Adaptive Coding for Compressed Video Sensing

Jian Wu[1], Yongfang Wang[1,2](✉), Yun Zhu[1], and Yuan Shuai[1]

[1] School of Communication and Information Engineering, Shanghai University,
Shanghai 200072, China
{wjgxw,yfw}@shu.edu.cn, wjtcw@hotmail.com

[2] Key Laboratory of Advanced Display and System Application,
Ministry of Education, Shanghai 200072, China

Abstract. In this paper, we propose an Adaptive Coding based Compressed Video Sensing (ACCS) scheme for Distributed Video Coding. Our scheme mimics the traditional video coding method and performs the mode decision both at the encoder and the decoder. At the encoder, the ACCS divides the frame blocks into three categories: the SKIP mode, INTER mode and COMBINED mode according to the residual of the blocks, and the adaptive sampling rate is chosen for these modes. At the decoder, we adopt different decoding methods for different modes. For the COMBINED mode, we apply adaptive decoding scheme by exploiting the intra-frame and inter-frame sparsity. Experimental results show that the proposed algorithm outperforms existing state-of-the-art video CS approaches at a very low sampling rate.

Keywords: Compressed sensing · Distributed video coding
Adaptive measurements · Sparse representations

1 Introduction

The resolution of today's video is much higher than before, and it brings huge challenges to the limited network. What's more, the attractive 3D videos have increasingly come into the public sight, which give people even better quality of experience. How to capture and compress the videos becomes a big problem. High Efficiency Video coding (HEVC) [1], as the latest video coding scheme, has a very high compression efficiency by exploiting the spatial and temporal structure for the video sequences, but it is not suited for the inexpensive video recording devices such as cellphones, wireless video cameras, which have limited computing capability and battery capacity.

Compressed sensing (CS) [2], as a novel signal processing theory, can acquire a signal at a sampling rate much lower than Nyquist rate via linear projection onto a random basis, and the original signal can be reconstructed through optimization method with high probability from some random measurements under certain conditions.

Given a signal $\mathbf{x} \in R^N$ with length N, it can be called K -sparse in a domain $\boldsymbol{\Psi}$ when K entries in its transform coefficients $\boldsymbol{\theta} = \boldsymbol{\Psi}^T\mathbf{x}$ are nonzero, and $\boldsymbol{\Psi}$ is an orthonormal basis here. CS attempts to reconstruct the K -sparse signal vector from a relatively small number of samples with linear projection $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}$, and the size of $\boldsymbol{\Phi}$ is

Y. Zhao et al. (Eds.): ICIG 2017, Part II, LNCS 10667, pp. 197–205, 2017.
https://doi.org/10.1007/978-3-319-71589-6_18

M by N. The dimension of measurement vector $\mathbf{y}$ is M, and the values of the nonzero coefficients can be well recovered if $M \ll N$.

In CS, instead of encoding all the coefficients of a signal, we only encode the M measurements and the reconstruction problem can be solved with the following l_0 minimization method

$$\min\|\boldsymbol{\Psi}\mathbf{x}\|_0, s.t. \quad \mathbf{y} = \boldsymbol{\Phi}\mathbf{x} \tag{1}$$

where $\|\mathbf{x}\|_0$ is a pseudo-norm (l_0-norm), which equals the number of nonzero elements in vector $\mathbf{x}$. Minimizing the number of nonzero entries is difficult. Instead, the optimization problem can be solved with l_1 minimization method

$$\min\|\boldsymbol{\Psi}\mathbf{x}\|_1, s.t. \quad \mathbf{y} = \boldsymbol{\Phi}\mathbf{x} \tag{2}$$

This is a convex optimization problem and can be solved easily via subspace pursuit [3] or CVX toolbox [4] and so on.

CS has a great potential in image and video applications for its low complexity and low power consumption. In CS based image compression techniques, an image is divided into n small blocks, then each one is rearranged into a vector $\mathbf{x}$. Next, the vector is sampled with the measurement matrix $\boldsymbol{\Phi}$. Finally, the original image can be reconstructed by many state-of-the-art algorithms such as BCS-SPL [5], SGSR [6], GSR [7] and so on.

