IEICE TRANS. INF. & SYST., VOL.E100–D, NO.9 SEPTEMBER 2017

PAPER Special Section on Picture Coding and Image Media Processing

A New Scheme of Distributed Video Coding Based on Compressive Sensing and Intra-Predictive Coding

Shin KURIHARA[†a)], ***Member***, **Suguru HIROKAWA**[†], ***Nonmember***, ***and*** **Hisakazu KIKUCHI**[†], ***Fellow***

SUMMARY Compressive sensing is attractive to distributed video coding with respect to two issues: low complexity in encoding and low data rate in transmission. In this paper, a novel compressive sensing-based distributed video coding system is presented based on a combination of predictive coding and Wyner-Ziv difference coding of compressively sampled frames. Experimental results show that the data volume in transmission in the proposed method is less than one tenth of the distributed compressive video sensing. The quality of decoded video was evaluated in terms of PSNR and structural similarity index as well as visual inspections.

key words: *video compression, distributed coding, compressive sensing*

1. Introduction

Recent years the Internet of things (IOT) is rapidly in progress. The development of video applications of using wireless cameras is active. For example, wireless visual sensor networks, wearable devices, remote medical equipments and vehicle-mounted cameras that are connected with cloud computing are promising areas of application. In those applications, a few of critical problems are the cost of a front-end terminal, the data amount between an encoder and a decoder, and the power consumption in battery-driven devices.

Distributed video coding (DVC) is advantageous for making a low-complexity/low-power encoder, since complex tasks such as motion estimation and motion compensation (ME/MC) can be shifted away from an encoder to a decoder [1]–[4]. Instead, a DVC decoder is powered by distributed decoding that uses the temporal/spatial correlation available from side information.

On the other hand, compressive sensing (CS) [5], [6] is attractive for the reduction in transmission data over channels. Based on the assumption of a sparse representation of a signal of interest, the original signal is recovered in the CS decoder. Many studies are available for the sparse recovery [7]–[9].

A framework of distributed compressive video sensing (DCVS) [10] is a combination of DVC and CS. It is attractive to realize low-complexity and low-power dissipation encoding and at the same time to save the required transmission data. Video data is compressively captured and transmitted to a decoder, and poor information is boosted by joint decoding with side information before sparse reconstruction [11]–[16]. In many studies, DCVS offers a simple structure in which every frame is equally subjected to CS, and nothing more is developed in the encoder. This leads to a dramatical reduction in the encoder complexity. However, since a simple data set obtained by CS observations is sent to a receiver in many DCVS systems, there is a room to save the amount of transmission data.

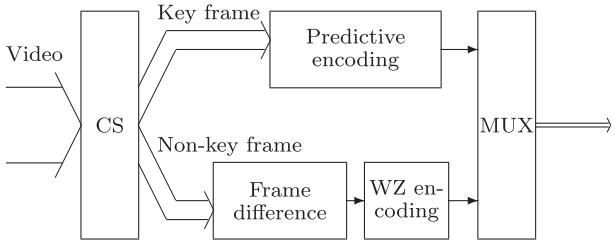


Fig. 1 Illustration of the proposed encoder.

The objectives of this work are twofold. The primary goal is a reduction in transmission data over channels. The other is to keep the encoder complexity at an acceptable level and hence to provide a framework for a low-power dissipation encoder. In the proposed architecture, every frame is compressively sampled with CS, before it is subjected to either predictive coding or distributed coding. In particular, the key frames after CS measurement are fed to intra-predictive coding which is very efficient for a spatially correlated image.

Figure 1 illustrates the proposed encoder, where a CS-measurement takes place for every frame to save the number of samples. As a result, a significant reduction in transmission data is achieved. The encoder complexity can increase than pure DCVS encoders, but can stay at a reasonable level compared to popular video codecs, since ME/MC are out of use. This architecture is hence advantageous to reduce the computational burden in the encoder and the communication traffic over transmission channels. The key frames are encoded by predictive coding, specifically by H.26x intra-predictive coding, while the non-key frames are encoded by differential Wyner-Ziv (WZ) encoding [2], [17]. The proposed encoder is hence viewed as a combination of compressive sensing, predictive coding, and distributed coding. Conceptually, it may be viewed as a fusion of compressive sensing and DVC.

The rest of the paper is organized as follows. Related works are overviewed in Sect. 2. A new codec system is proposed in Sect. 3. Experimental results are given in Sect. 4,

Manuscript received December 16, 2016.
Manuscript revised April 14, 2017.
Manuscript publicized June 14, 2017.

[†]The authors are with Department of Electrical and Electronic Engineering, Niigata University, Niigata-shi, 950–2181 Japan.
a) E-mail: f11l503g@mail.cc.niigata-u.ac.jp

DOI: 10.1587/transinf.2016PCP0009

Copyright © 2017 The Institute of Electronics, Information and Communication Engineers

where the quality of decoded video is evaluated in terms of PSNR and structural similarity index (SSIM) as well as visual inspections. Conclusions follow in Sect. 5.

2. Related Works

A few related works are overviewed.

2.1 Distributed Video Coding (DVC)

DVC has been proposed to reduce the complexity in video encoding [1], [4]. Successive frames of a video sequence are encoded independently at the encoder, while they are subjected to a joint decoding at the decoder with a help of correlation among neighboring frames.

Key frames are encoded and decoded using a conventional intra frame codec such as DCT or H.26x intra prediction. Non-key frames are treated by Wyner-Ziv encoding. At the DVC encoder, the Wyner-Ziv bits are generated by quantization and channel encoding of the inter-frame difference. The decoder generates side information (SI) based on Wyner-Ziv theorem. It is the interpolated frames generated by ME/MC of the neighboring key frames and previously decoded non-key frames. The decoder assumes a statistical model as a virtual correlation channel. The decoder uses the received Wyner-Ziv bits and SI to perform channel decoding of non-key frames.

