Dynamic measurement rate allocation for distributed compressive video sensing

Hung-Wei Chen, Li-Wei Kang, Chun-Shien Lu*
Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC

ABSTRACT

We address an important issue of fully low-cost and low-complexity video encoding for use in resource limited sensors/devices. Conventional distributed video coding (DVC) does not actually meet this requirement because the acquisition of video sequences still relies on the high-cost mechanism (sampling + compression). Recently, we have proposed a distributed compressive video sensing (DCVS) framework to directly capture compressed video data called measurements, while exploiting correlations among successive frames for video reconstruction at the decoder. The core is to integrate the respective characteristics of DVC and compressive sensing (CS) to achieve CS-based single-pixel camera-compatible video encoder. At DCVS decoder, video reconstruction can be formulated as a convex unconstrained optimization problem via solving the sparse coefficients with respect to some basis functions. Nevertheless, the issue of measurement rate allocation has not been considered yet in the literature. Actually, different measurement rates should be adaptively assigned to different local regions by considering the sparsity of each region for improving reconstructed quality. This paper investigates dynamic measurement rate allocation in block-based DCVS, which can adaptively adjust measurement rates by estimating the sparsity of each block via feedback information. Simulation results have indicated the effectiveness of our scheme. It is worth noting that our goal is to develop a novel fully low-complexity video compression paradigm via the emerging compressive sensing and sparse representation technologies, and provide an alternative scheme adaptive to the environment, where raw video data is not available, instead of competing compression performances against the current compression standards (*e.g.*, H.264/AVC) or DVC schemes which need raw data available for encoding.

Keywords: Compressive sensing, sparse representation, distributed compressive video sensing, measurement rate allocation, single-pixel camera, dictionary learning, distributed video coding, low-complexity video coding.

1. INTRODUCTION

Low-complexity video coding has been potentially applicable for several emerging applications, such as video conferencing with mobile devices and wireless visual sensor networks (WVSN)[1]. Since the low-complexity restriction for a video device, efficient video compression is challenging. In particular, distributed video coding (DVC)[1] based on the principle of distributed source coding (DSC)[2] has been recently proposed to reduce video encoding complexity to the order of that for still image encoding via shifting major encoding burden to the decoder. Nevertheless, even for still image encoding, it is required to capture huge amounts of raw image data first, followed by performing some transformation operator, which is also memory- and computation-intensive[3-4]. With the advent of the compressive sensing (CS)-based single-pixel camera architecture[5], CS is an emerging technology and enables to directly and efficiently capture compressed image data via randomly projecting raw image data to obtain linear and non-adaptive measurements. The image can then be reconstructed at the decoder via solving the convex optimization problem or using some iterative greedy algorithms[6-7] from the captured data measurements.

To directly capture compressed video data, a compressive video sensing framework[8] has been proposed to individually capture and reconstruct each compressed video frame. Recently, compressive video sensing integrating both DVC and CS characteristics has emerged as a new way to directly capture compressed video data via random projection at a low-complexity encoder while performing CS reconstruction together with exploiting correlations among successive frames at a high-complexity decoder[9-12]. A general structure is to divide a video sequence into several key frames and CS frames. Each key frame can be individually compressed and reconstructed while each CS frame can be individually compressed and conditionally reconstructed. We have proposed a distributed compressive video sensing (DCVS) framework[9], where an efficient initialization and several stopping criteria were proposed to improve and speedup the employed convex optimization algorithm for CS frame reconstruction with respect to the discrete wavelet transform (DWT) basis. In

*lcs@iis.sinica.edu.tw; phone 886-2-2788-3799 ext. 1513; fax 886-2-2782-4814. This work was supported in part by National Science Council, Taiwan, R.O.C, under Grant NSC 97-2628-E-001-011-MY3.

Visual Communications and Image Processing 2010, edited by Pascal Frossard, Houqiang Li, Feng Wu, Bernd Girod, Shipeng Li, Guo Wei, Proc. of SPIE Vol. 7744, 77440I · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.863094

addition, a DVC algorithm using CS was proposed[11], where at the decoder, each block in a CS frame is reconstructed with respect to the basis (dictionary) formed from a set of spatially neighboring blocks of previous decoded neighboring key frames. Similarly, a distributed compressed video sensing (DISCOS) framework was also proposed[12], where the major core is also to assume each block in a CS frame can be sparsely represented with respect to the dictionary formed from a set of spatially neighboring blocks of previous decoded neighboring key frames. Here, we denote the two above-mentioned schemes[11-12] as the "local dictionary"-based scheme for their major core employing the local blocks extracted from the neighboring frames as the dictionary for each block in a CS frame.

Similar to rate control/allocation for conventional video coding[13] or DVC[14], measurement rate allocation is very critical for a block-based CS video encoder. Here, the measurement rate (*MR*) for a signal (*e.g.*, an image or an image block) is defined as:

$$MR = \frac{M}{N}, \tag{1}$$

where N is the length of the signal (*e.g.*, the number of pixels in a block or an image), M is the number of measurements, *i.e.*, the number of acquired samples, and $M < N$.

