PAPER

# Distributed Compressed Video Sensing with Joint Optimization of Dictionary Learning and $l_1$-Analysis Based Reconstruction

Fang TIAN[†], Jie GUO[†], *Nonmembers*, Bin SONG[†a)], *Member*, Haixiao LIU[†], *and* Hao QIN[†], *Nonmembers*

**SUMMARY** Distributed compressed video sensing (DCVS), combining advantages of compressed sensing and distributed video coding, is developed as a novel and powerful system to get an encoder with low complexity. Nevertheless, it is still unclear how to explore the method to achieve an effective video recovery through utilizing realistic signal characteristics as much as possible. Based on this, we present a novel spatiotemporal dictionary learning (DL) based reconstruction method for DCVS, where both the DL model and the $l_1$-analysis based recovery with correlation constraints are included in the minimization problem to achieve the joint optimization of sparse representation and signal reconstruction. Besides, an alternating direction method with multipliers (ADMM) based numerical algorithm is outlined for solving the underlying optimization problem. Simulation results demonstrate that the proposed method outperforms other methods, with 0.03-4.14 dB increases in PSNR and a 0.13-15.31 dB gain for non-key frames.

***key words:*** *compressed video sensing, realistic signal characteristics, dictionary learning, alternating direction method with multipliers*

## 1. Introduction

Conventional video coding schemes, such as H.264/AVC and HEVC, often require much more computation for the encoder than the decoder. This asymmetry is well suited for broadcasting or for steaming video-on-demand systems where video is compressed once and decoded many times. However, in many application scenarios, e.g., video surveillance, and camera phones, a video codec with simple encoder and complex decoder is desired. Distributed video coding (DVC) [1] is such a solution for scenarios described above. For a video sequence, frames are independently encoded and then decoded collectively [2], [3], and consequently, we can shift the significant computational burden from the encoder to the decoder. More recently, compressed sensing (CS) theory [4], [5] has become widely popular since it offers a novel approach to gather data, which combines signal sampling and compression into one step and thus lowers the needed quantity of measurements. In this sense, the CS framework is suitable for DVC because of the low cost of power consumption and computation.

Recently, distributed compressed video sensing (DCVS) [6]–[13] has arisen as a novel method to acquire video by using random measurements at the encoder and undertaking joint recovery in the decoding process. The key issue of DCVS is about the method to employ the spatial/temporal redundancy existing in videos to obtain efficient reconstruction in the decoding end. In 2009, Prades-Nebot et al. [6] firstly presented a typical DCVS architecture, where a video sequence is split into two groups, namely the key frames and non-key (NK) ones. Key frames are independently coded; while NK frames are encoded using a CS based encoder. In [7], a novel two-phase measurement acquisition scheme with the inter-frame sparsity model was proposed for DCVS. However, in these methods, a large amount of raw video data is still required for key frames encoding employing conventional compression algorithms, which then wastes valuable resources.

Another DCVS framework was previously presented [8], [9], where sparse representation was achieved through the K-SVD dictionary learning algorithm [12]. Specially, in combination with the side information, a dictionary is learned using samples extracted from the former reconstructed frames. After obtaining the trained dictionary, we use the traditional CS reconstruction algorithm to recovery the NK frames. However, in the above papers, signal reconstruction and sparse representation are designed as independent tasks [8], [9], and then there is much space to improve in the field of resources consuming, for the reason that sparse representation has been embedded in the dictionary learning process. Meanwhile, there are some other papers about video coding based on CS [14]–[21]. For example, a motion-aware decoding method for compressively sampled videos and a maximum frame rate video acquisition framework were proposed in [14] and [15] respectively. In [16], the challenges involved in the transmission of CS-based video over wireless multimedia sensor networks were discussed by Pudlewski et al., and a cross-layer system that jointly controls the video encoding rate, the transmission rate. Additionally, a method to generate dictionary for video sampling based on CS was proposed in our previous work [17], an adaptive alternating direction method of multipliers (ADMM) with its application to compressed video sensing was presented in [18], [19], and more recently, a joint sampling rate and bit-depth optimization framework was proposed in [20]. Nevertheless, how to use the redundant information of temporal and spatial relations by means of side information (SI) at the decoding end to get efficient sparse dictionary, and signal recovery from limited numbers of measurements, is substantially unexplored.

Based on this situation, we present a spatiotemporal dictionary learning (DL) based restore method for DCVS.

Manuscript received September 12, 2015.
Manuscript revised December 21, 2015.
Manuscript publicized January 21, 2016.

