Distributed Compressed Video Sensing Based on Recurssive Least Square Dictionary Learning

Samad Roohi
Faculty Member
Computer Arts
Islamic Art University of Tabriz
Tabriz, Iran
Email: s.roohi@tabriziau.ac.ir

Jafar Zamani
Researcher
Biomedical Engineering
Amirkabir University of Technology
Tehran, Iran
Email: zamani.jafar@aut.ac.ir

Bagher B. Shotorban
Faculty Member
Computer Arts
Islamic Art University of Tabriz
Tabriz, Iran
Email: b.bahram@tabriziau.ac.ir

Abstract—**In this paper, we propose a method for distributed compressed video sensing (DCVS) based on dictionary learning. The proposed method divides the video sequences into group of pictures (GOP). Each GOP includes a key-frame following by a CS-frame. Compressed sensing (CS) is used to exploit spatial redundancy of frames. At encoder side Key-frames are sampled using random projection methods. To acquire much sparser version of CS-frames, basis extracted from CS-frame itself is used as a sparsifying basis. Sampling rate for key-frames and CS-frames are respectively adjusted to 0.5 and 0.25. At decoder side each frame reconstruction formulated as an l_1 − minimization problem. For each CS-frame, motion compensation interpolation method is applied on previous reconstructed key-frames to generate side information (SI). A dictionary is learned from SI and is used as a basis function in order to compensate low sample rate of CS-frames based of recursive least square dictionary learning algorithm (RLS-DLA). The results comparison with iterative least square dictionary learning algorithm (ILS-DLA) and K-SVD algorithm shows that the proposed method performs better than dictionaries learned by other methods.**

Keywords-compressed sensing; distributed video coding; dictionary learning; sparsifying basis; RLS-DLA

I. Introduction

Conventional video coding approaches are based on high-complexity techniques [1]. In these approaches, video signal is sensed and then compressed. Performing sensing and compressing disjointedly, will cause raw pixel data in the sensing stage that will be ignored in compressing stage. Acquiring data that will be discarded in latest stage, wastes most valuable allocated resources especially in resource limited applications. With the emersion of compressed sensing (CS) [2], and introducing the idea of single pixel camera [3], sensing and compressing steps have been merged into one step which samples and compresses a sparse video signal at a sub-Nyquist rate. It also promise that under certain conditions, the sensed video data will be reconstructed with fair quality.

Recently, the idea of CS has been spread to conventional distributed video coding. In distributed compressed video sensing (DCVS), the compressed video data for each frame taken directly via random projecting the raw data to linear, non-adaptive measurements. Reconstruction has done using solving the l_1 –minimization problem extracted from sensed measurements incorporate with utilizing inter-frame correlation [4]. In order to exploit the inter-frame correlation, video frames are classified to “key-frames” and “CS-frames”. Each “key-frame” has compressed and reconstructed independently from other frames, whereas each “CS-frame” will be sampled in low-rate compared with “key-frames” and reconstructed respect to successive previous reconstructed “key-frames”. In [5] a block-based selective video sampling approach has been proposed where, frames of video stream categorized to reference frames and non-reference frames. Each reference frame were sampled based on conventional video compressing techniques such as MPEG. Non-reference frames are divided non-overlapping blocks of same size. Sparsity of each block is predicted by latest reference frame. CS was applied to the blocks that identified as sparse, whereas the remaining blocks were sampled fully. In [6], we have introduced a low-complexity DCVS framework. where, to obtain a fair quality, each “CS-frame” has been constructed using compressed data respect to generated side information (SI) from latest pair of reconstructed “key-frames”. In [7] a dictionary learning-based DCVS has been proposed. where, video stream are divided to key-frames and CS-frames. Key-frames are sampled and reconstructed using common CS frame-based techniques. CS-frames are sampled using block-based random projection. To reconstruct each CS-frame, a dictionary was built using two of previous reconstructed key-frames in combination with SI generated from them. K-SVD was applied to learn the dictionary.

In this paper we propose a novel method for DCVS. The proposed compression scheme divides the video sequence into “key-frames” and “CS-frames”. Key-frames are sampled and recovered using common CS techniques. CS-frames were compressively sampled with a sampling rate much less than key-frames. To reconstruct each CS-frame, a dictionary is built using SI. SI can be generated by motion compensated interpolation from previous reconstructed key-frames. Dictionaries are learned in 9/7 wavelet domain using recursive least square dictionary learning algorithm (RLS-DLA) [8]. The simulation results illustrates that using RLS-DLA although significantly reduces the number of samples, it preserves video quality in fair level. The main contributions of this work are that dictionary learning is done by RLS-DLA on SI generated from previous reconstructed key-frames to obtain an efficient sparsifying basis for constructing CS-frames.