Nevertheless, to keep the complexity of a CS-based video encoder be low, a unique characteristic is that CS can "directly" capture compressed video data without temporally storing the raw data. Hence, it is hard to accurately perform measurement rate allocation for each block without accessing the raw data. To the best of our knowledge, this issue was only roughly mentioned in the compressive video sensing framework[8], where each block is determined to be either sparse or non-sparse by predicting the sparsity based on the previous reference frame (or key frame) being conventionally/fully sampled and transformed using the block-based discrete cosine transform (DCT). Each sparse block is compressively sampled whereas each non-sparse block is fully sampled. The major problems of this approach include: (i) it is required to periodically support fully sampled reference frame whose raw data are needed to be temporally stored and some transformation operation performed is required, which indeed violate the original intention of CS-based data compression and cannot be compatible with the CS-based single-pixel camera[5]; and (ii) the measurement rate allocation is too rough to only allocate either a certain rate or full rate for each block.

In this paper, we propose a novel block-based distributed compressive video sensing (DCVS) framework with feedback channel supported, which is extended from our recently developed global dictionary-based DCVS[15]. We focus on studying dynamic measurement rate allocation for DCVS, which can adaptively adjust measurement rates by estimating the sparsity of each block via feedback information. Note that the support of feedback channel is usually a common assumption in most DVC researches[1]. The major characteristics of our DCVS include: **(i) Dynamic measurement rate allocation:** the target average measurement rate for each frame can be properly allocated to each block in the frame based on the estimated sparsity via feedback information. **(ii) CS-based single-pixel camera-compatible:** only CS random projection process is individually performed for each frame or each block, which is compatible with the single-pixel camera architecture[5]. In the frameworks[11,12], it is required to support standard MPEG-X/H.26X intra-frame encoder to encode each key frame (similar to conventional I frame), which is more complex and incompatible with the single-pixel camera architecture[5]. **(iii) Global-dictionary based sparse representation:** to reconstruct a frame, a global dictionary, trained from a set of blocks extracted from the neighboring reconstructed frames together with the side information generated from them, is used as the basis of each block. The major advantages of our DCVS include: (a) more efficient utilization of available measurement rates; (b) the basis for a frame can be adaptively constructed based on neighboring reconstructed frames, which is better than using fixed basis (*e.g.*, DWT or DCT basis); (c) extracting more blocks globally for dictionary training can provide better basis for representing blocks with large motions; and (d) even if the qualities of the training blocks from neighboring frames are not good enough, the trained dictionary may still provide good basis for the blocks in a frame. The fact can be similarly explained by dictionary-based image denoising based on the dictionary trained from the blocks extracted from a noisy image itself[16,17]. In the works[11,12], for each block in a CS frame, a set of local (spatially neighboring) blocks are extracted from the neighboring reconstructed key frames to form its basis without training. Such local dictionary-based basis may not work very well for block with (very) large motion. In addition, such schemes highly rely on the qualities of neighboring reconstructed key frames. The performance may be degraded due to poorly reconstructed neighboring key frames. Other technical comparisons can be found in Table 1 of Sec. 4. An additional property is the inherent computational secrecy of measurements[18] which can be only reconstructed at the decoder via the same secret key for constructing measurement matrix as the one used in random projection (data

acquisition) at the encoder, and, hence, our previous CS-based image security technology[19] can be directly applied to support the security of our DCVS.

The rest of this paper is organized as follows. The overviews of distributed video coding (DVC), compressive sensing (CS), and sparse representation are given in Sec. 2. The proposed dynamic measurement rate allocation for our block-based DCVS with feedback channel is described in Sec. 3. Simulation results are presented in Sec. 4, followed by conclusions in Sec. 5.

2. BACKGROUND

2.1 Distributed video coding

In distributed video coding (DVC)[1], the statistical dependency between a frame W and its side information I is modeled as a virtual correlation channel, where I can be viewed as a noisy version of W. At the encoder, without performing motion estimation, the compression of W can be achieved by transmitting only part of the parity bits derived from the channel-encoded version of W. The decoder uses the received parity bits and the side information I derived from previous decoded frames to perform channel decoding to correct some "errors" in I for the reconstruction of W. In our DCVS, the side information for a CS frame is incorporated in training dictionary (basis) for this frame.

2.2 Compressive sensing

Assume that an orthonormal basis matrix (dictionary) Ψ with size $N \times N$ can provide a K sparse representation for a real value signal x with length N, *i.e.*, $x = \Psi\theta$, where θ with length N can be well approximated using only $K << N$ non-zero entries. Compressive sensing (CS)[3] states that x can be accurately reconstructed by taking only

$$M = \mathrm{O}\left(K \times \log\left(\frac{N}{K} \right) \right), \tag{2}$$

where $K < M << N$, linear and non-adaptive measurements from the random projection as

$$y = \Phi x, \tag{3}$$

where y is a measurement vector with length M, Φ is an $M \times N$ measurement matrix that is incoherent with Ψ. More specifically, the M measurements in y are random linear combinations of the entries of x, which can be viewed as the compressed version of x. The reconstruction of x can be formulated as a convex unconstrained optimization problem described in Sec. 3.1.

In our DCVS, the dictionary Ψ trained from selected blocks for each CS frame is an overcomplete learned dictionary[16], not orthonormal, and, hence, the CS theory cannot be entirely applied[20]. However, by using the measurement matrix Φ[6] randomly generated from some distribution, the incoherence between Φ and Ψ should be usually high enough[11,12].