2.2 Compressive Sensing

CS is very attractive in imaging applications, especially for low-power and low-resolution imaging devices [5]. We assume that a real valued signal $\boldsymbol{x}$ of length N is represented with an $N \times N$ basis matrix $\boldsymbol{A}$ as

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{1}$$

where an N-dimensional sparse vector $\boldsymbol{s}$ is well approximated by at most K non-zero entries in such a way that $K \ll N$. CS states that $\boldsymbol{x}$ is accurately reconstructed by taking only M linear and non-adaptive measurements, where M satisfies

$$K < M = O(K \log(N/K)) \ll N. \tag{2}$$

The measurement process is described by

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x}, \tag{3}$$

where $\boldsymbol{y}$ is an M-dimensional observation vector, and $\boldsymbol{C}$ is an $M \times N$ measurement matrix that is incoherent with $\boldsymbol{A}$. One thus obtains

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{s}. \tag{4}$$

M entries in $\boldsymbol{y}$ are random linear combinations of the entries of $\boldsymbol{s}$, which are viewed as a compressed and encrypted data of $\boldsymbol{x}$. To reconstruct $\boldsymbol{s}$ from $\boldsymbol{y}$, the convex optimization problem is solved through a minimization on the ℓ_1 norm as

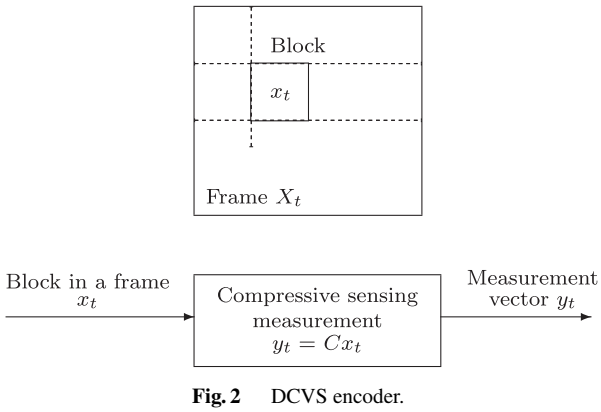


Fig. 2 DCVS encoder.

follows.

$$\boldsymbol{s}' = \arg\min_{s} \left\{ \|\boldsymbol{s}\|_1 \,\middle|\, \boldsymbol{y} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{s} \right\}, \tag{5}$$

where $\boldsymbol{s}'$ is an estimate of $\boldsymbol{s}$. The primal-dual interior point method [7] can be applied to solve the CS problem. Any other minimization techniques may be used [6]. The reconstruction of $\boldsymbol{x}$ is finally obtained as $\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{s}'$.

2.3 Gradient Projection for a Sparse Reconstruction

The gradient projection for a sparse reconstruction (GPSR) is one of the fast iterative algorithms [9]. It is used to solve a convex unconstrained optimization problem arising in compressed sensing. The problem is of the form,

$$\boldsymbol{s}' = \arg\min_{\boldsymbol{s}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{A}\boldsymbol{s}\|_2^2 + \lambda \|\boldsymbol{s}\|_1 \right\}, \tag{6}$$

where λ is a Lagrange multiplier.

2.4 Distributed Compressive Video Sensing

The theory of distributed compressive video sensing (DCVS) was presented by Kang and Lu [10]. The DCVS endoder does not mind if it is a key frame or non-key frame. A video sequence is simply identified with a collection of frames. The DCVS encoder is illustrated in Fig. 2, where ME/MC are absent. Since CS measurement is applied to every frame, no matter what it is a key frame or not, the structure of the encoder is simple. DCVS assumes none of information among successive frames. Data compression is simply gained by a reduction in dimension through the CS measurement, where the measurement matrix is a randomized block Hadamard transform.

A typical DCVS decoder is illustrated in Fig. 3. At the decoder, a key frame is reconstructed with their modified GPSR. A non-key frame is reconstructed by GPSR with a help of a picture estimate for that frame. The picture estimate is generated by ME/MC-based interpolation from the latest reconstructions of a pair of neighboring key frames. Since this forms a good initial estimate for GSPR, the algorithm reaches a convergence after a few iterations.

2.5 Intra Prediction of H.264/AVC

Intra prediction of H.264/AVC exploits the spatial correlation between adjacent coding blocks [18]. The current block is predicted by adjacent pixels in the upper and left blocks that are already decoded. H.264/AVC offers a rich set of prediction patterns for intra prediction: nine prediction modes for 4×4 and 8×8-luma blocks and four prediction modes for 16 × 16-luma blocks. Each mode follows a specific prediction direction and the predicted samples are obtained by a weighted average of decoded values of neighboring blocks. The standard intra prediction of H.264/AVC uses the Lagrange optimization in order to counterbalance the compromise between the cost for bit streams and the quality of video for a given quantization parameter (QP) in such a way that

$$\min\{J|\, J = D + \lambda R\}, \tag{7}$$

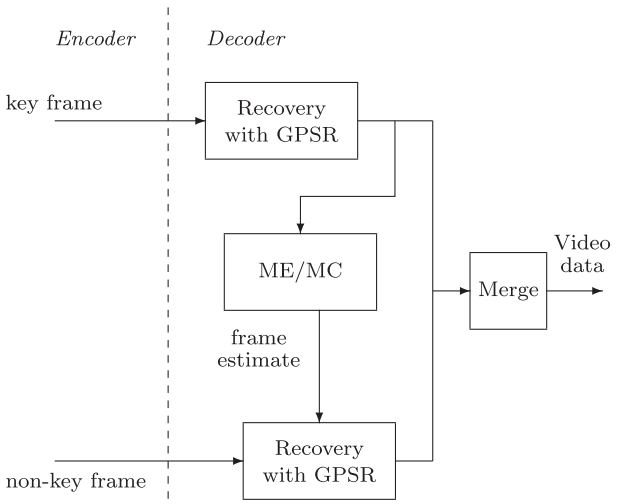


Fig. 3 DCVS decoder.

where D is a measure of distortion and R is the cost for the bits per block. λ is a Lagrange multiplier, which is a function of QP.