[†] The authors are with the State Key Laboratory of Integrated Services Networks, Xidian University, China.
a) E-mail: bsong@mail.xidian.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2015EDP7373

The target is to enhance performance through the joint optimization of sparse representation and signal reconstruction at the decoder, and reserving the relatively low complexity in computation at the encoding end at the same time. On this point, we develop a novel DL based reconstruction method, wherein both the DL model and the signal recovery with correlation constraint are included in the optimization problem. Specially, in our method, the dictionary is learned from overlapping blocks of the video sequence along both the spatial and temporal directions, and thus providing a sparser representation over the fixed analytical transforms. The correlation noise between the current frame and its SI is also considered in the video signal recovery from undersampled measurements, in order to avoid over-sparse solutions which could lead to lower visual quality. Furthermore, our proposal is an $l_1$-analysis based reconstruction method which finds the estimated signal in the pixel domain directly, instead of firstly finding the sparse coefficients with respect to the sparsifying basis and then recovering the frame via a synthesis operator as in the conventional methods [8], [9].

Another contribution of this paper is an iterative algorithm to solve the underlying optimization problem. The idea behind it is to examine an efficient and adaptive recovery method that could exploit more signal characteristics that go beyond simple sparsity, and while being rich enough to capture the complexity of modern big data and scalable enough to process huge datasets in a parallelized fashion. In this way, we present an ADMM [22], [23] based numerical algorithm, wherein DL is included in the iteration process as one single step, and consequently, achieving the joint optimization of sparse representation and signal reconstruction.

Finally, it is noteworthy that we mainly concentrate on exploring an efficient DCVS scheme with joint optimization of sparse representation and signal reconstruction, which offers a new simple video acquisition (and compression) framework and an alternative algorithm fit for the scenario in which original video signal cannot be obtained, not to compete the compression performance against the current video compression standards, which require original data to perform encoding. The remainder of this paper is organized as follows. Background information is provided in Sect. 2. The presented joint DL and signal recovery optimization problem and the ADMM-based numerical algorithm are described in Sect. 3. Section 4 presents the DCVS decoder architecture with DL based reconstruction. Experimental results are presented in Sect. 5. Conclusions are given in Sect. 6.

## 2. Background

### 2.1 Compressed Sensing

The concept of CS is firstly presented by Candes, Tao [4] and Donoho [5]. The new theory enables to efficiently and directly acquire signals with (random) linear projections and reconstruct them by settling a convex optimization problem. More precisely, given a discrete signal $f$ with length $N$, and its coefficients $\Psi$ with respect to the orthonormal basis $\Psi \in R^{N\times N}$, we say signal $f$ is $K$ -sparse with respect to $\Psi$ if only $K$ coefficients are non-zero. In this way, we could effortlessly compress the signal through encoding the values and positions of these non-zero coefficients. However, this process is extremely inefficient, since only $K$ coefficients are delivered by the encoder system while $N$ signals samples have to be acquired. The CS theory claims that the signal $f$ could be collected using the following linear random projections:

$$y = \Phi f \tag{1}$$

in which $y \in R^M$ is the sampled measurement ($M < N$), $\Phi \in R^{M\times N}$ is the measurement matrix, and the ratio between the height and width of the measurement matrix is defined as the measurement rate ($MR$), i.e,

$$MR = \frac{M}{N} \tag{2}$$

Thus the recovery of the sparse coefficients (with respect to $\Psi$) can be done by finding the set of coefficients that agrees with the measurements, and especially, with the minimum $l_0$ norm i.e.:

$$\min \|x\|_0 \quad \text{s.t.} \quad y = \Phi\Psi x \tag{3}$$

where $x$ is the sparse coefficients of $f$ with respect to $\Psi$.

Intractable as the problem is NP-hard for typical values of $N$, it is still solvable if the product of $\Phi$ and $\Psi$, denoting $A = \Phi\Psi$, obeys the Restricted Isometry Property (RIP) of order $K$ [5], that is $(1-\delta_K)\|s\|_2^2 \le \|As\|_2^2 \le (1+\delta_K)\|s\|_2^2$, which holds for all $K$-sparse vectors $s$ for a small isometry constant $0 < \delta_K < 1$. And then the signal can be recovered by solving the following unconstrained optimization problem Eq. (4).

$$\min \|x\|_1 \quad \text{s.t.} \quad y = \Phi\Psi x \tag{4}$$

This convex optimization problem, namely basis pursuit, can be recast as a linear program to be efficiently solved with the available optimization algorithms. However, the complexity of Eq. (4) still makes it impractical for many practical applications. In this way, many iterative greedy techniques have also been proposed to solve the above problem in the literature, e.g., matching pursuit (MP). It has been proven that MP could successfully reconstruct the compressively sampled signal with high probability. The other greedy algorithms such as orthogonal matching pursuit (OMP), stage-wise orthogonal matching pursuit (StOMP), and Subspace Pursuit have also been shown to attain similar guarantees to those of their optimization-based counterparts. Besides, it should be noted that many signals of interest in practice are often "approximately" sparse rather than "exactly" sparse, i.e., the transforming coefficients are generally different to zero and only few of them have significant values. In this case, the solution to Eq. (4) could still reconstruct the most sparse coefficients as revealed in the CS theory [4], [5], and hence, provide a good approximation of the original signal.