2.3 Sparse representation

Given an overcomplete dictionary $\mathrm{D} = \left\{ \left[\mathrm{d}_p \right]_{N\times 1} \right\}_{p=1,2,\ldots,P} \in \mathrm{R}^{N\times P}, N \le P$, containing P prototype signal atoms $\left[\mathrm{d}_p \right]$, a signal $\mathrm{u} \in \mathrm{R}^N$ can be represented as a sparse linear combination of these atoms, which is $\left\| \mathrm{u} - \mathrm{D}\alpha \right\|_2 \le \varepsilon$, where $\alpha \in \mathrm{R}^P$ is the sparse representation coefficients of u and $\varepsilon \ge 0$ is an error tolerance. The sparsest representation α can be solved as[16]:

$$\min_{\alpha} \left\| \alpha \right\|_0 \text{ subject to } \left\| \mathrm{u} - \mathrm{D}\alpha \right\|_2 \le \varepsilon, \tag{4}$$

where $||\alpha||_0$ is the l_0 norm of α, counting the number of nonzero coefficients of α.

3. PROPOSED DYNAMIC MEASUREMENT RATE ALLOCATION FOR OUR DCVS

3.1 Problem formulation

In our DCVS shown in Figure 1, a video sequence consists of several GOPs (group of pictures), where a GOP consists of a key frame followed by some CS frames. At DCVS encoder, given a target average measurement rate, we want to

perform optimal measurement rate allocation to each frame (or block) before performing random projection (data acquisition). Then, each frame (or block) x can be compressed via random projection to get its measurement vector $y = \Phi x$, where Φ is a measurement matrix[6]. At DCVS decoder, the reconstruction of x from y and Φ can be formulated as:

$$\min_{\theta} \frac{1}{2}\|y - A\theta\|_2^2 + \tau\|\theta\|_1, \tag{5}$$

where θ is a set of sparse coefficients with respect to a basis Ψ that is incoherent with Φ, $x = \Psi\theta$, $A = \Phi\Psi$, τ is a non-negative parameter, $\|v\|_2$ is the ℓ_2 norm of v, and $\|v\|_1$ is the ℓ_1 norm of v. Eq. (5) indicates a convex unconstrained optimization problem, which can be solved via certain iterative algorithm[7]. For reconstructing various types of frames, different basis functions Ψ or trained dictionaries will be employed, as described later in Secs. 3.3~3.5.

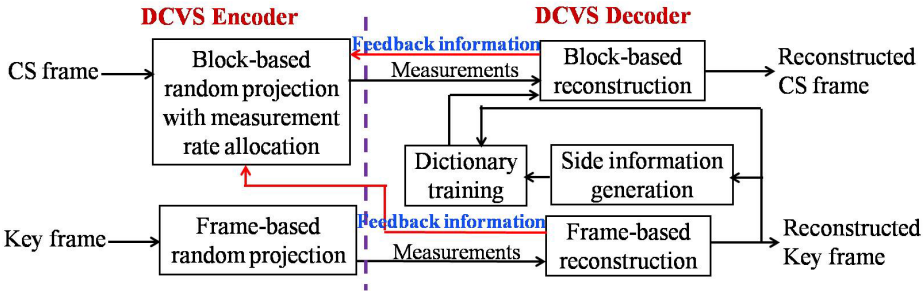

Figure 1. Proposed DCVS with dynamic measurement rate allocation.

3.2 DCVS encoder with dynamic measurement rate allocation

At DCVS encoder shown in Figure 1, without performing motion estimation, each key frame x_t viewed as a column vector with length N is compressed via frame-based random projection as $y_t = \Phi x_t$, where y_t is the measurement vector with length M_t, $M_t < N$, forming the compressed version of x_t, which will be transmitted to the decoder. Φ is an $M_t \times N$ measurement matrix[6] described later. Given a target average measurement rate MR_{ave}, we simply set the measurement rate MR_t of each key frame x_t to MR_{ave}. Hence, the number of measurements of a key frame x_t is $M_t = N \times MR_{ave}$.

On the other hand, each CS frame x_t consisting of B non-overlapping blocks, b_{ti}, $i = 1, 2, \ldots, B$, is compressed via block-based random projection by individually projecting each b_{ti} viewed as a column vector with length N_b via $y_{ti} = \Phi b_{ti}$, where y_{ti} is the measurement vector with length M_{ti}, $M_{ti} < N_b$, and Φ is an $M_{ti} \times N_b$ measurement matrix[6]. The vectors y_{ti}, $i = 1, 2, \ldots, B$, forming the compressed version of x_t, will be transmitted to the decoder.

Similar to key frame, we set the measurement rate MR_t of a CS frame x_t to MR_{ave} which will be adaptively allocated to each block b_{ti} in the frame based on its estimated sparsity via feedback information. Recall from Eq. (2) that the number of required measurements for reconstructing a block highly depends on the sparsity of the block. Hence, sparser blocks need fewer measurements whereas less sparse blocks need more measurements. Nevertheless, at the encoder, no raw block data can be available and the basis for a CS frame cannot be known which is adaptively constructed at the decoder. Hence, we propose to estimate the sparsity of a block based on its spatially co-located block in the previous reconstructed frame at the decoder. Then, the estimated number of measurements for compressively sampling current block can be obtained from the feedback information, addressed in Sec. 3.6.