3. Proposed Codec Architecture

The architecture of the proposed codec system is illustrated in Fig. 4. The underlying concept is to treat the CS-measurement of every frame as an ordinary picture so that predictive coding may be applied for efficient video transmission. The observation data for a key frame is encoded by the intra prediction. On the other hand, the observation data for a non-key frame is encoded by WZ coding. Since the data of both observations are efficiently compressed in the encoder, it is possible to reduce the data volume in transmission.

3.1 Encoding and Decoding of Key Frames

At the encoder, a key frame X_k is observed by block-based compressive sensing, where a block is of $N_b \times N_b$ pixels. The elements of a block are expressed by $\boldsymbol{x}_k$. The N_b^2 elements in a block is reshaped into an N-dimensional column vector, where $N = N_b^2$. A CS measurement is developed onto the column vector.

$$\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k = \boldsymbol{C}\boldsymbol{A}\boldsymbol{s}_k \tag{8}$$

is an M-dimensional vector. $\boldsymbol{C}$ is an $M \times N$ measurement matrix, where $M/N < 1$. $\boldsymbol{A}$ consists of a set of DCT basis vectors.

All of the entries of the observed block-measurement signals $\{\boldsymbol{y}_k\}$ are concatenated to form an image Y_k. As illustrated in Fig. 5, the vertical dimension of a compressed frame Y_k is shrunk to M/N of the original one. The proposed way for vector packing is a brute force convenience. As observed in the magnified view in the figure, the spatial structure is considerably destroyed, but the spatial correlation is still preserved to a significant extent.

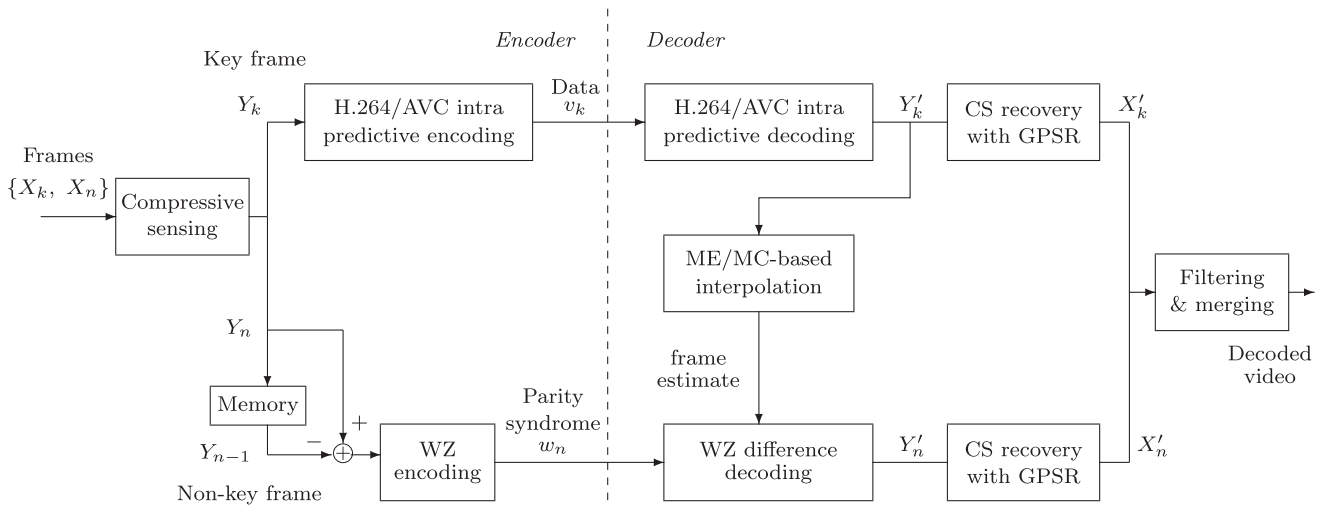


Fig. 4 The proposed codec architecture.

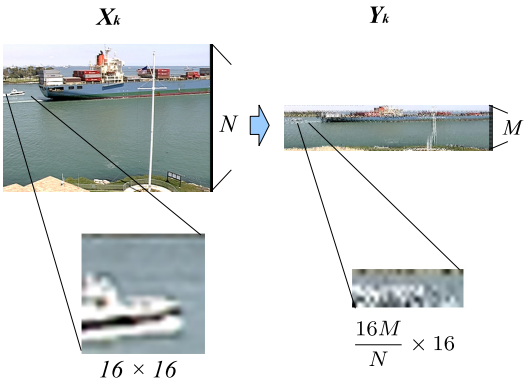


Fig. 5 The first frame of *container*. Left: original. Right: after packing of blockwise compressively sensed data into a smaller rectangular block. The top and bottom rows show the raw picture and a magnified view, respectively.

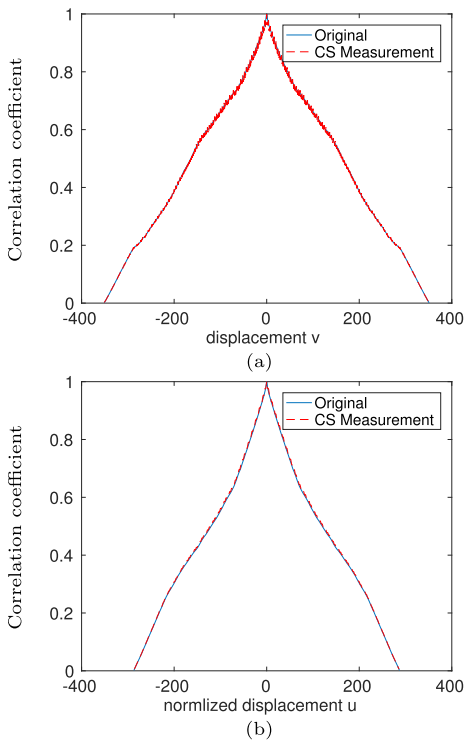


Fig. 6 The spatial correlation in the 3rd CS frame of *hall* in CIF. Horizontal direction (a), and vertical direction (b).