### 2.2 Distributed Video Coding

DVC is based on two theorems that are Slepian-Wolf and Wyner-Ziv of the information theory. In a classical DVC framework, the video sequence is divided into several groups of pictures (GOP), and each GOP has been made up by one key frame and multiple Wyner-Ziv (WZ) frames. Besides, key frames are encoded and decoded using conventional video compression standards such as H.264/AVC. Meanwhile, the compressed version of a WZ frame $f_{WZ}$ is derived by transferring the parity bits which comes from the channel-encoded version of $f_{WZ}$ without performing motion estimation. At the decoding end, the side information $f_{SI}$ is created by motion compensation in advance, and then the decoding end utilizes the received parity bits and $f_{SI}$ to recover $f_{WZ}$. By exploiting the temporal and spatial redundancy of videos at the decoding end, the major computational complexity of DVC in the encoding end is transferred to the decoder.

This paper concentrates on DCVS that integrates strengths of both CS and DVC. We put forward to jointly optimize dictionary learning and signal recovery, as well as capture the inter-frame correlation existed in video. At this point, we propose a spatiotemporal DL based reconstruction scheme for DCVS, and develop an ADMM based numerical algorithm to figure out the underlying optimization problem.

## 3. Joint Optimization of Dictionary Learning and Reconstruction

### 3.1 Joint Optimization Problem

The typical DCVS system is used in this paper (to be shown in Fig. 1). The key frame $f_K$ is projected and recovered employing the orthonormal basis $\Psi$ and the conventional CS reconstruction algorithm Eq. (4). The NK frame $f_{NK}$ is segmented into some non-overlapping blocks. The $i$th block is vectorized as $f_{NK,i} \in R^N$ ($i = 1, 2, \dots, B$ and $B$ is the number of blocks) and projected using the random measurement matrix $\Phi \in R^{M\times N}$, i.e., $y_{NK,i} = \Phi f_{NK,i}$. Then, the global NK frame acquisition process can be formulated as follows:

$$\begin{bmatrix} y_{NK,1} \\ y_{NK,2} \\ \vdots \\ y_{NK,B} \end{bmatrix} = \begin{bmatrix} \Phi & & & \\ & \Phi & & \\ & & \ddots & \\ & & & \Phi \end{bmatrix} \begin{bmatrix} f_{NK,1} \\ f_{NK,2} \\ \vdots \\ f_{NK,B} \end{bmatrix} \tag{5}$$

$$or \quad y_{NK} = \Phi_{NK} f_{NK}$$

where $y_{NK} = [y_{NK,1}^T, y_{NK,2}^T, \dots, y_{NK,B}^T]^T$, $f_{NK} = [f_{NK,1}^T, f_{NK,2}^T, \dots, f_{NK,B}^T]^T$ and $\Phi_{NK} = diag\{\Phi, \Phi, \dots, \Phi\}$.

Inspired by the work of Wang et al. for medical resonance imaging [24], we first investigate the dictionary learning model for DCVS reconstruction. Let $R_i$ be the operator that extracts a block from the video frame, i.e., $f_{NK,i} = R_i f_{NK}$. Then the DL problem is expressed as follows

$$\min_{D,\{x_i\}} \frac{1}{2} \sum_i \|R_i f_{NK} - Dx_i\|_2^2 + \mu \sum_i \|x_i\|_0 \tag{6}$$

in which $D$ is the "global" dictionary for all the blocks in $f_{NK}$.

