Journal of Information Hiding and Multimedia Signal Processing ©2014 ISSN 2073-4212
Ubiquitous International Volume **5**, Number **1**, January **2014**

Adaptive Distributed Compressed Video Sensing

Xue Zhang[1,3], Anhong Wang[1*], Bing Zeng[2], and Lei Liu[3]

[1]Institute of Digital Media and Communication
Taiyuan University of Science and Technology, Taiyuan 030024, China
13112059@bjtu.edu.cn
*wah_ty@163.com

[2]The Hong Kong University of Science and Technology
Hong Kong SAR, China
eezeng@ust.hk

[3]Institute of Information Science
Beijing Jiaotong University, Beijing 100044, China
13112059@bjtu.edu.cn;12112061@bjtu.edu.cn

Received May, 2013; revised September, 2013

Abstract. *Compressed sensing is a state-of-the-art technology which can significantly reduce the number of sampled data in sparse signal acquisition. This paper studies the distributed compressed sensing (DISCOS) of video signals. To this end, we propose adaptive adjustments to the block-based (local) measurement rate, the frame-based (global) measurement rate, and the sparse dictionary size, thus forming an adaptive DISCOS scheme (aDISCOS). Two adjustments on measurement rates are based on the spatial and temporal sparsity that is obtained through an analysis on the block-type and the inter-frame motion, while the sparse dictionary size is adjusted according to the motion information. All analyses are implemented at the decoder side and the analysis results are sent back to the encoder via a feedback channel, yielding a low-complexity encoding (to meet the equirement of a distributed coding scheme). Simulation results show that the proposed aDISCOS achieves a superior rate-distortion performance as well as better visual quality, when compared with the original DISCOS scheme.*
Keywords: Compressed sensing, distributed compressed sensing, adaptive sampling, sparse dictionary

1. **Introduction.** In the up-link communication of low-power video capturing (via mobile cameras, wireless visual sensor networks, etc.) where the computing power is limited, people usually would like to design a simple encoder but leave a big complex to the decoder side. For tasks like this, the distributed video coding (DVC) [1-2] has been proposed with a combination of an independent encoder but a joint decoder applied to individual video frames. According to this framework, many computation intensive operations such as motion estimation and prediction have been shifted from encoder to decoder, thus offering a good solution to the aforementioned scenario. However, like a conventional image encoder, typical DVC encoders still need to do a large amount of computations such as intra-prediction, transform, quantization, and entropy coding.

More recently, the theory of compressed sensing (CS) has initiated a tremendous wave in the sparse signal processing community [3-6]. Owing to the sparseness − an intrinsic property in many signals in practice (including image and video signals), CS can

sample and compress a sparse signal at a sub-Nyquist rate while still enabling a nearly exact reconstruction, and based on the random measurement, it offers the better error-resilience at the same time[7-8]. In practice, the CS sampling (for data acquisition) can be implemented as simple as generating a (pseudo) random matrix and performing some multiplications, thus fitting very well to the DVC scenario.

Given the fact that the two aforementioned theories share a common principle of maintaining a low-complexity encoder, ***d***istributed ***c***ompressed ***v***ideo ***s***ensing (DCVS) integrating both DVC and CS characteristics has emerged as a new way to directly capture CS-sampled video data via random projection while performing the CS reconstruction together with exploiting correlations among successive frames at a high-complexity decoder [9-10]. In this scheme, video frames are classified into "key" frames and "CS" frames. Each key frame is coded independently (by any conventional intra-frame coding scheme); whereas each block in a CS frame is just CS-sampled at the encoder side and then reconstructed at the decoder side with respect to the basis (dictionary) formed from a set of spatially neighboring blocks of previously decoded neighboring key frames. We notice that, in the DCVS scheme, acquisition of each block/frame is always implemented at a fixed rate, without considering the diversified contents in various blocks within a frame or inter-frame correlations among frames. To overcome this drawback, DCVS schemes employing a dynamic measurement rate allocation have been proposed in [11-13]. Nevertheless, they only apply varying CS sampling rates on different blocks or CS frames, ignoring local or global information preserved by block-level and frame-level measurement, respectively.