The auto correlation coefficients along the horizontal and vertical directions are plotted in Fig. 6, where the horizontal axis is the displacement between pixels. In the plots, it is a surprise to observe that the blockwise compressively-sampled observation and the original frame show the identical spatial auto correlations. This is a supporting evidence why intra predictive coding is applied to the CS observation data regardless of its nature of random sampling.

It is now possible to apply H.264/AVC intra prediction to the compressed frame Y_k. Y_k is converted into a luma/chroma signal in YCbCr 4:2:0 before intra predictive coding. The encoded data v_k is sent to a decoder [19]. At the decoder, a picture block Y'_k is decoded by intra predictive decoding of v_k. The decoded Y'_k is divided into blocks to form a set of vectors, $\{\boldsymbol{y}'_k\}$. A sparse reconstruction is developed on $\boldsymbol{y}'_k$ as follows.

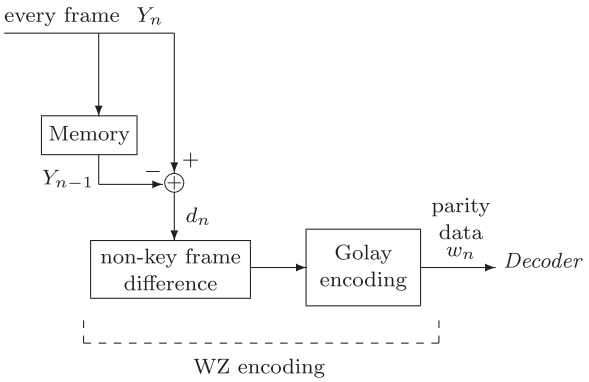


Fig. 7 Encoding of non-key frames.

$$\boldsymbol{s}'_k = \arg\min_{\boldsymbol{s}_k} \left\{ \frac{1}{2}\|\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{A}\boldsymbol{s}_k\|_2^2 + \lambda\|\boldsymbol{s}_k\|_1 \right\}. \tag{9}$$

The decoded key frame X'_k is reconstructed from the solution vectors $\{\boldsymbol{x}'_k\}$ that are obtained by a substitution of $\boldsymbol{s}'_k$ into Eq. (1).

3.2 Encoding and Decoding of Non-Key Frames

At the encoder, a non-key frame X_n is observed by the block-based compressive sensing as same as a key frame. The resulting data is formed into a vertically-shrunk pseudo frame, Y_n. As illustrated in Fig. 7, the difference between the successive frames is fed to Golay encoding [20]. The non-key frame difference d_n is calculated by

$$d_n = Y_n - Y_{n-1}. \tag{10}$$

d_n is encoded by Golay coding, and the parity syndrome w_n is generated to be sent to a decoder. The difference (residual) coding is the simplest predictive coding. An intra-frame or inter-frame difference in between the neighboring and previous pixel can be a good choice for data compression. The former has been applied to key frames but in a more sophisticated prediction. The latter is here applied to non-key frames. The basic idea is common to WZ residual coding in Ref. [3].

At the decoder, ME/MC-based interpolation is applied to key frame pictures to obtain the estimates of non-key frames. The algorithm of motion estimation is the simple and efficient search (SES) [21]. The decoding process is illustrated in Fig. 8, where the frame estimate is refined by error-correcting Golay decoding of the frame difference to obtain Y'_n. Finally, the decoded non-key frame X'_n is obtained by GPSR of Y'_n.

3.3 Post Processing

For every frame, a Gaussian lowpass filter is applied to sup-

press the block noise created by a block-based CS reconstruction. The lowpass filtering (LPF) is followed by unsharp masking (USM) to compensate blurred effects and to emphasize the sharpness around edges. Finally, the decoded key and non-key frames are merged to form a video sequence.

Figure 9 shows a pair of pictures for the 8th frame of *hall* sequence. Parts (a), (b), and (c) show the filtering-free reconstruction, the filtered one, and that after LPF and USM, respectively. Block artifacts are visible in part (a), whereas they are absent in part (b) at the cost of blurring effect. In (c), the sharpness has been visually improved around edges. As for objective figures of merit, PSNR and SSIM have increased by 1.36dB and 0.2, respectively. In conclusion, the visual quality of decoded frames is improved by the post processing.

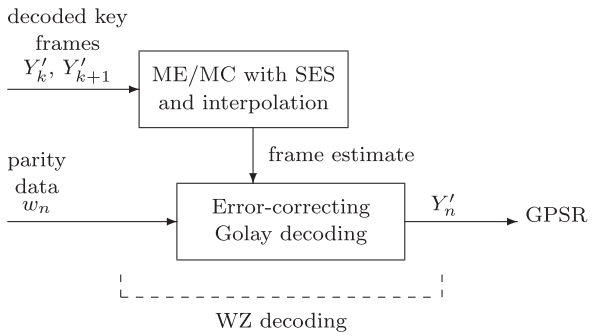


Fig. 8 Decoding of non-key frames.