As for the application of CS to video compression [8–12], the sparsity between the successive frames with the classic transform domain (e.g. DWT, DCT) is exploited based on the adjacent frames' correlation in video sequences. In [8], authors proposed a reconstruction model based on the idea that the total variation (TV) norm of the residual between the frame and its prediction, TV norm of the frame, l_0-norm of the frame in a certain transform domain are all very small. The similar idea appeared again in [9] which introduced the forward and the backward motion-compensated residuals. The support (location of large valued entries) is estimated based on the idea that the large valued entries belonging to the adjacent frames are located in almost the same place [10]. A hierarchical frame structure was proposed to exploit the correlation between the current frames and the reference frames in [11].

Unlike above methods, Distributed Compressed Video Sensing (DISCOS) framework is introduced in [13], which present the idea that the sparsest representation of a block is a linear combination of a few temporal neighboring blocks of previous reconstructed frames or nearby key frames. The same method also can be found in [14]. However, if the blocks are non-rigid objects (dancer in Fig. 1) whose shapes change a lot. And their sparsest representation is not a linear combination of a few temporal neighboring blocks, which means that it is unable to be well recovered even at a very high sampling rate by using DISCOS algorithm. As the result illustrated in Fig. 1, most of the regions in the frame are decoded perfectly, except for the dander in the background. Because the regions near the dancer change a lot and cannot be represented by the reference ones no matter how high the sampling rate is. So these blocks should be recovered by the image CS algorithms (INTRA). Moreover in [14, 15], a feedback

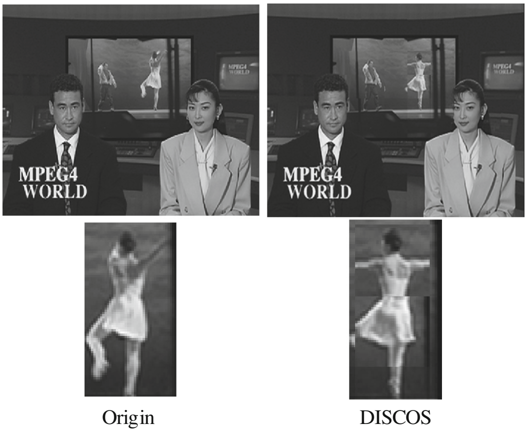


Fig. 1. The performance of DISCOS (the 5th frame and the sampling rate is 0.08)

channel was used to allocate the different measurement rates to each block at the encoder. Although it improved the image quality, it may lower the efficiency of the encoder.

Traditional video coding techniques, as we all know, can achieve high compression ratio by making complicated mode decision in the encoder. And can we perform the mode decision both in the encoder and decoder in the framework of CS? Therefore, we propose a new CS based video coding framework based on mode decision in this paper. We employ different measurement rates for different block modes at the encoder, and perform mode decisions at the decoder side.

The paper is organized as follows. Section 2 proposes CS video scheme based on adaptive coding. Experimental results are given in Sect. 3, and we conclude this paper in Sect. 4.

2 Adaptive Coding Based on Video CS

Motivated by the above analysis, we propose the adaptive coding for compressed video sensing scheme (ACCS). The architecture of our proposed ACCS framework is depicted in Fig. 2, and we will show the details in the next subsections.

2.1 Adaptive CS Video Encoder

At the encoder, firstly, video sequences are divided into group of pictures (GOP), and each GOP contains two categories: key frames (K-frames) and non-key frames (CS frames), as shown in Fig. 3 (K stands for key frames, and CS stands for non-key frames). Be different from [14], the K-frames and CS-frames in the paper are both coded using CS principles. K-frames are sampled at a high sampling rate, while CS-frames are sampled at adaptive sampling rate.

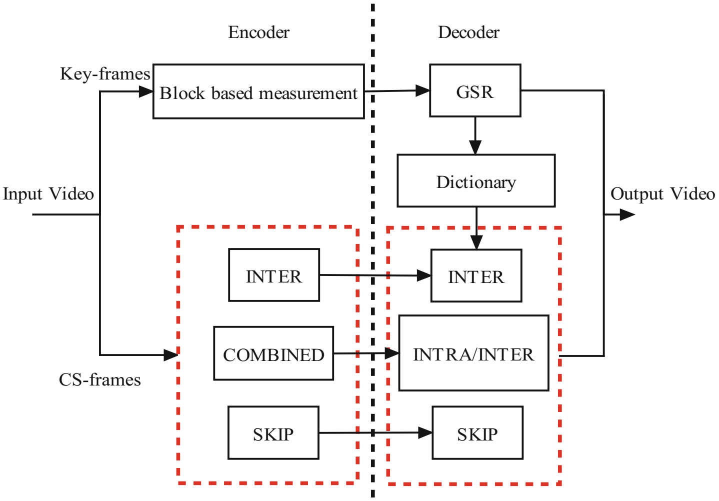


Fig. 2. Block diagram of the proposed CS scheme.