Here, the used measurement matrix Φ is the scrambled block Hadamard ensemble (SBHE) matrix[6], which takes the partial block Hadamard transform, followed by randomly permuting its columns. SBHE has been shown to satisfy the five requirements, including near optimal performance, university, fast computation, memory efficient, and hardware friendly. Therefore, it can be seen from Figure 1 that our DCVS encoder is indeed memory and computation efficient.

3.3 DCVS decoder for key frame reconstruction

At DCVS decoder, each key frame x_t can be reconstructed via solving the convex unconstrained optimization problem described in Eq. (5) as:

$$\min_{\theta_t} \frac{1}{2}\|y_t - A\theta_t\|_2^2 + \tau\|\theta_t\|_1, \tag{6}$$

where y_t is the received measurement vector, $y_t = \Phi x_t$, $A = \Phi\Psi$, Φ is the SBHE measurement matrix[6], Ψ is the DWT basis, θ_t is the sparse coefficients to be solved for x_t with respect to Ψ, and τ is a non-negative parameter. In DCVS, θ_t is solved via the "sparse reconstruction by separable approximation (SpaRSA)" algorithm[7] due to its superior efficiency. Other algorithms solving convex optimization problem or iterative greedy algorithms can also be employed. Finally, the key frame x_t can be reconstructed via $\tilde{x}_t = \Psi\tilde{\theta}_t$, where $\tilde{\theta}_t$ is the final solution obtained by SpaRSA. For individual reconstruction of a key frame, a general-purpose basis, DWT basis, for image representation is employed.

3.4 DCVS decoder for CS frame reconstruction

At DCVS decoder, each CS frame x_t can also be reconstructed via solving the convex unconstrained optimization problem for each block b_{ti}, $i = 1, 2, \ldots, B$, in x_t as

$$\min_{\alpha_{ti}} \frac{1}{2}\|y_{ti} - A_t\alpha_{ti}\|_2^2 + \tau\|\alpha_{ti}\|_1, \tag{7}$$

where y_{ti} is the received measurement vector with length M_{ti} for the block b_{ti}, viewed as a column vector with length N_b, $y_{ti} = \Phi b_{ti}$, $A_t = \Phi D_t$, Φ is the SBHE measurement matrix[6] with size $M_{ti}\times N_b$, D_t is the trained dictionary with size $N_b\times P$, $N_b \le P$, for x_t, described in Sec. 3.5, α_{ti} is the sparse coefficient vector with length P to be solved for b_{ti} with respect to the basis D_t, and τ is a non-negative parameter. Similarly, b_{ti} can be reconstructed via $\tilde{b}_{ti} = D_t\tilde{\alpha}_{ti}$, where $\tilde{\alpha}_{ti}$ is the final solution obtained by SpaRSA. That is, each block b_{ti} in x_t can be represented as a linear combination of the atoms (column vectors) in D_t. Finally, the CS frame x_t can be reconstructed by integrating $\tilde{b}_{ti}$, $i = 1, 2, \ldots, B$.

3.5 Dictionary training for CS frame reconstruction

If the basis for an image can be created based on the atoms of the image itself, this basis should provide much sparser representation for the image. Although, it is impossible to get the basis created from an image itself to be reconstructed at decoder, based on the general fact that the image contents of successive frames in the same scene of a video should be similar, a frame can be well-predicted based on its side information generated from the interpolation of its neighboring reconstructed frames, which has been successfully employed in DVC[1].

At DCVS decoder, for a CS frame x_t, its side information I_t can be generated from the motion-compensated interpolation of its previous and next reconstructed key frames, respectively, denoted by x_{t-j} and x_{t+j}. Then, we use the three frames, x_{t-j}, I_t, and x_{t+j} to train the dictionary (basis) for this CS frame x_t as follows. First, we extract Q training patches $u_i \in R^{N_b}$, $i = 1, 2, \ldots, Q$, from x_{t-j}, I_t, and x_{t+j}, where each frame is divided into several non-overlapping blocks. For each non-overlapping block in the three frames, we extract the 9 training patches including the nearest 8 blocks overlapping this block and this block itself, where each extracted patch can be viewed as a column vector with length N_b. Second, we apply the K-SVD algorithm[16] to these Q training patches to train the dictionary D_t with size $N_b\times P$, $N_b \le P$, for x_t, where D_t is an overcomplete dictionary containing P atoms. With respect to D_t, each block b_{ti} in x_t can be represented as a sparse coefficient vector α_{ti} whose length P is larger than or equal to that (N_b) of b_{ti}, but α_{ti} is usually very sparse, *i.e.*, $\|\alpha_{ti}\|_0 << N_b$. Using the trained dictionary for all the blocks of a CS frame can usually provide sparser representation for the frame than using a fixed DWT basis. The block diagram of our DCVS can be illustrated in Figure 2.

An illustrative example of the *Foreman* QCIF video sequence at measurement rate (*MR*, defined in Eq. (1)) = 0.3 shown in Figure 3 is used to demonstrate the efficiency of DCVS decoder, where the parameter settings are described in Sec. 4. Figure 3(a) and (b) show, respectively, an original CS frame (the 32nd frame), and its dictionary with size 256×256, where each atom (column vector) with length 256 in the dictionary is displayed as a block. Figure 3(c) and (d), respectively, show the reconstructed CS frame using the dictionary shown in Figure 3(b) and the frame-based DWT basis (treat this frame as a key frame). It can be observed from Figure 3 that using the trained dictionary can provide better CS frame reconstruction than using the DWT basis at the same *MR*.