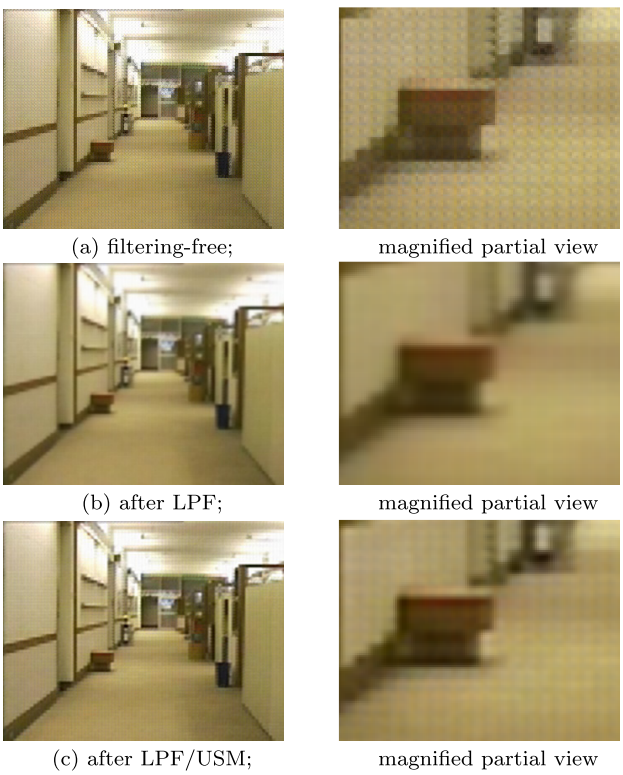


Fig. 9 The eighth decoded frame of *hall* in CIF. (a) filtering-free: PSNR 28.24dB, SSIM 0.60, (b) Gaussian lowpass filtered: PSNR 28.83dB, SSIM 0.79, and (c) Gaussian lowpass filter and unsharp masking (USM): PSNR 29.60dB, SSIM 0.80.

4. Experiments

4.1 Experimental Conditions

Image quality is evaluated by means of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [22] that shows high correlations with the subjective scores in image quality assessments [23]. Needless to mention, PSNR is a mean squared error over a picture and SSIM is sensitive to the erroneous activity around edges. Since there is none of good established objective measures for visual quality assessment, and since those two measures treat different types of error, objective assessments are developed on them. Remember that it does not matter which measure is better.

PSNR is evaluated in YCbCr 4:2:0. In this paper, the test video sequences are *container, foreman, tempete, hall, news, bus* and *bridge-close* in CIF (352 × 288-pixel), of which frame rate is 15fps. The GOP size is fixed at five frames per GOP. In block-based compressive sensing, the measurement block size is 4×4, and $N_b = 4$. The $M\times N$ observation matrix is a random matrix of which size is 8×16, and the compression ratio is $M/N = 0.5$. A sparse basis matrix $\boldsymbol{A}$ is formed by the backward DCT basis. These parameters are common for DCVS and the proposed method. The quantization parameter is set as $QP = 20$ on intra prediction in H.264/AVC. The macro-block size in ME is 8×8 and the algorithm is the simple and efficient search (SES) [21]. The standard deviation of the Gaussian filter is experimentally determined as $\sigma = 1.2$.

4.2 Experimental Results on Objective Measures

The transmission data volume of a key frame is compared between the proposed system and DCVS [10]. The result is listed in Table 1, where the rightmost column stands for

Table 1 PSNR, SSIM, and data size ratio with respect to a key frame. Every 16th key frame was tested and the average values are listed. The data size implies the data amount of a frame.

Test video	Method	PSNR in dB	SSIM	Data size ratio
container	DCVS	28.26	0.65	1.00
	Proposed	28.89	0.76	0.13
hall	DCVS	28.95	0.70	1.00
	Proposed	29.67	0.81	0.17
foreman	DCVS	31.10	0.67	1.00
	Proposed	32.02	0.81	0.18
tempete	DCVS	28.11	0.70	1.00
	Proposed	27.94	0.69	0.18
news	DCVS	28.83	0.81	1.00
	Proposed	29.73	0.82	0.12
bus	DCVS	26.59	0.68	1.00
	Proposed	27.11	0.66	0.17
bridge-close	DCVS	29.67	0.67	1.00
	Proposed	29.80	0.76	0.17

Table 2 PSNR, SSIM, and data size ratio with respect to a non-key frame. Every 8th frame was tested and the average values are listed. The data size implies the data amount of a frame.

Test video	Method	PSNR in dB	SSIM	Data size ratio
container	DCVS	28.23	0.65	1.000
	Proposed	28.45	0.73	0.001
hall	DCVS	29.31	0.70	1.000
	Proposed	29.60	0.80	0.002
foreman	DCVS	30.57	0.67	1.000
	Proposed	29.25	0.72	0.003
tempete	DCVS	28.26	0.57	1.000
	Proposed	26.79	0.63	0.005
news	DCVS	29.30	0.80	1.000
	Proposed	28.95	0.80	0.001
bus	DCVS	26.30	0.69	1.000
	Proposed	23.56	0.42	0.037
bridge-close	DCVS	29.63	0.67	1.000
	Proposed	29.76	0.73	0.001

Table 3 PSNR and bit rates with respect to a GOP. Six GOPs (i.e. 30 frames) were tested and PSNR denotes the average value.

Test video	Method	PSNR in dB	Bit rate in Mbps	Bit rate ratio
container	DCVS	28.25	9.12	1.00
	Proposed	28.57	0.30	0.03
hall	DCVS	29.27	9.12	1.00
	Proposed	29.04	0.32	0.04
foreman	DCVS	30.77	9.12	1.00
	Proposed	29.14	0.35	0.03
tempete	DCVS	28.22	9.12	1.00
	Proposed	26.66	0.37	0.04
news	DCVS	29.29	9.12	1.00
	Proposed	29.17	0.24	0.03
bus	DCVS	27.03	9.12	1.00
	Proposed	23.56	0.57	0.06
bridge-close	DCVS	29.65	9.12	1.00
	Proposed	29.70	0.31	0.03

the data size ratio divided by the data size of DCVS, and every 16th frame is subjected to the comparison. As seen in the table, the transmitted data of a key frame in the proposed system is less than one fifth of that in DCVS under the condition that both methods show competitive values in PSNR. It is observed that the values of structural similarity index (SSIM) are also similar for both methods. Precisely speaking, the proposed method results in higher SSIM than DCVS for most of cases.

Table 2 lists the same items as the previous table, but for a different case of non-key frame transmission where every 8th frame is subjected for experiment. The pairs of values of PSNR and SSIM between the proposed method and DCVS are almost the same. In contrast, the transmission data size of the proposed method is much smaller than that of DCVS. Typically it amounts to 0.1-3.7% of DCVS, and it is an evidence for its dramatical performance. In the cases for *bus*, the transmission data are around 4%, because fast motions are dominantly present in those sequences.