In this paper, we follow the idea proposed recently in [9] to implement an aDISCOS scheme equipped with a feedback channel. Notice that such a feedback channel is usually a common assumption in some DVC systems [1]. In our scheme, each source video frame is compressed independently by a number of random sampling operations (each being a simple and random linear projection) so as to keep the simplicity at the encoder side. On the other hand, all analyses will be conducted at the decoder side, leading to a joint and more complicated decoding to deliver a higher performance. Compared with the original DISCOS work [9], our contributions in this paper are summarized as follows.

- The state-of-the-art DISCOS scheme employs a fixed measurement (or sampling) rate in both block-level measurement and frame-level measurement for each CS-frame, and the sparse dictionary also keeps a constant size, which ignores the diversified contents in various blocks within a frame as well as temporal variations among frames. In this paper, we propose to adjust adaptively the block-based (local) and frame-based (global) measurement rates, as well as the sparse dictionary size in order to produce a better coding performance.
- The actual block-based and frame-based measurement rates are determined in our paper by estimating spatial and temporal sparsity, which are obtained at the decoder through some analyses on block type and inter-frame motion, respectively. The sparse dictionary size is also adjusted adaptively depending on the motion information.

2. **The DISCOS framework.** The DISCOS framework proposed in [9] is shown in Figure. 1. A source video sequence is divided into several GOPs (group of pictures), where a GOP consists of a key-frame followed by some CS-frames. Each key-frame is intra-coded by a conventional video coding method (such as MPEG or H.26x). CS-frames are compressively sampled by using two kinds of measurements, block-based (local) and frame-based (global) ones, and all measured data are transmitted to the decoder. The frame-based measurements is similar to the generic CS coding, i.e., each frame $\mathbf{p}_t$ (of size

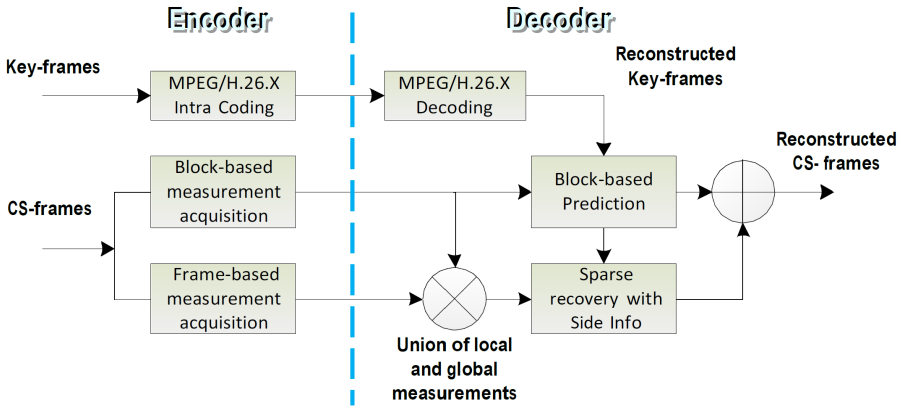


FIGURE 1. The DISCOS framework in [9]

N×N, t denoting the time) is first vectorized as $\mathbf{x}_t$ (with height N^2) and then compressed via a CS-sampling process as:

$$y_t = \Phi x_t \tag{1}$$

where $\mathbf{y}_t$ denotes the output measurement-vector of length M_t, Φ represents the $M_t \times N^2$ measurement (or sampling) matrix generated by the method of structurally random matrices (SRMs) [14]. The measurement rate for $\mathbf{x}_t$ is denoted as $R_t = M_t/N^2$.