II. Related Works

A. Compressed Sensing Theory

In the field of signal processing, the Nyquist-Shannon sampling theorem [9] states that to preserve information, sampling rate must be at least twice the highest signal frequency component. Using this method in most state of the art signal processing applications such as video processing applications, generates large amounts of data to transmission or storing.

CS is a novel framework that samples and compress data simultaneously in the sub-Nyquist sampling rate with small sacrifice in reconstructed signal [2], [10].In this part, we briefly review CS concept.

Suppose that each frame of a video sequence demonstrated as $\alpha = \alpha_1, \alpha_2, \dots, \alpha_n$. α, is k-sparse, if representation of it in an orthonormal basis has at most k nonzero entries:

$$\alpha_{n\times 1} = \Psi_{n\times n} \cdot s_{n\times 1}. \quad (1)$$

where Ψ is an orthonormal basis matrix (or dictionary) can provide a k-sparse representation for α and s is representation of α in Ψ where s can be well approximated using only $k \ll n$ non-zero entries. CS states that α could be reconstructed accurately using a relatively small number of non-adaptive linear projection measurements:

$$y_{m\times 1} = \Phi_{m\times n} . \Psi_{n\times n} \cdot s_{n\times 1}. \quad (2)$$

where y is measurement vector which can be considered as compressed version of α and Φ is a measurement matrix incoherent with Ψ. Since $m \ll n$, there are infinite solution for (2) , accordingly traditional methods such as least square are not able to solve it. CS states that if s is k-sparse in some known transform domain, i.e. $||s||_0 \leq k$, where $||.||_0$ is l_0 norm, then to reconstruct α, the following underdetermined problem should be solved:

$$y_{m\times 1} = \Phi_{m\times n} . \Psi_{n\times n} \cdot s_{n\times 1} = \Theta_{m\times n} . s_{n\times 1}. \quad (3)$$

where Θ is a $m \times n$ measurement matrix. If $m \geq 2k$ and Θ meets restricted isometry property (RIP) conditions [11]. Then (3) could be uniquely solved through finding the sparsest solution for:

$$\min ||s||_0 \quad s.t. \quad y = \Theta . s. \quad (4)$$

Numerical solution to solve (4) are unstable and NP-complete. In [11] it has been proved that if Θ meets restricted isometry property (RIP) conditions with parameter $\left(2k, \sqrt{2} - 1\right)$, then l_1 norm can efficiently approximates k-sparse signal using only $m \geq \left(ck \log\left(\frac{n}{k}\right)\right)$ where, $k < m \ll n$ measurements with computational complexity of $O(n^3)$.

B. Distributed Compressed Video Sensing

In conventional video coding systems, such as H.264/MPEG-x, the temporal correlation between successive frames is obtained using complex algorithms such as motion compensation in the encoder side. In Distributed video coding (DVC), correlation between two frames can be obtained by encoding separately and decoding jointly.

In distributed compressed sensing (DCS) [12], each frame of a video sequence is measured independently using CS and reconstructed jointly at decoder. In [6], we proposed a framework to simultaneously sensing and compressing video frames. We used various sampling matrices in encoder side and nonlinear reconstruction algorithm to reconstruct video frames in decoder side e.g. gradient projection for sparse representation (GPSR) and non-linear conjugate gradient algorithm (NLCG) [13].

C. Dictionary Learning

Dictionary learning is a topic in signal processing area to find a representation of original signal that approximates it with as few atoms as possible [8]. In sparse representation a vector $\mathcal{X}$ is represented or approximated as a linear combination of some few of the dictionary atoms. The approximation of $\mathcal{X}$ can be written as:

$$\mathcal{X}_{approx} = D\mathcal{W}. \quad (5)$$

where $\mathcal{W}$ is a vector of coefficients includes most of the entries equal to zero.

The key concept of dictionary learning is the choice of sparsifying basis (dictionary). There are many pre-specified sparsifying basis e.g. discrete wavelet transform (DWT), discrete cosine transform (DCT), curvelets and etc. These bases are computationally fast and simple for implementation although, they are not able to represent a signal efficiently. If the basis be a dictionary extracted from the image itself, it provide much sparser representation for the image [14]. Incoherency between dictionary D and Φ can be achieved by generating a random matrix via some known random distributions.