The proposed method and DCVS are again compared with respect to GOP encoding as shown in Table 3, where the bit rate is in bits per second rather than bits per GOP. Note that the frame rate is 15fps, and thus six GOPs are equal to the period of two seconds. As seen in the table, the proposed system offers much lower bit rates than DCVS. The improvement in bit rates owes to the differential compression of non-key frames, whereas the DCVS codec operations are executed on every frame, no matter which is a key frame or non-key frame.

4.3 Visual Inspections and Other Comprehensive Evaluations

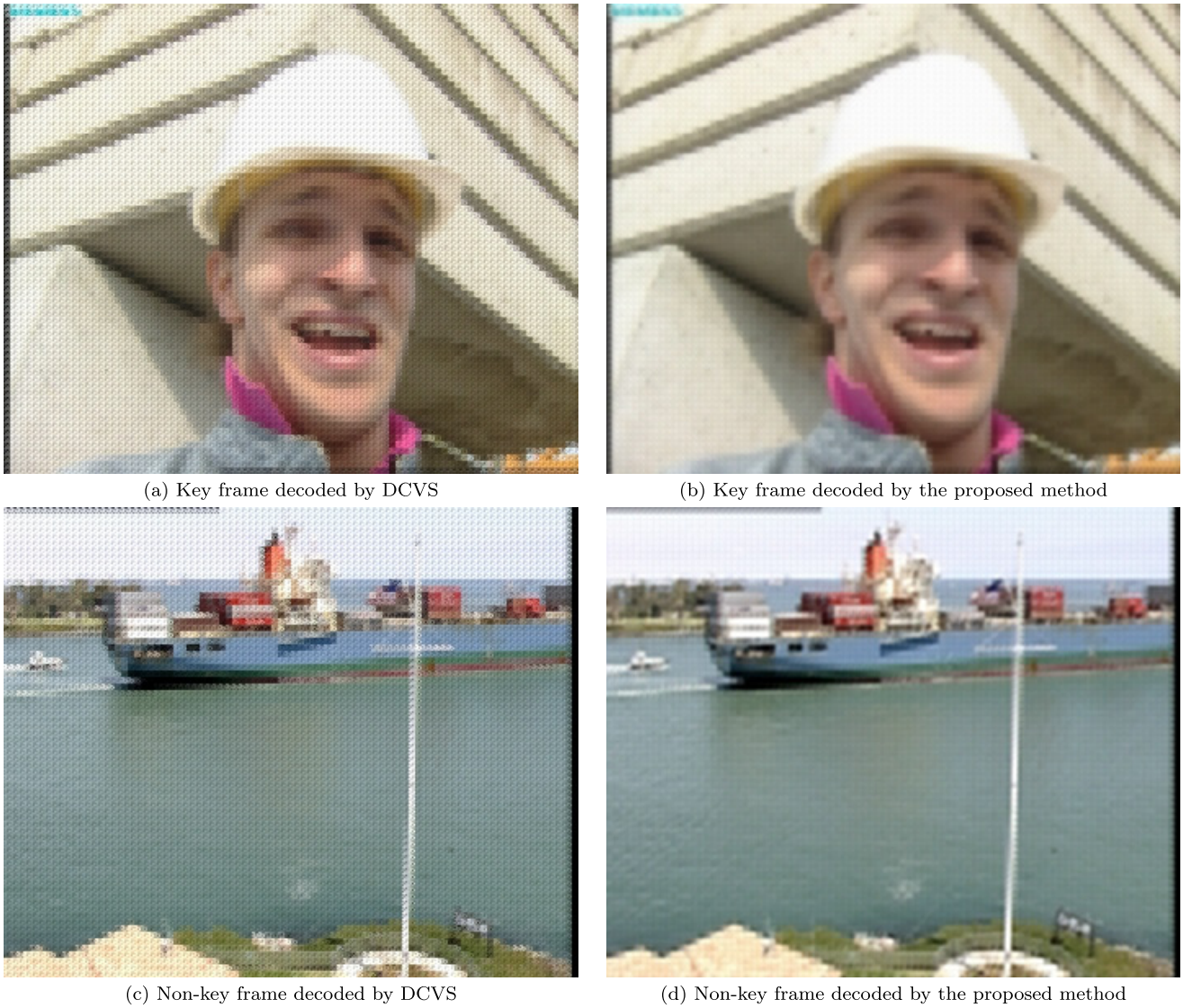
In Fig. 10, decoded key and non-key frames are shown in the top and bottom row, respectively. The decoded samples from DCVS and the proposed method are placed at the left and right column, respectively. Block artifacts are evident in the decoded pictures from DCVS as seen in the left column. In particular, the thin flag pole looks smeared and jaggy in the non-key frame decoded by DCVS as observed in (c) in spite of a considerably large data size. In contrast as seen in the right column, significant artifacts are imperceptible in both pictures decoded by the proposed method. Note that the data size for the non-key frame in (d) is as little as just 22k bits, whereas it is 204k bits in (c). As for key frames, the better visual performance of the proposed method owes to the high performance in H.264 intra-prediction. Alternately, as for non-key frames, the higher performance of the proposed method is gained by a low-distortion quantization in WZ difference encoding and the post processing.

Figures 11 and 12 show the performance comparisons over time with respect to PSNR and SSIM for *hall* and *foreman*, respectively. Blue and red plots represent DCVS and the proposed method, respectively.

In the case of lower motion video such as *hall*, both methods are competitive in PSNR, while the proposed method outperforms DCVS with respect to SSIM. Significant degradations were imperceptible in the video sequences decoded by the proposed method.

If fast motions are dominant as such as in *foreman*, high fluctuations in PSNR and SSIM are visually observed in the sequence decoded by the proposed method. The fidelity of decoded images decreases because of poor information for non-key frames.