The block-based measurements are also exploited to preserve local information that helps the decoder construct more accurate side information (SI) in DISCOS. Each CS-frame is first partitioned into non-overlapped blocks of size B × B. Then, each vectorized block x_i (where i stands for the block's index) is sampled with the same CS operator as:

$$y_i = \Phi_B x_i \tag{2}$$

where Φ_B is the measurement matrix. The equivalent sampling operator Φ appeared in Eq. (1) for the whole frame is a block-wise diagonal matrix composed by Φ_B.

At the decoder side, an independent reconstruction by using the necessary video decoding method is first carried out for all key-frames, while the reconstruction for CS-frames are much more complicated. As shown in Figure.1, block-based prediction is first achieved: each block in a frame is reconstructed via solving an l_1 minimization problem as:

$$\hat{\alpha}_i = \arg\min \|\alpha_i\|_1 \quad s.t. \quad y_i = \Phi_B \Psi_i \alpha_i \tag{3}$$

where y_i is obtained from (2), Ψ_i is a sparse basis matrix which can provide a sparse representation for x_i, i.e., $x_i = \Psi_i \cdot \alpha_i$. Instead of using a fixed linear transform (e.g., the block DCT), DISCOS uses a dictionary formed from a set of spatially neighboring blocks of previously decoded neighboring key-frames as the sparsifying matrix Ψ_i. Block-based prediction uses the sparsity adaptive matching pursuit (SAMP) [15] reconstruction algorithm to solve the l1-minimization. Then, DISCOS employs a sparse recovery with SI from its global measurements and its local block-based prediction to jointly reconstruct a CS-frame: to substract the measurement vector of an original CS-frame from that of a block-based prediction frame to form a new measurement vector of the prediction error. Finally, the CS-frame is recovered by adding the prediction error to the prediction frame, and the gradient projection for sparse reconstruction (GPSR) algorithm [16] is used.

3. **Adaptive distributed compressed video sensing (*a*DISCOS).** Both block-based measurements and frame-based measurements of each CS-frame employed a fixed measurement rate in the DISCOS scheme [9]. Apparently, it has ignored the diversified contents in various blocks (i.e., spatial sparsity) and inter-frame variations (i.e., temporal sparsity) within a video sequence. According to the CS theory, the measurement rate for a frame (or a block) can be made smaller when the temporal (or spatial) sparsity is larger, and vice versa. In addition, each block in a CS frame is reconstructed in [9] with respect to the dictionary formed from a set of spatially neighboring blocks of previously decoded neighboring key-frames. It implies that the block-based prediction quality will change with the different correlation between the preceding and following key-frames. When motion is high (generally, the correlation is low), one needs a bigger-size dictionary in order to exploit the correlation accurately. While in the case of low motion (high temporal correlation), the dictionary should be set smaller in order to save the computational complexity. Unfortunately, a fixed-size dictionary has been adopted in DISCOS, which ignored the diversity of temporal correlation and therefore the reconstruction quality would sacrifice to a certain extent.

To determine appropriate block-based and frame-based measurement rates for each CS-frame as well as a suitable sparse dictionary size for each block, one needs to carry out some analyses on the spatial and temporal sparsity. Nevertheless, one must note that these analyses should not be done at the encoder side: (1) in a practical CS scenario, raw video data not available because each original video frame stands only in the real world and (2) it would otherwise defeat the purpose of maintaining a low-complexity at the encoder side. To this end, this paper proposes to do such analyses at the decoder side. Especially, block classification and inter-frame motion analysis are performed to respectively estimate the spatial and temporal sparsity.