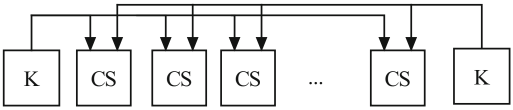
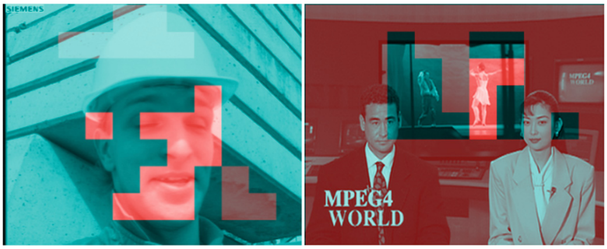


Fig. 3. GOP structure. Every CS-frame is recovered with two K-frames.

In this paper, we propose an adaptive sampling method for CS-frames by considering different characteristics of blocks in a frame. Firstly, we classify the blocks into three kinds of mode: SKIP, COMBINED and INTER based on the l_1 -norm of the residual between the current frame and the reference ones. If the residual is very small, which means the block changes very little with the co-located blocks in the reference frames, we only transmit a flag indicating the SKIP mode, which does not need any measurements so the sampling rate here is zero. If a block has very large change with respect to its co-located block, we will assign the COMBINED mode to it, which needs a high sampling rate and would be decoded both by INTRA and INTER. While the remainder blocks with minor change will take a small sampling rate, we refer to this coding mode as INTER mode, which decoded by INTER. The decoder schemes for three modes (SKIP, INTRA, INTER) will be explained in the next section. Figure 4 shows three kinds of mode of blocks in image, which is marked by different colors. The dark-red blocks show the COMBINED blocks, the light-red ones choose the INTER mode, and the rest blocks select the SKIP mode. The above description can be represented as follows

$$MODE = \begin{cases} SKIP, & difference < threshold1 \\ COMBINED, & threshold1 < difference < threshold2 \\ INTER, & difference > threshold1 \end{cases} \tag{3}$$



(a) (b)

Fig. 4. The mode decision: (a) foreman (b) news.

where *difference* means the l_0 -norm of the difference between the blocks in current frame and the references. $thresholds1$ and $threshold2$ are thresholds, respectively, which are set by experimental results.

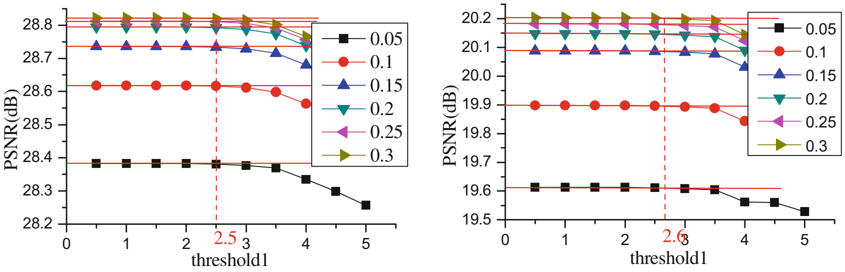
2.2 Adaptive CS Video Decoder

At decoder, we decode the K-frame by GSR, which can achieve the excellent performance in CS without using the information of the reference frames. For CS-frames, there are three kinds of blocks, and each of them will adopt different decoding schemes. The SKIP blocks can be decoded by copying the co-located block in the reference frames. The INTER blocks can be decoded by solving Eq. (2), whose redundant dictionary comes from the blocks near the co-located blocks in previously reconstructed frames. COMBINED blocks, which are the most complicated blocks, will perform the mode decision at the decoder. It recovered both by GSR (INTRA) and Eq. (2) (INTER). INTER blocks can be represented by the redundant dictionary sparsely, which is suitable for the blocks with complex local details but simple motion, while GSR cannot reconstruct the details finely at a low sampling rate, although it is a state-of-the-art scheme. As the blocks with complicated motion cannot be represented by the reference ones no matter how high the sampling rate is, they should be recovered by GSR. Then, we will decide which recovered block to be chosen by comparing the residual of the measurements based on the idea that the residual between the original ($\mathbf{x}$) and the prediction ($\hat{\mathbf{x}}$) is proportion to the residual of the measurements, as shown in (4). When the residual between the original measurements and the recovered measurements (by INTER) is smaller than the residual between the original and the recovered (by INTRA), we will choose block recovered by INTER and vice versa.

