, that is, SKIP mode, SINGLE mode and L1 mode.

In SKIP mode, the current coding block of even frames can be skipped at the CS encoder and the decoding is performed by copying the co-located block. Here, the two mean square error (MSE) values between the current block and its co-located blocks in its forward and backward key frames can be computed, which can be used to determine the SKIP mode. If one of the two MSE values is smaller than the threshold T_0, the block can be skipped and no measurements need to be transmitted. It is noted that the SKIP mode is improved compared with [9]. In [9] only one MSE value is calculated between the current block and its co-located block in its forward key frames while in this paper the two MSE values are computed using its co-located blocks in its forward and backward key frames. Here, the bi-directional MSE values can make better use of the temporal correlation in the original video sequences, which may lead to the performance improvement.

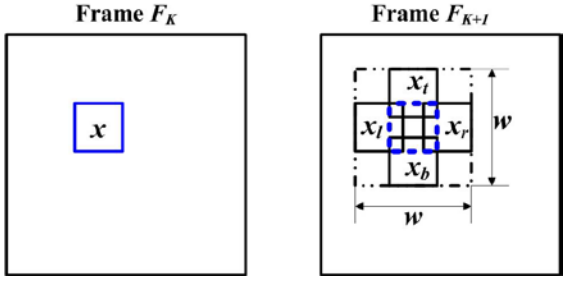


Fig. 2 Reference blocks in SINGLE mode

Here, we improved the SINGLE and L1 mode to avoid the feedback channel in [9]. In SINGLE mode, for the current block x in the frame F_K, four reference blocks x_t, x_b, x_l and x_r in F_{K+1} can be taken into account in the $w \times w$ square window, as shown in Figure 2. Then we can calculate the minimum MSE (MMSE) value between x and its four reference blocks. If the MMSE value is smaller than the threshold T_1, the block x can be encoded using M_S measurements. Otherwise, M_L ($M_L > M_S$) measurements are needed to encode the block x and this is called L1 mode. It is noted that the label with two bits are needed to distinguish the mode selected.

In SINGLE mode the decoder compares the received M_S measurements with the measurements generated by each block in the dictionary and selects the block with MMSE, which can decrease the complexity of the decoder. Furthermore, in L1 mode the decoder will perform optimized problem in (1) using M_L measurements. It is noted that M_S and M_L can be adaptively selected according to the object and background regions in the frames. The details of region adaptive measurements are introduced in sub-section C.

2.3 Region Adaptive Measurements

Fig. 3 Examples of the test video sequences with the object regions

As it is mentioned above, the blocks using SINGLE mode and L1 mode will be CS encoded into M_S and M_L measurements adaptively. Here, a novel approach is proposed to realize the measurements allocation in view of the object and background regions. In general, the object regions may be more complex to require more measurements for representation, while the background regions may be smoother to need fewer measurements. Figure 3 shows the examples of the original test video sequences (Foreman, Silent, Akiyo, Suzie) and the corresponding object regions. In Figure 3, the object regions can be adaptively achieved using the differences between the neighboring frames. From Figure 3, it can be found out that most object regions may appear around the center of the frames. In fact, it is also consistent with the human visual perception. When people shoot the objects with cameras, they often put their focuses on the middle of the images.

90	57	58	59	60	61	62	63	64	65	91
89	56	31	32	33	34	35	36	37	66	92
88	55	30	13	14	15	16	17	38	67	93
87	54	29	12	3	4	5	18	39	68	94
86	53	28	11	2	1	6	19	40	69	95
85	52	27	10	9	8	7	20	41	70	96
84	51	26	25	24	23	22	21	42	71	97
83	50	49	48	47	46	45	44	43	72	98
82	81	80	79	78	77	76	75	74	73	99

Fig. 4 An example of measurements allocation for the different regions.

According to the above analysis, the central regions can contain the most objects in the frames. Therefore, in the proposed scheme more measurements will be allocated for the central regions while fewer measurements for the background of the frames. Here the measurements will be progressively allocated from the central regions to the boundaries of the frame, which is shown in Figure 4. In Figure 4, assuming that the resolution of the frames is 176×144 and the size of the block is 16×16, the different amount of measurements will be allocated to the 99 blocks according to the location of the blocks in the frame. There are four kinds of measurements allocation for SINGLE mode and L1 mode respectively which have been denoted using four different colors in Figure 4. Furthermore, Table 1 shows the parameters of the measurements allocation in SINGLE and L1 mode. It is noted that the parameters are progressively adjusted, that is, $M_L(0) > M_S(0) > M_L(1) > M_S(1) > ... > M_L(3) > M_S(3)$.