3.6 Feedback information for dynamic measurement rate allocation

After reconstructing a key frame x_t, we want to exploit the sparsity of each block in x_t to estimate the sparsity of the spatially co-located block in the next CS frame x_{t+1} which will be immediately encoded at the encoder. Nevertheless, it should be noted that the basis (fixed DWT basis) of x_t is different from that (trained dictionary) of x_{t+1}. Hence, it is desired to find the sparse representation of each block in x_t with respect to the basis of x_{t+1}. Based on the assumption that two successive frames in a video should be similar, the sparsity with respect to the same basis of each corresponding pair

of blocks in the two frames should also be similar. Because the training basis of x_{t+1} depends on x_t and its succeeding key frame that is unavailable before compressing x_{t+1}, we use the dictionary trained in the previous GOP, which has existed at the decoder, to simulate the basis D_t of x_t and find the sparse representation of each block b_{ti} with respect to D_t by solving Eq. (4) to get α_{ti}. The simulated basis D_t should be similar to the real basis D_{t+1} used for reconstructing x_{t+1} to some extent if GOP size is small enough. We are also currently investigating the achievable performance by comparing with the performance upper bound when the next key frame is assumed to be available. Then, we use the sparse representation α_{ti} of b_{ti} to predict that ($\alpha_{(t+1),i}$) of the spatially co-located block $b_{(t+1),i}$ in x_{t+1}, $i = 1,2,\ldots,B$. Actually, it is not easy to use the number of nonzero coefficients (obtained by performing some CS reconstruction algorithm) of the sparse representation of a block to estimate its real sparsity. Alternately, based on the fact that the complexity and sparsity of an image are highly correlated[21], we propose to exploit the variance of the coefficients of each block to perform measurement rate allocation. Based on the variance of estimated $\alpha_{(t+1),i}$, denoted by $v_{(t+1),i}$, for $b_{(t+1),i}$ and the target measurement rate MR_{t+1} of x_{t+1}, we allocate the number of measurements for each block $b_{(t+1),i}$ as

$$M_{(t+1),i} = \frac{v_{(t+1),i}}{\sum_{j=1}^{B} v_{(t+1),j}} \times \left(MR_{t+1} \times N\right), \tag{8}$$

where N is the frame size. The allocation strategy implies that more complex (less sparse) blocks will be allocated more measurements, and vice versa. Then, the information including $M_{(t+1),i}$ will be sent back to the encoder via the feedback channel for compressively sampling x_{t+1}.

Furthermore, after reconstructing a CS frame x_t, if its next frame x_{t+1} is also a CS frame, we just use the variance of the coefficients of each block in x_t to estimate that of the spatially co-located block in x_{t+1} because the bases of the two CS frames in the same GOP are identical. Then, the measurements rate allocation can be similarly performed using Eq. (8), which will be sent back to the encoder for compressively sampling x_{t+1}.

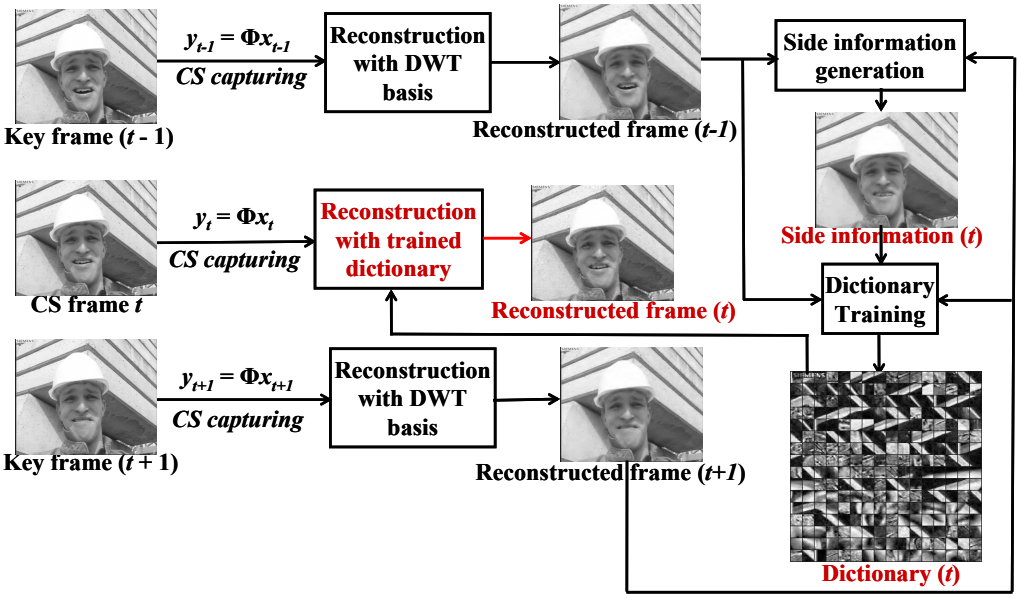

Figure 2. The block diagram of our DCVS.