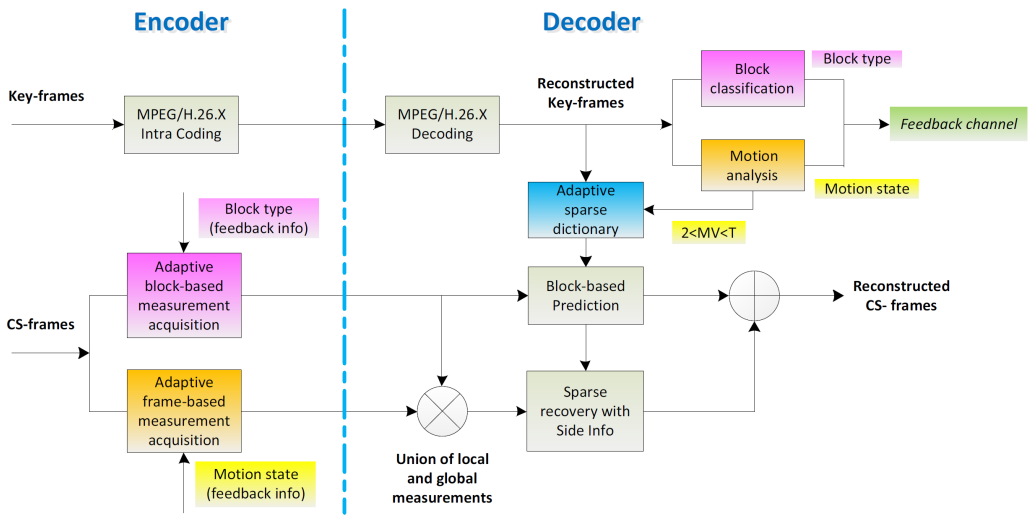


FIGURE 2. Our proposed *a*DISCOS scheme.

Now, let's present our proposed aDISCOS scheme. As shown in Figure. 2, it consists of an encoder with low-complexity and a decoder with high-complexity. When compared to the original DISCOS scheme, a major change happens at the CS-sampling as well as the decoding process of each CS-frame. In particular, the latter change makes the decoder in the current framework even more complicated, resulting in some delays if applied to a limited-power decoder.

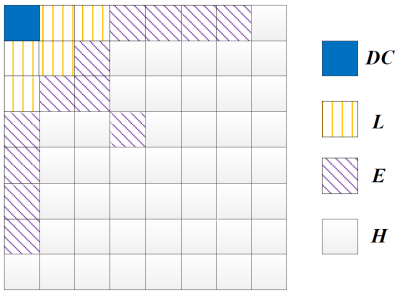


FIGURE 3. Block classification employed in our work: four indicative areas (with different marks) are fixed for all blocks.

3.1. **Block-level processing.**

A. *Block classification.* Generally, spatial sparsity of a video frame is highly correlated to the block style; thus a block classification model based on the DCT coefficients proposed in [17] is employed in our work. Here, a CS-frame is divided into 8×8 blocks and each block is DCT-transformed. Each block of DCT coefficients is divided into four indicative areas as shown in Figure. 3, where the absolute sums of the DCT values in these four areas are denoted as DC, L (low frequency), E (edge), and H (high frequency), respectively. Then, each block is assigned to the PLAIN, EDGE, or TEXTURE class according to relations between the values of L, E, and H, and some pre-determined thresholds, see [17] for the details.

B. *Dynamic bit-allocation for blocks.* As depicted in Figure. 2, each CS-frame goes through adaptive block-based measurements by estimating block type of each block. Based on the assumption that two successive frames in a video should be similar, the styles of each corresponding pair of blocks in the two frames should also be similar. Therefore, after reconstructing key-frames, we perform the block classification on them, and exploit the block type in key-frames to estimate the block co-located in the preceding and following frames which will be immediately encoded at the encoder. Then, we take advantage of a feedback channel to send the classification result back to the encoder. According to the estimated block type, bits are allocated dynamically to various blocks, which follow the ordering: "TEXTURE" (less sparse) > "EDGE" > "PLAIN" (more sparse).