The rate-distortion performance is compared in Fig. 13 among DVC [2], DCVS [10], and the proposed method. The rate-distortion plots on H.264/AVC (VM 9.5) and MPEG-2 are also overlayed for reference. The test sequence is *tempete* in CIF format. As seen in the figure, the proposed method operates at very low bit rates compared with DVC and DCVS. This means that the primary goal of this work has been successfully satisfied. Since the bit rate was mainly controlled by the decimation factor, M/N, in compressive sensing for DCVS and the proposed method, the operation curves are missing on some intervals along the operational bit rates. In contrast, conventional hybrid video codecs such as H.264 and MPEG-2 shows better performances even at very low bit rates at the expense of high complexity in encoding.



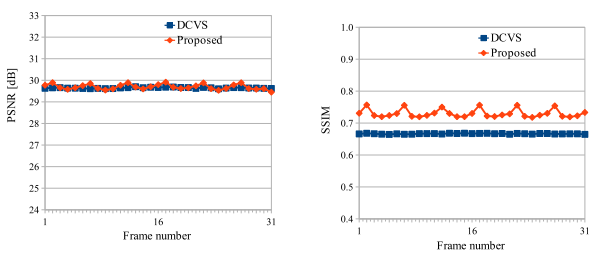
(a) Key frame decoded by DCVS

(b) Key frame decoded by the proposed method

(c) Non-key frame decoded by DCVS

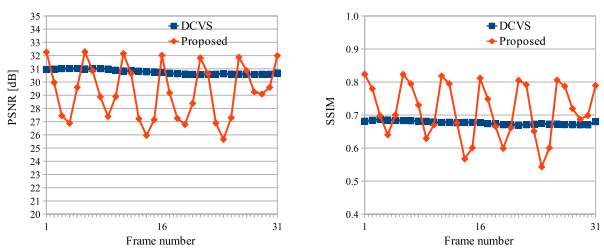
(d) Non-key frame decoded by the proposed method

Fig. 10 Decoded frames of *foreman* and *container* in CIF format for visual inspections. (a) 16th frame by DCVS; PSNR 30.43dB, SSIM 0.78. (b) 16th frame by the proposed method; PSNR 32.20dB, SSIM 0.85. (c) 8th frame by DCVS; PSNR 26.59dB, SSIM 0.41. (d) 8th frame by the proposed method; PSNR 28.75dB, SSIM 0.74.



(a) (b)

Fig. 11 Temporal variations in (a) PSNR and (b) SSIM for *hall*. Two plots of DCVS and the proposed are in blue and red, respectively.



(a) (b)

Fig. 12 Temporal variations in (a) PSNR and (b) SSIM for *foreman*. Two plots of DCVS and the proposed are in blue and red, respectively.

4.4 Computational Complexity and Others

The computational complexity of the proposed encoder exceeds that of DCVS because of the intra-predictive coding. It is however less than that of DVC due to the data reduction by compressive sensing.

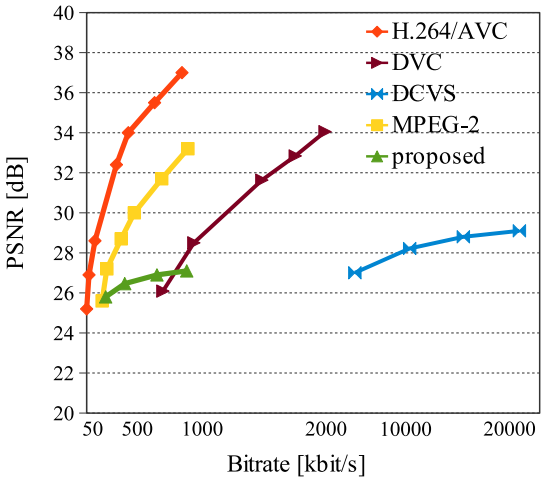


Fig. 13 The rate-distortion plots for *tempete* in CIF format.

Table 4 Encoding time. Six GOPs, i.e. 30 frames, were tested.

Method	Frames	Encoding time in sec	Total time in sec
Proposed	key	2.25	10.89
	non-key	8.64	
DCVS [10]	key	0.74	3.75
	non-key	3.01	
DVC [2]	key	5.23	20.12
	non-key	14.89	
H.264/AVC	key	13.56	59.64
	non-key	46.08	

Encoding time for six GOPs including 30 frames is compared among the proposed method, DCVS, DVC, and H.264 in Table 4. The proposed method needs a longer time by a few times than DCVS, because of intra predictive coding. On the contrary, it is twice faster than DVC, because DVC operates on the whole pixels in full resolution. As anticipated, the H.264 encoder requires the longest time.

The proposed decoder consists of H.264 decoding of key frames, ME/MC-based interpolation for side information for non-key frames, and iterative sparse reconstruction. The decoder is hence much more complex than the encoder. However, the expense for intra-frame decoding and ME/MC is much less than the conventional hybrid decoders with transform-based predictive decoding and ME/MC because of the decimation of raw samples by compressive sensing. Compared to the DCVS decoder, the proposed decoder is more complex due to H.264 decoding and ME/MC. However, it is of less (or competitive) complexity than (or to) DVC decoders where all pixels in a full resolution are processed. As for the WZ difference decoding including sparse reconstruction but excepting side information generation, it is not so heavy but is more time-consuming than arithmetic decoding in most of conventional decoders.

On the other hand, the transmission data over channels decreases, since the raw data being encoded shrinks by a factor of M/N and is further compressed by intra predictive coding and WZ difference coding. It is worth to mention that deterioration in visual quality is suppressed by post processing of LPF and unsharp masking. The proposed system is hence advantageous for wireless applications and hand-held terminals with low-power requirements.

5. Conclusions

A new compressive sensing-based distributed video coding has been presented. In the proposed scheme, the encoder is made of a block-based compressive sensing of a video frame and a pair of encoding units. One is the intra predictive coding for key frames, and the other is the Wyner-Ziv difference coding for non-key frames. The proposed scheme results in a great deal of reduction in transmission/storage data, while the encoding complexity stays at an intermediate level between compressive sensing-encoding with no prediction and distributed encoding. Experimental results showed that the transmitted data volume of the proposed scheme is less than one tenth to that of the conventional DCVS.