4. SIMULATION RESULTS

In this paper, several QCIF (frame size: 176×144) video sequences (51 Y frames for each) with GOP size = 2, and different measurement rates (*MR*s) were employed to evaluate the proposed DVCS with dynamic measurement rate allocation (denoted by **Proposed**). For training the dictionary for each CS frame consisting of several non-overlapping 16×16 blocks, the parameter settings are described as follows. The dictionary size was set to 256×256, *i.e.*, $N_b = 16\times16 =$

256 and P = 256 (atoms). In K-SVD[16], the number of iterations for training was set to 10 while the number of nonzero coefficients used to represent each signal (block) was set to 10. Basically, the two parameters should be adjusted to adapt to the contents of video sequences, which will be a subject for future work. According to our simulations, the performances will not exhibit significant changes when the two above-mentioned parameters for K-SVD are increased, which will increase the complexity of dictionary training based on K-SVD. For SpaRSA[7], its default parameter settings were used.

(a) (b) (c) (d)

Figure 3. Comparison of the luminance (Y) components of CS frame reconstruction between trained and fixed dictionaries: (a) The original 32nd frame; (b) the trained dictionary for (a); (c) the reconstructed 32nd frame with respect to the dictionary shown in (b) (PSNR=31.49dB); and (d) the reconstructed 32nd frame with respect to the frame-based DWT basis (PSNR=27.83dB).

In this paper, three compressive video sensing schemes without measurement rate allocation were used for comparison with our global dictionary training-based DCVS scheme with measurement rate allocation. The first one is our DCVS without measurement rate allocation (denoted by "**Proposed W/O**")[15]. The second one is a "**Frame-DWT**[6]" scheme. Under our DCVS architecture, all frames are treated as key frame (reconstructed with respect to the frame-based DWT basis). The third one is a "**Local-Dict**[11-12]" scheme. Based on our DCVS architecture, each block in a CS frame is reconstructed with respect to its corresponding local dictionary-based basis similar to the major core in the works[11-12]. Here, based on the work[11], the dictionary of each block in a CS frame includes the blocks extracted from the two spatially corresponding square 17×17 windows, respectively, in the two neighboring reconstructed key frames without needing dictionary training. The characteristics of our DCVS with measurement rate allocation (**Proposed**) and the Local-Dict schemes[11-12] are summarized in Table 1. Please note that we only implemented the major core of the schemes[11-12] instead of the full system for comparison. In the simulations, the key frames with the same index in a certain simulation of the three schemes are all kept to be the same. That is, the three schemes mentioned above will exhibit different capabilities to affect the qualities of CS frames. Currently, additional complexity for dictionary training is required for each CS frame, which is, however, usually acceptable in a DVC scenario supporting a high-complexity decoder, which may be further reduced for future work.

Table 1. Comparisons of the proposed DCVS and Local-Dict schemes.

Scheme	**Proposed**	**Local-Dict**[11-12]
Ingredients of dictionary	Training based on the extracted blocks from neighboring key frames and side information	Spatially neighboring blocks from neighboring key frames without training
Dictionary size	256 atoms	Spatially corresponding square window size × Number of neighboring key frames (17×17×2 = 578 atoms)
Number of dictionaries per CS frame	1	Number of blocks per CS frame (99 dictionaries for a QCIF CS frame)
Dictionary type	Global	Local
Decoding complexity per CS frame	Dictionary training by K-SVD + Sparse decoding for 256 coefficients per block	Sparse decoding for 578 coefficients per block
Measurement rate allocation	Yes	No

The average PSNR (dB) performances of CS frames at different *MR*s for the *News*, *Foreman*, and *Football* video sequences are shown in Tables 2-4, respectively, where it can be observed that the PSNR performances of the proposed DCVS can outperform the three schemes for comparison, especially at lower *MR*s and for large-motion sequences. In our scheme (**Proposed**), available measurement rates can be more efficiently utilized. It can also be observed from Table 4 that the PSNR performances obtained from the four evaluated schemes are somewhat poor (< 25 dB). The major reasons include: (i) the frame contents of the *Football* sequence are somewhat complex, which may not be exactly sparse signals with respect to most bases, and (ii) the motions of the sequence are very large so that it is hard to find a good dictionary for a CS frame from its neighboring key frames. It is worth noting that the dictionary training of our DCVS can reveal some "denoising" capability to obtain a basis better than that of the Local-Dict scheme[11-12] without relying on dictionary training. It should be noted that the ranges of PSNR values presented in this paper are lower than those presented in the papers[11-12]. The major reason is that in the papers[11-12], each key frame is encoded using the H.264/AVC encoder which is very efficient, but also very complex and single-pixel camera-incompatible, resulting in better basis for CS frame reconstruction, while in this paper, all frames are encoded based on compressive sensing.

Table 2. The performances of the *News* sequence.

MR(%)	10	20	30	40
Proposed	**21.01**	**24.75**	**27.43**	**28.94**
Proposed W/O[15]	16.44	23.75	26.67	28.65
Local-Dict[11-12]	15.09	22.18	25.74	28.12
Frame-DWT[6]	14.85	21.87	23.93	26.24

Table 3. The performances of the *Foreman* sequence.

MR(%)	10	20	30	40
Proposed	**23.41**	**26.33**	**28.22**	**29.92**
Proposed W/O[15]	16.98	25.90	27.87	29.68
Local-Dict[11-12]	14.80	23.94	26.82	29.40
Frame-DWT[6]	13.58	22.29	24.06	26.25

Table 4. The performances of the *Football* sequence.