III. The Proposed Method

In contrast with traditional video coding approaches, DVCS implements acquiring video data and compression in a unified task using random projecting of each frame at a low-complexity encoder. As a result, CS transfers the complexity to decoder side which is more acceptable in most of the modern video applications, e.g. visual sensor networks. In the following parts we investigate the architecture of our proposed method then we discuss two sides of proposed method separately.

A. The Architecture of Proposed Method

The main structure of proposed method is based on our previous work in [6] . The Architecture of proposed method in this paper is shown in Fig. 1 and each stage is explained subsequently:

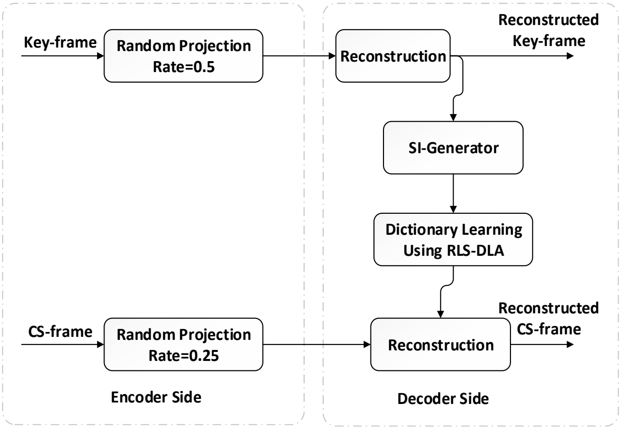


Figure 1. The architecture of proposed method

B. Description of Encoder Side

The proposed method, divides a sequence of video into several group of pictures (GOP) at encoder side. Each GOP includes a key-frame followed by some CS-frames. Each key frame compressively sampled via CS random projection. We use scrambled block Hadamarad ensemble (SBHE) matrix as measurement matrix, which uses partial block Hadamard transform followed by randomly permuting its columns. DWT used as sparsifying transform domain. As our previous work [6] the sampling rate for key-frames are set to 0.5.

To sample the CS-frames, we use a dictionary learning approach formulation proposed in [8]. In this approach, each CS-frame f_i represented as a vector X, directly generated from non-overlapping patches of the image. Let X be a matrix of columns vectors x_i and C a matrix of coefficients with c_i as columns. Respect to X and C, the dictionary learning problem can be formulated as a hard optimization problem:

$$\{D_{opt}, C_{opt}\} = \arg \min_{D,C} ||C||_0 + \Upsilon || - DC||_2^F. \qquad (6)$$

where C is sparse matrix obtained from sparse approximation of X using dictionary D. Solution for (6) could be found using order recursive matching pursuit (ORMP). We apply frame based random projection as $y = \Phi . C$ to obtain measurement vector. Using dictionary learning approach, the sampling rate for CS-frames are set to about 0.25. Then y is transmitted to decoder side as compressed version of f_i.

C. Description of Decoder side for Key-Frames

At the decoder side, each key-frame is reconstructed using GPSR, which solves the following convex unconstrained optimization problem:

$$\min_{s_t} ||y_t + \Theta s_t||_2^2 + \tau||s_t||_1. \qquad (7)$$

where $y_t \epsilon R^m$ is received measurement vector, $\Theta = \Phi . \Psi$, Φ is the measurement matrix, Ψ is the DWT basis, τ is a non-negative parameter and s_t is the sparse term coefficient vector. (7) Could be solved via SpaRSA algorithm [15].

TABLE I. COMPARISIONS OF RESULTS

Video Sequences	Method	SNR	PNSR	SSIM
Average of results for two successive CS-frame of Coastguard	***ILS-DLA***	15.85	33.66	0.91
	K-SVD	16.72	34.11	0.93
	RLS-DLA	17.92	36.94	0.96
Average of results for two successive CS-frame of Foreman	***ILS-DLA***	17.02	35.54	0.93
	K-SVD	17.52	37.22	0.94
	RLS-DLA	18.73	40.62	0.97