$$\mathbf{y} - \hat{\mathbf{y}} = \mathbf{\Phi}(\mathbf{x} - \hat{\mathbf{x}}) \tag{4}$$

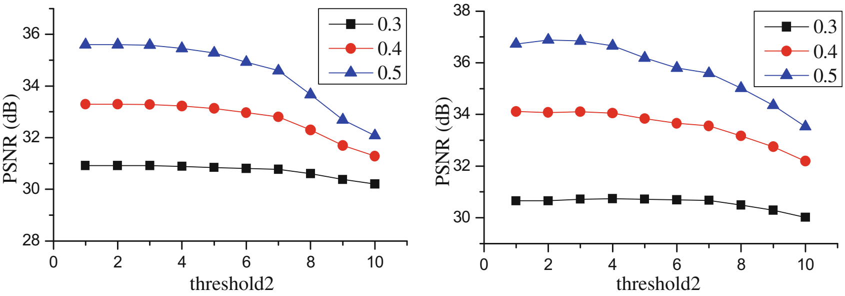
3 Experiment Results and Analysis

To evaluate the performance of our proposed scheme, several CIF video sequences are used: news, foreman, football, hall, coastguard, mobile, which represent for small, moderation and large movement videos. In our experiments, the CS measurements are obtained by using random matrix with Gaussian i.i.d entries at the block level, and the size of block (BS) is set to 32 * 32. *threshold*1 and *threshold*2 determine the coding modes for blocks. In order to determine the *threshold*1 and *threshold*2, we design a lot of experiments as follows. Firstly, we do not use the *threshold*2, and let the *threshold*1 vary from 0.05 * BS * BS to 5 * BS * BS. We can find from Fig. 5 that the PSNR is high when *threshold*1 < 3, and we choose *threshold*1 = 2 ∗ BS ∗ BS by considering all sequences. Secondly, in order to determine *threshold*2, we set the *threshold*1 to 2 * BS * BS and *threshold*2 varies from 1 * BS * BS to 10 * BS * BS. As shown in Fig. 6, we choose *threshold*2 = 8 ∗ BS ∗ BS by considering all sequences. In order to improve the compression ratio, the GOP size is set to 8, the first frame and the eighth frame are K-frames and others are CS-frames, as shown in Fig. 3. We assumed K-frames are losslessly available at the decoder in our experiments, as used in [14]. The high sampling rate (used in COMBINED mode) is five times the low one (used in INTER mode). We compare our proposed algorithm with two state-of-the-art image/video CS methods including GSR [7] and DISCOS [13]. GSR is an excellent still-image CS approach but smooths the details of the image. DISCOS makes full use of the inter-frame sparsity but it will introduce the blocking artifacts. Coding efficiency is measured using PSNR and sampling rate.



(a) (b)

Fig. 5. Recovery performance changes with *threshold*1 in different sequences: (a) coastguard (b) football.



(a) (b)

Fig. 6. Recovery performance changes with *threshold*2 in different sequences: (a) coastguard (b) football.