3.2. **Frame-level processing.**

A. *Motion analysis.* In order to assess temporal sparsity within neighboring key-frames, we proposed to perform an inter-frame motion analysis. Here, the conventional block matching motion estimation is first applied to the current key-frame p and the referenced key-frame p_{ref} (i.e., preceding key-frame), and motion vectors (MVs) of all blocks can be obtained, e.g., $(\Delta x_i,\Delta y_i)$ is the MV of block i. Then, we extract the maximum absolute value component, i.e., $V_{i\max} = \max(|\Delta x_i|, |\Delta y_i|)$. Given an integer T (T=7 in our experiments), we count how many blocks meet imax $V_{i\max} > T-$denoted as n. Based on n, one can determine the motion state of the current key-frame p:

$$motionstate = \begin{cases} 1 & \text{if } n \geq K \\ 0 & \text{if } n < K \end{cases} \tag{4}$$

where the threshold K will be determined by experiments and related to the movement existed in source video sequence . When the *motion state* is 1, it means

that the motion of the current key-frame is high and thus the correlation with its preceding key-frame is low.

B. *Dynamic bit-allocation for blocks.* At the decoder side, the inter-frame motion analysis is performed using the reconstructed neighboring key-frames, and the resulted *motion state* will be sent back to the encoder side via a feedback channel, as shown in Figure. 2. According to the *motion state*, the bit-allocation for frames will have dynamic characteristics as follows: if the *motion state* of the current key-frame is 1, it means that a potential scene change exists - denoted as the large-motion GOP. In such a scenario, all CS-frames need to be sampled at a higher measurement rate. In addition, experiments tell us that: global sampling at low rates often lacks of the effectiveness. As a result, when a CS-frame in a GOP without large changes is sampled at a lower rate, frame-based measurement will be ignored completely, thus only keeping the block-based measurements, in order to reduce the computational burden.

3.3. **Adaptive sparse dictionary.** Through motion analysis on the reconstructed neighboring key-frames at decoder, the maximum absolute value component $V_{i\max}$ of each block in key-frames will be obtained. Then, V is calculated as:

$$V = \begin{cases} 2 & if \quad V_{i\max} \leq 1 \\ V_{i\max} & if \quad 1 < V_{i\max} \leq T \\ T & if \quad V_{i\max} > T \end{cases} \tag{5}$$

Based on V of each block in the current key-frame, one can adaptively adjust the sparse dictionary size used for the recovery of co-located blocks in CS-frames from the previous GOP. The generation of adaptive sparse dictionary is depicted in Figure. 4. Obviously, V controls the search window size and the sparsifying matrix ψ_i is an adaptive dictionary combined by vectorized blocks in neighboring key-frames. Experimental results demonstrated that our adaptive sparse dictionary scheme is powerful than the fixed dictionary to some extent and reduces the computational complexity.

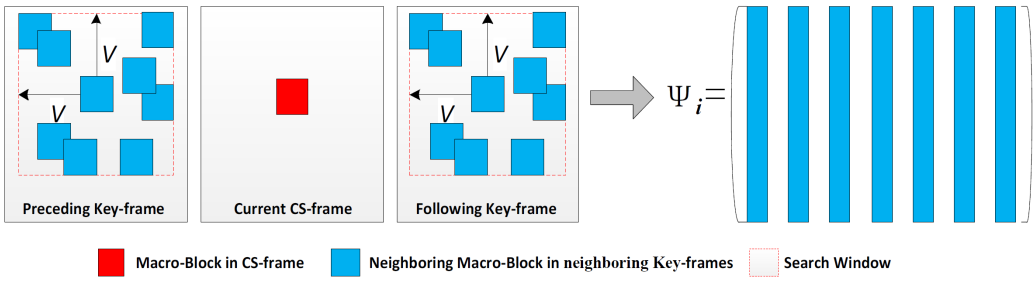


FIGURE 4. The generation of adaptive sparse dictionary for a block i in a CS-frame, it assumes that the block can be predicted using a linear combination of (vectorized) neighboring blocks in preceding and following key-frames.