Table 1 Table Type Styles

Blocks	1-9	10-25	26-49	50-99
SINGLE mode (M_S)	$M_S(0)$	$M_S(1)$	$M_S(2)$	$M_S(3)$
L1 mode (M_L)	$M_L(0)$	$M_L(1)$	$M_L(2)$	$M_L(3)$

Although Figure 4 shows an example of the measurements allocation, it can be easily extended to general cases. The basic idea is that the measurements allocation is consistent with spiral scan from the center of the frames.

2.4 Bi-directional Dictionary

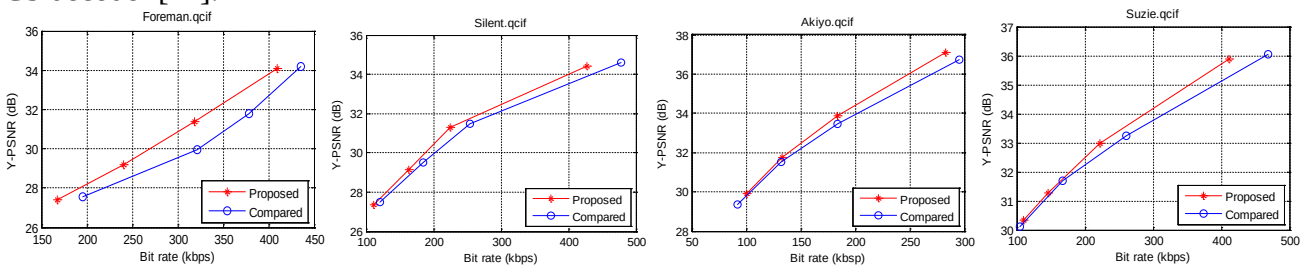
The matrix Ψ should be chosen to maximize the sparsity of the signals so that it can reduce the number of measurements to be transmitted. In most CS applications the Ψ is built by the fixed orthonormal basis such as Discrete Wavelet Transform (DWT) or Discrete Cosine Transform (DCT). In view of the temporal correlation of the video signal, the current block can be predicted by the blocks in its reference frames. Therefore, the current block can be considered as the sparse signal when it is represented as a linear combination of the reference blocks. At the decoder in [9], the dictionary Ψ of each block is built by picking blocks from recently decoded frames whose position lie in a square window of $w \times w$ pixels centered in the position of x. Since temporal correlation is very important for the video compression, the current block may be predicted more accurately by the forward and backward reference blocks. Therefore we make use of temporal correlation to design a bi-directional dictionary. For example, if the current block in the even frame F_k, both the reference blocks in the odd forward frame F_{k-1} and backward frame F_{k+1} are picked to built the bi-directional dictionary. Each reference block can be organized as a column vector in Ψ_k. Here, if the dictionary from the forward frame F_{k-1} can be achieved as Ψ_{k-1} and the other dictionary from the backward frame F_{k+1} as Ψ_{k+1}, then the proposed bi-directional dictionary is as follows.

$$\Psi_k = [\Psi_{k-1} \quad \Psi_{k+1}] \tag{2}$$

As it is shown in Equation (2), bi-directional dictionary Ψ_k can maintain better temporal correlations of the original video than the single directional dictionary Ψ_{k-1} or Ψ_{k+1}, which may lead to better reconstruction of CS decoder.

3. Experimental Results

Next, we will discuss both the rate-distortion performance and the visual quality between the proposed scheme and other relative scheme combined DVC and CS. Here, the standard video sequences “Foreman”, “Silent”, “Akiyo” and “Suzie” with QCIF format are used for the experiments. To make a fair comparison, the standard codec is H.264 codec with the version JM 10.2 and the coding mode of the key frame is pure intra-coding for simplicity. Furthermore, the same convex programming is used at the CS decoder [11].