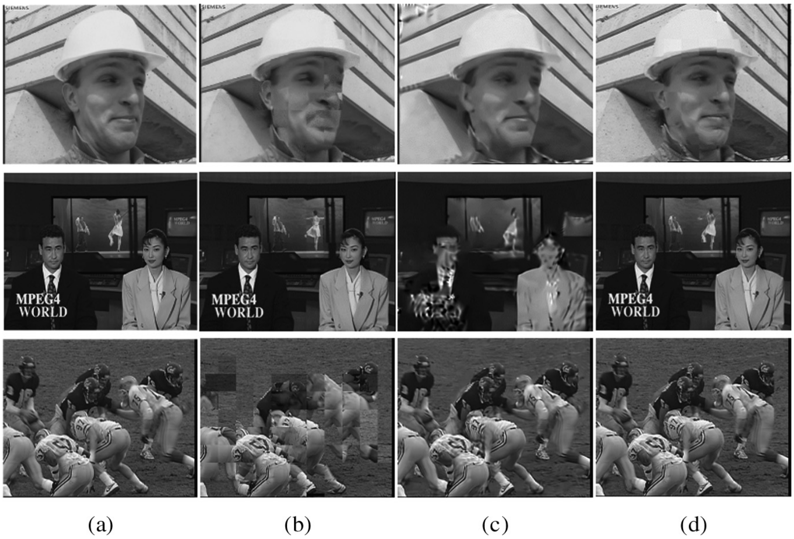
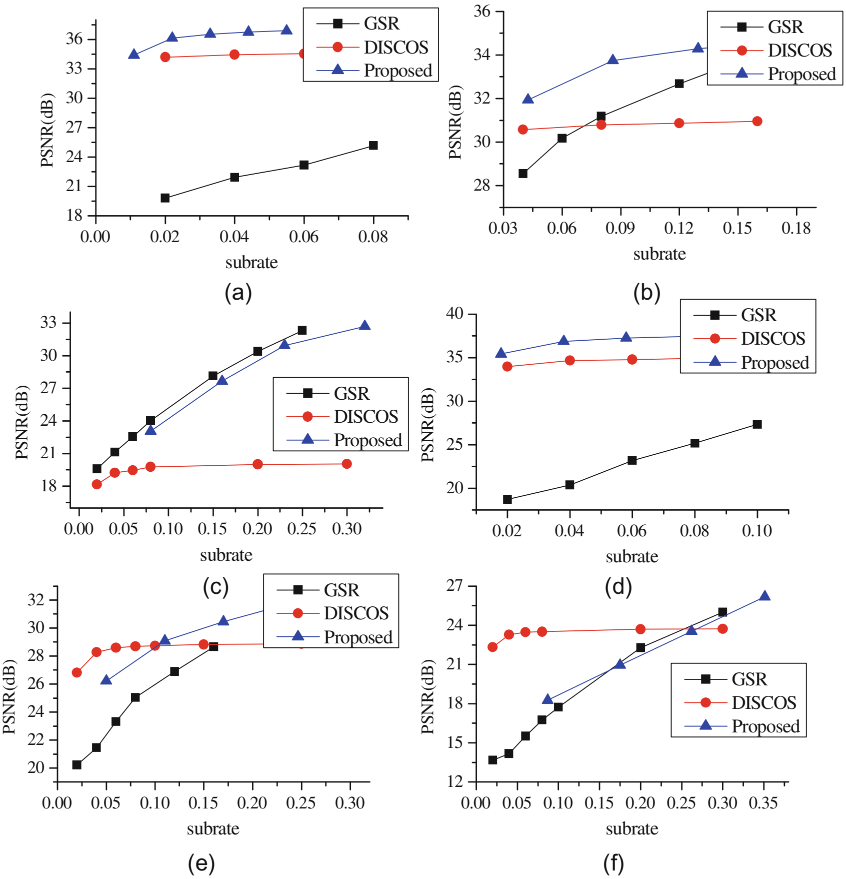


Fig. 7. Visual quality comparison with different algorithm for sequences (the 5th frame): (a) origin, (b) DISCOS [13], (c) GSR [7], (d) Proposed.

Figure 8 shows the rate-distortion curves of the proposed and the other approaches of all the test sequences. We can see that the rate-distortion performance of our approach is superior to GSR and DISCOS in most cases. Especially, for news and hall sequences with non-rigid objects, the proposed method achieves better coding performance compared to other two methods. For football sequence with large/complicated movement, our scheme will choose the INTRA mode in most cases, so the performance is similar to GSR. For Mobile sequence with large movement, different to football sequence, DISCOS achieves the best performance when the sampling rate is very low. This is due to the fact that mobile sequence has large motion but little shape-changing. Our scheme may choose the wrong encoding mode when the sampling rate is very low. However with the increase of sampling rate, our method achieves better performance than DISCOS. Besides, PSNR gain for DISCOS is less with the sampling rate increased [14]. Because the blocks have simple motion can be recovered perfectly even at a very low sampling rate, while the blocks have complicated motion which are not sparse in redundant dictionary cannot be reconstructed well.