4. **Simulation results.** We choose the DISCOS method proposed in [9] as the comparison benchmark. However, we modify the block size to 8×8. Notice that it usually leads to a better performance if a larger block size is used (e.g., 32×32 is used in [9]). For all experimental results presented below, we tried to maintain all selections as simple as possible so as to focus on solely demonstrating the effectiveness of adopting a dynamic bit-allocation strategy. The test signals are the first 100 frames of three CIF (frame size:

352×288, luminance only) video sequences: *Foreman*, *CoastGuard* and *News*, with GOP size=4.

Since there is no difference on each key-frame between the existing DISCOS method and our scheme, the following comparison focuses on the non-key frames.

- We implement our proposed *a*DISCOS scheme first, in which different measurement rate (MR) and variable sparse dictionary size are employed. We book-keep the total rate consumed in sampling each CS-frame, including block-based and frame-based measurement rate. We take average over all CS-frames to obtain an
- According to the obtained EMR, we then implement the DISCOS scheme proposed in [9] (i.e., each CS-frame is applied both local block-based and global frame-based measurements at a fixed rate exactly equal to EMR/2) to facilitate a fair comparison.

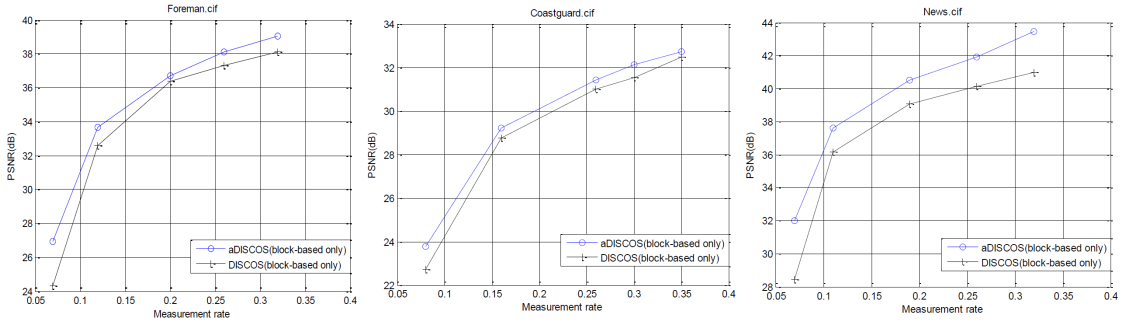
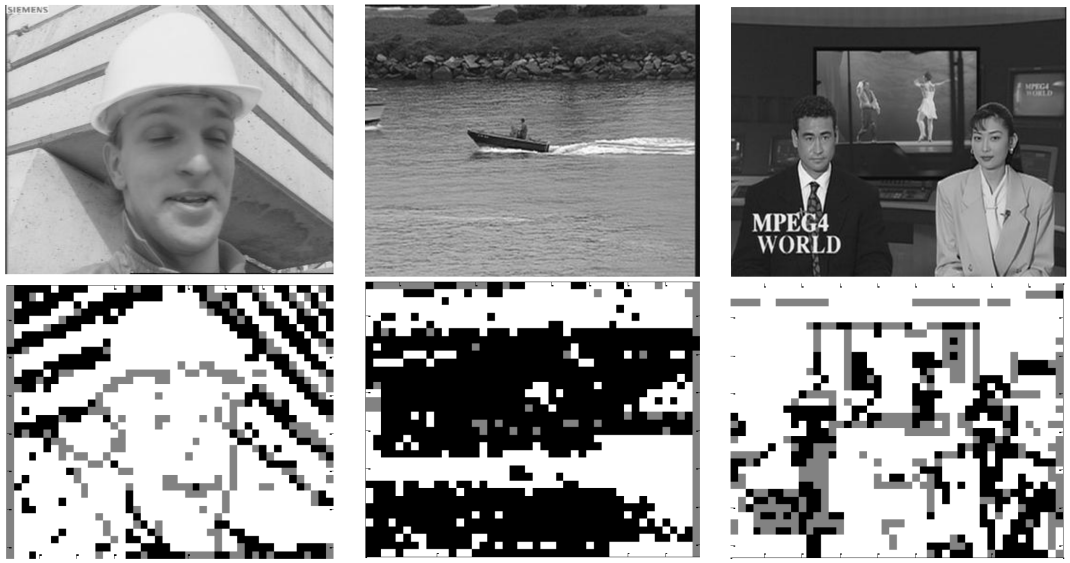


FIGURE 5. Comparison of our *a*DISCOS and fixed-rate DISCOS with only block-based measurements.