MR(%)	10	20	30	40
Proposed	**20.10**	**21.63**	**23.40**	**24.95**
Proposed W/O[15]	17.11	21.08	22.53	23.85
Local-Dict[11-12]	15.08	18.45	19.47	20.72
Frame-DWT[6]	15.68	20.10	22.08	24.00

5. CONCLUSIONS

In this paper, a distributed compressive video sensing (DCVS) framework via global dictionary-based sparse coding with measurement rate allocation is proposed to directly capture compressed video for CS-based single-pixel camera architecture. The simulation results have shown that the available measurement rates can be more efficiently utilized and the trained global dictionary can provide better basis for video reconstruction than using the DWT basis and local dictionary-based basis. For the future works, several important issues need to be investigated in depth for achieving a complete CS-based video coding system are descried as follows. (i) Frame-level measurement rate allocation: The available measurements should be adaptively allocated to each frame based on its sparsity. (ii) Measurement rate allocation without needing feedback channel. (iii) Adaptive measurement matrix learning: If a measurement matrix can be adaptively learned based on the characteristics of current signal to be captured[20], the number of captured measurements should be reduced while preserving a certain performance. (iv) Measurement quantization[22]: Real measurement values should be properly quantized to get the best tradeoff between the number of quantization levels and quantization loss. (v) Bit allocation and entropy coding for measurements[22-23]. (vi) Fast dictionary training at the decoder. (vii) More efficient algorithm solving the convex optimization problem. (viii) More robust algorithm solving the convex optimization problem against quantization errors and transmission errors or other error resilience techniques. (ix) More accurate side information generation: If more accurate side information for a CS frame can be generated, the trained dictionary can provide much sparser representation for this frame, resulting in better compression performance.

On the other hand, it has been shown that the compression efficiency of CS currently cannot be comparable with traditional compression techniques[23-24]. We think the major reason is that most image/video data are not really sparse signals. That is, it is hard to find the optimal basis to represent an image or a video frame at decoder without knowing the real raw data. If the basis for an image can be created based on the atoms of the image itself, this basis should provide much sparser representation for the image. Even though it is impossible to get the basis created from an image itself to be reconstructed at decoder, our DCVS try to find good basis of the current image to be reconstructed via dictionary training for the atoms extracted from the neighboring reconstructed frames together with the generated side information. If the training samples for dictionary training can be more comprehensive (*e.g.*, with training samples extracted from more temporal and/or interview reference frames and the side information generated from them), better basis should be obtained. Hence, we believe that CS will succeed in low-complexity image/video compression if the important issues described in the previous paragraph can be well-solved.

In addition, a unique characteristic of CS is to directly capture compressed data (measurements) without temporally storing the complete raw data. This characteristic is beneficial to applications with limited resources for data acquisition[23], such as wireless sensor networks[25] and low-power mobile device. Although the process of data reconstruction from measurements is currently more computationally expensive, some applications have been shown to be accomplished in measurement domain, such as image retrieval[26] and video surveillance operations[27-28]. At meanwhile, CS can provide computational security[18] and, hence, the above-mentioned applications can be performed in secure/private domain. CS can also be suitable to applied to develop security technologies[19,29]. On the other hand, CS and sparse representation technologies have been shown to be useful in developing several image/video post-processing techniques[16,17,30-35], such as denoising, deblurring, demosaicking, enhancement, restoration, super-resolution, and inpainting, which can be used to further enhance reconstructed image quality. CS has also been applied to data transmission over networks[25,36]. For further applications, CS and sparse representation have been applied to face recognition[37] and object recognition[38]. In conclusion, compressive sensing and sparse representation technologies can be applicable to multimedia data acquisition, compression, transmission, security, post-processing, and several applications, which are worthy to be further investigated.

REFERENCES

1. Girod, B., Aaron, A., Rane, S. and Rebollo-Monedero, D., "Distributed video coding," Proceedings of the IEEE 93(1), 71-83 (2005).
2. Wyner, A. and Ziv, J., "The rate-distortion function for source coding with side information at the decoder," IEEE Trans. on Information Theory 22(1), 1-10 (1976).
3. Candes, E. J. and Wakin, M. B., "An introduction to compressive sampling," IEEE Signal Processing Magazine 25(2), 21-30 (2008).
4. Romberg, J., "Imaging via compressive sampling," IEEE Signal Processing Magazine 25(2), 14-20 (2008).
5. Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F. and Baraniuk, R. G., "Single-pixel imaging via compressive sampling," IEEE Signal Processing Magazine 25(2), 83-91 (2008).
6. Gan, L., Do, T. T. and Tran, T. D., "Fast compressive imaging using scrambled hadamard ensemble," Proc. European Signal Processing Conf., (2008) (Matlab codes available from http://thanglong.ece. jhu.edu/CS/).
7. Wright, S. J., Nowak, R. D. and Figueiredo, M. A. T., "Sparse reconstruction by separable approximation," IEEE Trans. on Signal Processing 57(7), 2479-2493 (2009) (Matlab codes available from http://www.lx.it.pt/~mtf/ SpaRSA/).
8. Stankovic, V., Stankovic, L. and Cheng, S., "Compressive video sampling," Proc. European Signal Processing Conf., (2008).
9. Kang, L. W. and Lu, C. S., "Distributed compressive video sensing," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 1169-1172 (2009).
10. Stankovic, V., Stankovic, L. and Cheng, S., "Compressive image sampling with side information," Proc. IEEE Int. Conf. on Image Processing, 3037-3040 (2009).
11. Prades-Nebot, J., Ma, Y. and Huang, T., "Distributed video coding using compressive sampling," Proc. Picture Coding Symposium, (2009).
12. Do, T. T., Chen, Y., Nguyen, D. T., Nguyen, N., Gan, L. and Tran, T. D., "Distributed compressed video sensing," Proc. IEEE Int. Conf. on Image Processing, 1393-1396 (2009).