D. Description of Decoder side for Key-Frames

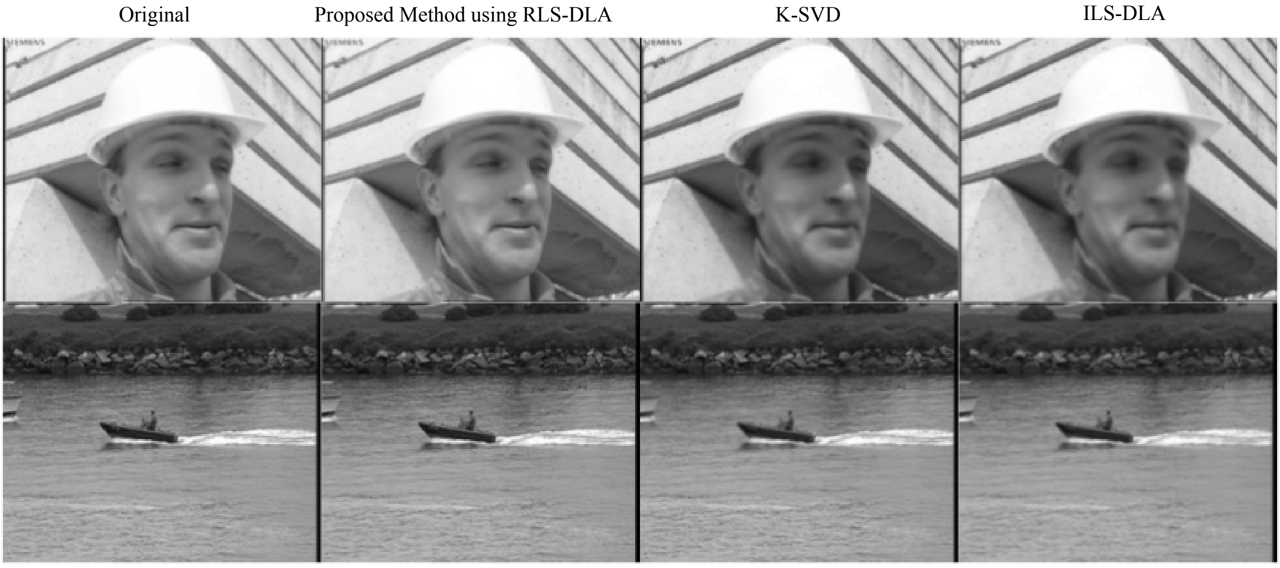
As mentioned before, extracting sparsifying basis from image itself, results in much sparser representation for the image despite the fact that it is impossible to acquire such a basis in decoder side from the measurements. Dictionaries learned from neighboring images is the best way to obtain such basis.

To reconstruct each CS-frame f_t, its SI is generated via SI-generation methods such as motion-compensated interpolation of previous key-frame f_{t-1} and next key-frame as f_{t+1}. A dictionary is learned using recursive least square dictionary learning algorithm (RLS-DLA) on SI as follow. First we extract non-overlapping patches from generated SI. Each patch is made into a training vector simply by lexicographically ordering of the pixels. The training vectors that are selected in a random way from the set of training images, are presented for RLS-DLA algorithm. The initial dictionary D_0 is made using the K first random training vectors. The current dictionary continuously will be updated at each step. In RLS-DLA they are defined a time step ‘ i ’, the matrix X_i of size $N \times i$, $W_i = [w_1, w_2, \ldots, w_i]$ of size $K \times i$ and $C_i = (W_i W^T)^{-1}$ as well as D_i which is the least square minimization of $|| X_i - DW_i||_F^2$. In each step a new training vector x_i is provided and the corresponding weights w_i are found using the latest dictionary D_{i-1} and a vector selection algorithm [8]. Updating rule is as follow:

$$C_i = C_{i-1} - \lambda u u^T. \qquad (8)$$

$$D_i = D_{i-1} - \lambda r_i u^T. \qquad (9)$$

where $u = C_{i-1} w_i$ and $\lambda = 1/(1 + w_i^T u)$, $r_i = x_i - D_{i-1} w_i$ is the representation error. Introducing an adaptive forgotten factor μ_i in step i, results in the less dependency of dictionary on the initial dictionary and improving convergence properties of it. The training process will be done in two stages. First we obtain w_i by a Matching Pursuit algorithm using a stop criteria on approximation error. The second stage updates the dictionary and C_i. As depicted in Fig. 1, after obtaining dictionary, we use it as basis matrix to reconstruct the CS-frames.