(a) (b) (c)

FIGURE 6. Original frames and their block classification maps for the first frame of three test video sequence ("white" for PLAIN, "black" for TEXTURE, and "grey" for EDGE).

Figure. 5 shows the comparison of our *a*DISCOS and the fixed-rate DISCOS [9] with only block-based measurements at five quality levels. The numerical values on the x-axis denote the MR while those on the y-axis represent the average reconstruction quality (PSNR in dB) of CS-frames. From the curve, we can find that our adaptive block-based

method yields a significant improvement for *Foreman*, *Coast-Guard* and *News* (1.15 dB, 0.56 dB, and 2.14 dB on average, respectively) due to the consideration of the spatial sparsity. Here, the block types of the first two CS-frames in a GOP follow that of the preceding key-frame; while the third CS-frame follows the next key-frame. The block-type maps for the first frame of three test video sequences are presented in Figure. 6.

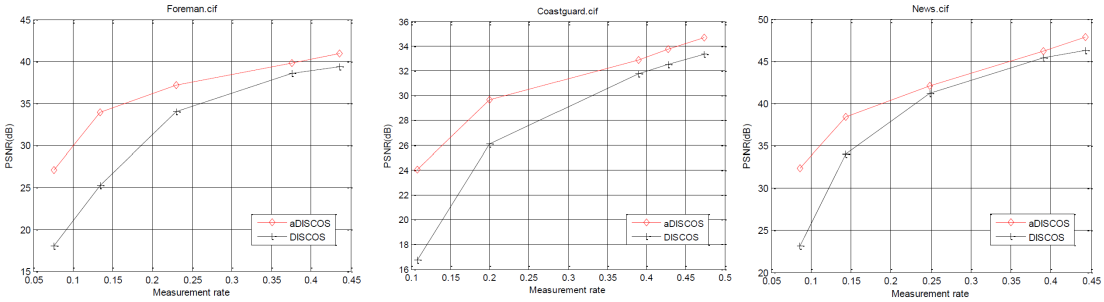


FIGURE 7. Comparison between our *a*DISCOS and the fixed-rate DISCOS.

Figure.7 shows the performance comparison between our *a*DISCOS (which considers both spatial and temporal sparsity) and the fixed-rate DISCOS [9] scheme for the same test sequences. On average, our proposed scheme achieves an improvement of about 4.72 dB, 2.91 dB, and 3.38 dB in PSNR over DISCOS with a fixed measurement rate, respectively.

To have some visual comparisons, we show in Figure. 8 some reconstructed frames for *Foreman* at EMR=0.134, where one CS-frame is from a GOP with low motion and the other is from a large-motion GOP, from which one can perceive a very noticeable improvement by using our *a*DISCOS method.

(a1) PSNR=36.33dB (a2) PSNR=25.90dB (b1) PSNR=34.93dB (b2) PSNR=24.75dB

FIGURE 8. Reconstructed frames by (1) *a*DISCOS and (2) DISCOS at the same EMR= 0.134, and (a) is from a GOP with low motion, (b) is from a GOP with large motion.