Acknowledgements

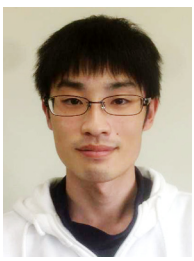
A part of this work was supported by JSPS Grants-in-Aid for Scientific Research, No. 16K06343.

References

[1] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," Proc. IEEE, vol.93, no.1, pp.71–83, 2005.
[2] A. Aaron, S.D. Rane, E. Setton, and B. Girod, "Transform-domain wyner-ziv codec for video," Proc. SPIE VCIP, San Jose, CA, USA, Jan. 2004.
[3] A. Aaron, D. Varodayan, and B. Girod, "Wyner-ziv residual coding of video," Proc. Picture Coding Symp., Beijing, China, pp.28–32, April 2006.
[4] 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.
[5] D.L. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, 2006.
[6] J. Romberg, "Imaging via compressive sampling," IEEE Signal Processing Magazine, vol.25, no.2, pp.14–20, 2008.
[7] Wright and S.J, Primal-dual Interior-point Methods, Society for Industrial and Applied Mathematics, 1987.
[8] R. Puri and K. Ramchandran, "Prism: A "reversed" multimedia coding paradigm," Proc. IEEE Int. Conf. Image Processing, Barcelona, Spain, 2003.
[9] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," IEEE J. Selected Topics in Signal Processing, vol.1, no.4, pp.586–597, 2007.
[10] L.-W. Kang and C.-S. Lu, "Distributed compressive video sensing," IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp.1169–1172, IEEE, 2009.
[11] T.T. Do, Y. Chen, D.T. Nguyen, N. Nguyen, L. Gan, and T.D. Tran, "Distributed compressed video sensing," Conf. Information Sciences and Systems (CISS), pp.1–2, 2009.
[12] J. Prades-Nebot, Y. Ma, and T. Huang, "Distributed video coding using compressive sampling," Proc. Picture Coding Symposium, Chicago, Illinois, USA, pp.1–4, May 2009.
[13] L. Liu, A. Wang, Z. Li, and K. Zhu, "An improved distributed compressive video sensing based on adaptive sparse basis," First Int.

Conf. Robot, Vision and Signal Processing, pp.137–140, 2011.
[14] Z. Xue, W. Anhong, Z. Bing, and L. Lei, "Adaptive distributed compressed video sensing," Journal of Information Hiding and Multimedia Signal Processing, vol.5, pp.98–106, Jan. 2014.
[15] F. Tian, J. Guo, B. Song, H. Liu, and H. Qin, "Distributed compressed video sensing with joint optimization of dictionary learning and l_1-analysis based reconstruction," IEICE Trans. Inf. & Syst., vol.E99-D, no.4, pp.1202–1211, April 2016.
[16] J. Xu, Y. Qiao, and Q. Wen, "Rate-distortion optimized distributed compressive video sensing," IEICE Trans. fundamentals, vol.E99-A, no.6, pp.1272–1276, June 2016.
[17] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," IEEE Trans. Information Theory, vol.22, no.1, pp.1–10, Jan. 1976.
[18] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits Syst. Video Technol, vol.13, no.7, pp.560–576, 2003.
[19] S. Kurihara and H. Kikuchi, "An improvement of key frame processing by an integration of compressive sensing and intra prediction of H.264/AVC," IEEE Region 10 Symposium (TENSYMP), Kuala Lumpur, Malaysia, pp.584–587, 2014.
[20] M.J.E. Golay, "Notes on digital coding," Proc. IRE, vol.37, p.657, 1949.
[21] J. Lu and M.L. Liou, "A simple and efficent search algorithm for block-matching motion estimation," IEEE Trans. Circuits & Systems For Video Technology, vol.7, no.2, pp.429–433, April 1997.
[22] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Processing, vol.13, no.4, pp.600–612, April 2004.
[23] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," Advances of Modern Radioelectronics, vol.10, pp.30–45, 2009.

Shin Kurihara received his B.S. and M.S. degrees in electrical engineering from Niigata University, Japan, in 2006 and 2008, respectively. Currently, he is a Ph.D. student in Department of Electrical and Electronic Engineering, Niigata University. His research interests lie in distributed video coding and compressive video sensing.



Suguru Hirokawa received his B.S. degrees in electrical engineering from Niigata University, Japan, in 2016. Currently, he is a M.S. student in Department of Electrical and Electronic Engineering, Niigata University. His research interests are in distributed video coding and compressive video sensing.

Hisakazu Kikuchi received B. E. and M. E. degrees from Niigata University, Niigata, in 1974 and 1976, respectively, and Dr. Eng. degree in electrical and electronic engineering from Tokyo Institute of Technology, Tokyo, in 1988. From 1976 to 1979 he worked at Information Processing Systems Laboratory, Fujitsu Ltd., Tokyo. Since 1979, he has been with Niigata University, where he is a Professor in the Department of Electrical and Electronic Engineering. During the 1992 academic year, he was a visiting scholar in the Electrical Engineering Department, University of California, Los Angeles, USA. He holds a visiting professorship at Chongqing University of Posts and Telecommunications and Nanjing University of Information Science and Technology, both in China, since 2002 and 2005, respectively. His research interests include digital signal processing and image processing. He is a Fellow of IEICE and a Member of ITE (Institute of Image Information and Television Engineers of Japan), IIEEJ (Institute of Image Electronics Engineers of Japan), and IEEE. He served the general co-chair of ITC-CSCC 2011 in Korea, the chair of Circuits and Systems Group, IEICE, in 2000 and the general chair of Digital Signal Processing Symposium, IEICE, in 1998 and Karuizawa Workshop on Circuits and Systems, IEICE, in 1996, respectively.