(a) Foreman.qcif (b) Silent.qcif . (c) Akiyo.qcif (d) Suzie.qcif.

Fig. 5 Comparison of rate-distortion performance

In Figure 5, the comparison of rate-distortion performance has been shown for different video sequences. Here, in the compared scheme, the equal measurements have been adopted in each block for SINGLE mode and L1 mode, that is, $M_S(0) = M_S(1) = ... = M_S(3)$ and $M_L(0) = M_L(1) = ... = M_L(3)$. From the figures, it can be seen that the proposed scheme has achieved better rate-distortion performance than the compared scheme. The reason for this is that adaptive measurements are applied in view of the different regions. Due to adaptive measurements, more measurements are allocated for the complex contents of the object and fewer measurements for the simple background.

(a) 224 kbps, 31.4746 dB (b) 253 kbps, 31.5021 dB (c) 184 kbps, 33.9516 dB (d) 183 kbps, 33.7956 dB

Fig. 6 Comparison of visual quality. (a) and (c): proposed scheme; (b) and (d): compared sch

Figure 6 has shown the comparison of visual quality for the tested video sequences, the 16th frame of "Slient" and 20th frame of "Akiyo". From this figure, it can be seen that although the bit rates and PSNR values are comparable, the better visual quality has been achieved by the proposed scheme, especially in the regions labeled by yellow circles. Therefore, the adaptive measurements have improved the visual quality compared with the equal allocation.

4. Conclusions

We proposed a novel DVC framework combined with CS. At the encoder, flexible mode selection of the CS encoder is applied to improve coding efficiency. Furthermore, in view of the different regions, adaptive measurements are allocated for different blocks of the video frames, which have achieved both better rate-distortion performance and visual quality. At the decoder, the bi-directional dictionary is proposed as the sparse basis to improve the recovery quality of CS. Due to the characteristics of DVC and CS, the proposed video coding framework may be promising for the applications of WVSN.

Acknowledgements

Training program for excellent young teachers in Colleges and universities of Guangdong province (project number: Yq2013186); Guangzhou Education Science "12th Five-Year" planning project (project number: 12A160)

References

[1] J. Yick, B. Mukherjee and D. Ghosal, "Wireless sensor network survey," Computer Networks, vol.52, pp. 2292–2330, 2008.

[2] ITU-T Recommendation H.264, International Standard ISO/IEC 14496-10: Advanced Video Coding for Generic Audiovisual Services, International Telecommunication Union: Geneva, Switzerland, January 2012.

[3] V. Lappalainen, A. Hallapuro and T.D. Hamalainen, "Complexity of optimized H.26l video decoder implementation," IEEE Transactions on Circuits Systems for Video Technology, vol.13, pp. 717–725, 2003.

[4] B. Girod, A.M. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed video coding," Proceedings of IEEE, vol.1, pp. 71–83, 2005.

[5] F. Pereira, L. Torres, C. Guillemot, T. Ebrahimi, R. Leonardi and S. Klomp, "Distributed video coding: selecting the most promising application scenarios," Signal Processing: Image Communication, vol. 23, pp. 339–352, 2008.

[6] E.J. Candes and M.B. Wakin, "An introduction to compressive sampling," IEEE Signal Processing Magazine, vol. 25, pp.21–30, 2008.

[7] L.W. Kang, C.S.Lui, "Distributed compressive video sensing," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 09), Taibei, pp. 1169–1172, 2009.

[8] T. T. Do, Y. Chen, D. T. Nguyen, N. Nguyen, L. Gan and T. D. Tran, “Distributed compressed video sensing,” IEEE Int. Conf. on Image Processing (ICIP 09), Cairo, Egypt, pp.1393-1396, 2009.
[9] J. Prades-Nebot, Y.Ma, T. Huang, “Distributed video coding using compressive sampling,” Picture Coding Symposium (PCS 09), Chicago, USA, pp. 165–168, 2009.
[10] H. Bai, W. Lin, M. Zhang, A. Wang, Y. Zhao, Multiple Description Video Coding Based on Human Visual System Characteristics, IEEE Trans. on Circuits and Systems for Video Technology, 24(8), 1390-1394, 2014.
[11] Matlab codes: http://dsp.rice.edu/cs.