Figure 7 shows the visual quality comparison for the foreman and news at the average rate about 0.04 (the actual sampling rates of our proposed scheme are 0.044 and 0.043 in foreman and news respectively) and for football at the sampling rate about 0.2. From Fig. 7, we can see that the proposed algorithm provides better visual quality than others. Table 1 provides the corresponding average execution time of various algorithms for reconstructing a frame in different sequences. These data are obtained using Matlab on a computer with Intel i5-3230, 2.6G CPU and 4 GB memory. From Table 1, we can see that the complexity of the proposed algorithm can be the medium one in most cases. Therefore, the proposed method provides a good tradeoff between visual quality and complexity.



(a) (b) (c) (d) (e) (f)

Fig. 8. Recovery performance comparison with different algorithms: (a) news, (b) foreman, (c) football, (d) hall, (e) coastguard, (f) mobile.

Table 1. Reconstruction time in minutes.

	Foreman (0.04)	News (0.04)	Football (0.2)
DISCOS	0.4	0.4	0.6
GSR	92.1	102.8	105.0
Proposed	13.3	3.0	42.5

4 Conclusion

In this paper, a new ACCS method has been proposed, which makes full use of the intra-frame and inter-frame sparsity. Our algorithm also exploits the fact that when the sequences have large or non-rigid motion, the sparsest representation of a block in video sequences is not a linear combination of a few temporal neighboring blocks that are in nearby key frames. Experimental results show that the proposed algorithm can achieve excellent performance even at a very low sampling rate, and outperforms existing state-of-the-art image/video CS approaches.

Acknowledgments. This work was supported by Natural Science Foundation of China under Grant Nos. 61671283 and 61301113, and Shanghai National Natural Science under Grant No. 13ZR14165.

References

1. Sze, V., Budagavi, M., Sullivan, G.J.: High Efficiency Video Coding (HEVC): Algorithms and Architectures. Integrated Circuit and Systems. Springer, Cham (2014)
2. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
3. Dai, W., Milenkovic, O.: Subspacepursuit for compressive sensing signal reconstruction. IEEE Trans. Inf. Theory **55**(5), 2230–2249 (2009)
4. Grant, M., Boyd, S.: CVX: MATLAB software for disciplined convex programming, version 2.0 beta, September 2013. http://cvxr.com/cvx
5. Gan, L.: Block compressed sensing of natural images. In: 15th International Conference on Digital Signal Processing, pp. 403–406, Cardiff (2007)
6. Zhang, J., Zhao, D., Jiang, F., Gao, W.: Structural group sparse representation for image compressive sensing recovery. In: IEEE Data Compression Conference, pp. 331–340, Snowbird (2013)
7. Zhang, J., Zhao, D., Gao, W.: Group-based sparse representation for image restoration. IEEE Trans. Image Process. **23**(8), 3336–3351 (2014)
8. Chang, K., Qin, T., Tang, Z.: Reconstruction of compressed-sensed video using compound regularization. In: 15th IEEE International Conference on Multimedia and Expo, pp. 1–6, Chengdu (2014)
9. Asif, M.S., Fernandes, F., Romberg, J.: Low-complexity video compression and compressive sensing. In: Asilomar Conference on Signals, Systems and Computers, pp. 579–583 (2013)
10. Mansour, H., Yilmaz, Ö.: Adaptive compressed sensing for video acquisition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3465–3468 (2012)
11. Che, W., Gao, X., Fan, X., et al.: Spatial-temporal recovery for hierarchical frame based video compressed sensing. In: IEEE International Conference on Image Processing, pp. 1110–1114 (2015)
12. Chang, K., Ding, P.L.K., Li, B.: Compressive sensing reconstruction of correlated images using joint regularization. IEEE Sig. Process. Lett. **23**(4), 449–453 (2016)
13. Do, T.T., Chen, Y., Nguyen, D.T., et al.: Distributed compressed video sensing. In: IEEE International Conference on Image Processing, pp. 1393–1396 (2009)
14. Prades-Nebot, J., Ma, Y., Huang, T.: Distributed video coding using compressive sampling. In: Picture Coding Symposium, pp. 1–4 (2009)
15. Ran, L., Zongliang, G., Ziguan, C., Minghu, W., Xiuchang, Z.: Distributed adaptive compressed video sensing using smoothed projected landweber reconstruction. China Commun. **10**(11), 58–69 (2013)