13. Kwon, D. K., Shen M. Y. and Kuo, C.-C. Jay, "Rate control for H.264 video with enhanced rate and distortion models," IEEE Trans. on Circults and Systems for Video Technology 17(5), 517-529 (2007).
14. Brites, C. and Pereira, F., "Encoder rate control for transform domain Wyner-Ziv video coding," Proc. IEEE Int. Conf. on Image Processing, 5-7 (2007).
15. Chen, H. W., Kang, L. W. and Lu, C. S., "Distributed compressive video sensing via global dictionary-based sparse coding," submitted to Proc. IEEE Int. Conf. on Image Processing (2010).
16. Aharon, M., Elad, M. and Bruckstein, A. M., "The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation," IEEE Trans. on Signal Processing 54(11), 4311–4322 (2006) (Matlab codes available from http://www.cs.technion. ac.il/~ronrubin/software.html).
17. Elad, M. and Aharon, M., "Image denoising via sparse and redundant representations over learned dictionaries," IEEE Trans. on Image Processing 15(12), 3736–3745 (2006).
18. Rachlin, Y. and Baron, D., "The secrecy of compressed sensing measurements," Proc. Allerton Conf. on Communication, Control, and Computing, (2008).
19. Kang, L. W., Lu, C. S. and Hsu, C. Y., "Compressive sensing-based image hashing," Proc. IEEE Int. Conf. on Image Processing (2009).
20. Duarte-Carvajalino, J. M. and Sapiro, G., "Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization," IEEE Trans. on Image Processing 18(7), 1395-1408 (2009).
21. Perkio, J. and Hyvarinen, A., "Modelling image complexity by independent component analysis, with application to content-based image retrieval," Proc. Int. Conf. on Artificial Neural Networks (2009).
22. Dai, W., Pham, H. V. and Milenkovic, O., "Distortion-rate functions for quantized compressive sensing," Proc. IEEE Information Theory Workshop on Networking and Information Theory (2009).
23. Goyal, V. K., Fletcher, A. K. and Rangan, S., "Compressive sampling and lossy compression," IEEE Signal Processing Magazine 25(2), 48-56 (2008).
24. Schulz, A., Velho, L. and da Silva, E. A. B., "On the empirical rate-distortion performance of compressive sensing," Proc. IEEE Int. Conf. on Image Processing (2009).
25. Duarte, M. F., Wakin, M. B., Baron, D. and Baraniuk, R. G., "Universal distributed sensing via random projections," Proc. ACM/IEEE Int. Conf. on Information Processing in Sensor Networks, 19-21 (2006).
26. Divekar A. and Ersoy, O., "Compact storage of correlated data for content based retrieval," Proc. Asilomar Conf. on Signals, Systems and Computers (2009).
27. Cevher, V., Sankaranarayanan, A., Duarte, M. F., Reddy, D., Baraniuk, R. G. and Chellappa, R., "Compressive sensing for background subtraction," Proc. European Conf. on Computer Vision (2008).
28. Cossalter, M., Tagliasacchi, M. and Valenzise, G. "Privacy-enabled object tracking in video sequences using compressive sensing," Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (2009).
29. Tagliasacchi, M., Valenzise, G. and Tubaro S., "Hash-based identification of sparse image tampering," IEEE Trans. on Image Processing 18(11), 2491-2504 (2009).
30. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. and Yan, S., "Sparse representation for computer vision and pattern recognition," to appear in Proceedings of the IEEE.
31. Elad, M., Figueiredo, M. A. T. and Ma Y., "On the role of sparse and redundant representation in image processing," to appear in Proceedings of the IEEE.
32. Mairal, J., Elad, M. and Sapiro, G., "Sparse representation for color image restoration," IEEE Trans. on Image Processing 17(1), 53-69 (2008).
33. Mairal, J., Sapiro, G. and Elad M., "Learning multiscale sparse representations for image and video restoration," SIAM Multiscale Modeling and Simulation 7(1), 214-241 (2008).
34. Fadili, M. J., Starck, J. L. and Murtagh, F., "Inpainting and zooming using sparse representations," The Computer Journal 52(1), 64-79 (2009).
35. Yang, J., Wright, J., Huang, T. and Ma, Y., "Image super-resolution via sparse representation," to appear in IEEE Trans. on Image Processing.
36. Haupt, J., Bajwa, W. U., Rabbat, M. and Nowak, R., "Compressed sensing for networked data," IEEE Signal Processing Magazine 25(2), 92-101 (2008).
37. Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. and Ma, Y. "Robust face recognition via sparse representation," IEEE Trans. on Pattern Analysis and Machine Intelligence 31(2), 210-227 (2009).
38. Yang, A. Y., Gastpar, M., Bajcsy, R. and Sastry, S. S., "Distributed sensor perception via sparse representation," to appear in Proceedings of the IEEE.