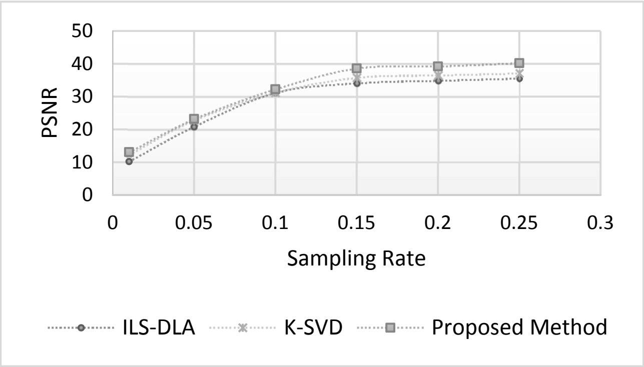
Original Proposed Method using RLS-DLA K-SVD ILS-DLA

Figure 2. Comparison of proposed method with two known dictionary learning method for two CS-frame of foreman and coastguard video sequences. As you see the proposed method outperforms other methods in quality

IV. Results

In this section, we evaluate the performance of proposed method on the two well-known video sequences ("Foreman" and "Coastguard") with a CIF resolution of 352×288 pixels and GOP=2. Experiments were performed using Matlab 2014b, on a computer with Intel® Core^{TM} $i7$, 4.0 GHz processor with 16 GB of RAM. To evaluate the proposed method, three applicable quality assessors, the signal-to-noise ratio (SNR), the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) were employed. The key-frames are sampled via the SBHE [SBHE citation] and constructed using the method proposed in [16], with measurement rate equal to 0.5. To reconstruct the Cs-frames which is sampled in a rate much less than key-frames (sampling rate ≈ 0.25), we use a dictionary learning approach proposed based on [8]. We use the dictionary generated from related SI as a basis to restore the sparse representation of image. The dictionary size is set to 64×440, corresponding to 8-by-8 patches of image in three level dyadic $9/7$ wavelet in DWT domain. We randomly pick 1500 patches from each of training images in transform domain. To apply RLS-DLA, a new training vector selected randomly in each iteration. ORMP is used for vector selection. Learning error limit, adjusted using target PSNR equal to 38 dB.

TABLE II. Comparision Results For the Foreman CS-Frames



The results of proposed method compared with two dictionary-learning based image and video compression in TABLE. 1 and Fig.2. The first compared method is DCVS based on dictionary learned by K-SVD and the second one is DCVS based on dictionary learned by ILS-DLA.

Results in TABLE. I, TABLE. II and Fig.2 shows that using a dictionary based approach based on RLS-DLA outperforms K-SVD and ILS-DLA in quality measures.

V. References

[1] T. Sikora, "The MPEG-4 video standard verification model," *Circuits Syst. Video Technol. IEEE Trans. On*, vol. 7, no. 1, pp. 19–31, 1997.

[2] D. L. Donoho, "Compressed sensing," *Inf. Theory IEEE Trans. On*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *Signal Process. Mag. IEEE*, vol. 25, no. 2, pp. 83–91, 2008.

[4] L.-W. Kang and C.-S. Lu, "Distributed compressive video sensing," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 1169–1172.

[5] V. Stankovic, L. Stankovic, and S. Cheng, "Compressive video sampling," in *Signal Processing Conference, 2008 16th European*, 2008, pp. 1–5.

[6] S. Roohi, M. Noorhosseini, J. Zamani, and H. S. Rad, "Low complexity distributed video coding using compressed sensing," in *Machine Vision and Image Processing (MVIP), 2013 8th Iranian Conference on*, 2013, pp. 53–57.

[7] H.-W. Chen, L.-W. Kang, and C.-S. Lu, "Dictionary learning-based distributed compressive video sensing," in *Picture Coding Symposium (PCS), 2010*, 2010, pp. 210–213.

[8] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *Signal Process. IEEE Trans. On*, vol. 58, no. 4, pp. 2121–2130, 2010.

[9] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[10] E. J. Candes and M. B. Wakin, “An Introduction To Compressive Sampling,” *Signal Process. Mag. IEEE*, vol. 25, no. 2, pp. 21 –30, Mar. 2008.

[11] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Math.*, vol. 346, no. 9, pp. 589–592, 2008.

[12] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, *Distributed compressed sensing*. 2005.

[13] S. Roohi, J. Zamani, M. Noorhosseini, and M. Rahmati, “Super-resolution MRI images using Compressive Sensing,” in *2012 20th Iranian Conference on Electrical Engineering (ICEE)*, 2012, pp. 1618 –1622.

[14] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, “Learning overcomplete dictionaries based on atom-by-atom updating,” *Signal Process. IEEE Trans. On*, vol. 62, no. 4, pp. 883–891, 2014.

[15] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, “Sparse reconstruction by separable approximation,” *Signal Process. IEEE Trans. On*, vol. 57, no. 7, pp. 2479–2493, 2009.

[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,” *Sel. Top. Signal Process. IEEE J. Of*, vol. 1, no. 4, pp. 586–597, 2007.