5. **Conclusions.** We introduced in this paper an adaptive distributed compressed sensing (*a*DISCOS) scheme for video signals in which both local block-based and global frame-based measurement rates as well as the sparse dictionary size can be adjusted adaptively. One unique feature is that our analyses on the spatial and temporal sparsity are carried out at the decoder side, and the analyses results are sent back to the CS encoder. Therefore, the nature of maintaining a low-complexity encoding is well preserved, which makes it very suitable for low-power mobile video capturing, such as mobile camera and wireless sensor networks. Experimental results demonstrated that the proposed *a*DISCOS clearly

outperforms the existing DISCOS scheme with a fixed measurement rate. Nevertheless, the measurement rates are chosen manually in this work, and our future works are to come up with some rules to automatically determine these numbers or formulate an optimization.

Acknowledgment. This work was supported in part by Sino-Singapore JRP(2010DFA11 010), National Natural Science Foundation of China (No. 61073142, No. 61272262, No. 61210006, No. 61272051), Doctor Startup Foundation of TYUST (No. 20092011), International Cooperative Program of Shanxi Province (No. 2011081055), The Shanxi Provincial Foundation for Leaders of Disciplines in Sci-ence (20111022)AShanxi province Talent Introduction and Development Fund (2011), Shanxi Provincial Natural Science Foundation (2012011014-3), Beijing Natural Science Foundation (No. 4102049) and New Teacher Foundation of State Education Ministry (No. 20090009120006).

REFERENCES

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

[2] M. H. Taieb, J.Y. Chouinard, D. Wang, K. Loukhaoukha, and G. Huchet, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 3. no. 1, pp. 1-11, 2012.

[3] D. L. Donoho, Compressed sensing, *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289-1306, 2006.

[4] E. J. Candes, and T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies?, *IEEE Trans. Information Theory*, vol. 52, no. 12, pp. 5406-5425, 2006.

[5] E. J. Candes, and M. B. Wakin, An introduction to compressive sampling, *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21-30, 2008.

[6] J. Romberg, Imaging via compressive sampling, *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 14-20, 2008.

[7] L. L. Tang, J. S. Pan, H. Luo, and J. B. Li, Novel watermarked MDC system based on SFQ algorithm, *IEICE trans. Communications*, vol. 95, no. 9, pp. 2922-2925, 1990.

[8] H. C. Huang, S. C. Chu, J. S. Pan, C. Y. Huang, and B. Y. Liao, Tabu search based multi-watermarks embedding algorithm with multiple description coding, *Journal of Information Sciences*, vol. 181, no, 16, pp. 3379-3396, 2011.

[9] T. T. Do, Y. Chen, D. T. Nguyen, N. Nguyen, L. Gan, and T. D. Tran, Distributed compressed video sensing, *Proc. of The 16th IEEE International Conference on Image Processing*, pp. 1381-1384, 2009.

[10] E. W. Tramel, and J. E. Fowler, Video compressed sensing with multihypothesis, *Proc. of Data Compression Conference*, pp. 193-202, 2011.

[11] J. Prades-Nebot, Y. Ma, and T. Huang, Distributed video coding using compressive sampling, *Proc. of the 27th conference on Picture Coding Symposium*, pp. 165-168, 2009.

[12] H. W. Chen, L. W. Kang, and C. S. Lu, Dynamic measurement rate allocation for distributed compressive video sensing, *Proc. of SPIE*, vol. 7744, pp. 77440I-1-77440I-10, 2010.

[13] Z. Liu, H. V. Zhao, and A. Y. Elezzabi, Block-based adaptive compressed sensing for video, *Proc of The 17th IEEE International Conference on Image Processing*, pp. 1649-1652, 2010.

[14] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, Fast and efficient compressive sensing using structurally random matrices, *IEEE Trans. Signal Processing*, vol. 60, no. 1, pp. 139-154, 2012.

[15] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, Sparsity adaptive matching pursuit for practical compressed sensing, *Proc. of The 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 581-587, 2008.

[16] 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 Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586-597, 2007.

[17] H. H. Y. Tong, and A. N. Venetsanopoulos, A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking, *Proc. of International Conference on Image Processing*, pp. 428-432, 1998.