Then, based on the aforementioned DL model, we formulate the joint optimization problem for DCVS reconstruction as

$$\min_{f_{NK},D,\{x_i\}} \frac{1}{2}\|\Phi_{NK} f_{NK} - y_{NK}\|_2^2 + \frac{\lambda_1}{2} \sum_i \|R_i f_{NK} - Dx_i\|_2^2$$
$$+\mu \sum_i \|x_i\|_0 + \lambda_2 h(f_{NK}) \tag{7}$$

where $\lambda_1$ and $\lambda_2$ are tuning parameters, and $h(f)$ is an additional regularizer. The first term in Eq. (7) enforces the data fidelity in the measurement domain. The second and third terms correspond to the DL model. The selection of $h(f)$ can be any suitable sparsifying transforms, such as discrete wavelet transform (DWT) or discrete cosine transform (DCT). Nevertheless, in the DVC framework, it is known that the correlation between $f_{NK}$ and its SI $f_{SI}$ can be modeled as the correlation noise model (CNM) that follows the Laplacian distribution [1]. The similar sparse distribution can also be found in the DCVS framework [10]. Consequently, in this paper, we choose the frequency correlation noise model to characterize $h(f)$:

$$h(f) = \|\Psi^T(f - f_{SI})\|_1 \tag{8}$$

In other words, the sparsity requirement of CS is achieved under the video correlation constraint in Eq. (8), and thus more signal structures are leveraged in the DCVS framework to improve the reconstruction visual quality. Furthermore, it should be noted that the proposed optimization problem Eq. (7) is actually an analysis-based method, which finds the estimated $\hat{f}_{NK}$ directly from Eq. (7). It is quite different from the conventional synthesis-based schemes Eq. (4) in literatures, wherein one first finds the sparsest possible coefficient $\hat{x}$ and then the solution to $\hat{f}_{NK}$ is derived via a synthesis operation (i.e., $\hat{f}_{NK} = D\hat{x}$).

### 3.2 ADMM-Based Numerical Algorithm

It is challenging to simultaneously find all unknown variables in the objective function Eq. (7) which is not differentiable due to the $l_0$ and $l_1$ term. In this section, we focus on the variable splitting and augmented Lagrangian methods, and develop an ADMM based iterative algorithm for solving the optimization problem Eq. (7). Although ADMM was originally proposed in the mid-1970s by Glowinski et al. [22] and Gabay et al. [23], it has been widely known until recently with the development of large-scale distributed computing systems and massive optimization problems, especially for problems arising in sparse recovery. For example, a split augmented Lagrangian shrinkage algorithm was proposed in [25]. It aims to transform the unconstrained problem into a constrained one via variable splitting trick, and then attack this constrained problem using ADMM. Besides, a fast algorithm was presented for total variation-based image deblurring in [26] and an adaptive-ADMM algorithm with support and signal value detection was presented in [18], [19].

Here, we first introduce an auxiliary variable $d = \Psi^T(f - f_{SI})$ to decouple the $l_1$ term from other parts and obtain the following equivalent problem (here we use $f$ to denote the NK frame for the sake of brevity):

$$\begin{aligned}\min_{f,D,\{x_i\}} \frac{1}{2}\|\Phi_{NK}f - y\|_2^2 + \frac{\lambda_1}{2}\sum_i \|R_i f - Dx_i\|_2^2 \\ +\mu\sum_i \|x_i\|_0 + \lambda_2\|d\|_1 \qquad (9)\\ \text{subject to } d - \Psi^T(f - f_{SI}) = 0\end{aligned}$$

The scaled augmented Lagrangian function of (9) is:

$$\begin{aligned}L_p(f, D, x_i, d, u) = \frac{1}{2}\|\Phi_{NK}f - y\|_2^2 + \frac{\lambda_1}{2}\sum_i \|R_i f - Dx_i\|_2^2 \\ +\mu\sum_i \|x_i\|_0 + \lambda_2\|d\|_1 + \frac{\rho}{2}\|d - \Psi^T(f - f_{SI}) + u\|_2^2 \qquad (10)\end{aligned}$$

where $u$ is called the scaled dual variable of the Lagrangian multiplier and $\rho$ is the penalty parameter. Based on ADMM, we derive the following iteration scheme:

$$\{D^{k+1}, x_i^{k+1}\} = \arg\min_{D,x_i} \frac{\lambda_1}{2}\sum_i \|R_i f^k - Dx_i\|_2^2 + \mu\sum_i \|x_i\|_0 \quad (11)$$

$$\begin{aligned}f^{k+1} = \arg\min_f \frac{1}{2}\|\Phi_{NK}f - y\|_2^2 + \frac{\lambda_1}{2}\sum_i \|R_i f - D^{k+1}x_i^{k+1}\|_2^2 \\ + \frac{\rho}{2}\|d^k - \Psi^T(f - f_{SI}) + u^k\|_2^2 \qquad (12)\end{aligned}$$

$$d^{k+1} = \arg\min_d \lambda_2\|d\|_1 + \frac{\rho}{2}\|d - \Psi^T(f^{k+1} - f_{SI}) + u^k\|_2^2 \quad (13)$$

$$u^{k+1} = u^k + d^{k+1} - \Psi^T(f^{k+1} - f_{SI}) \quad (14)$$

It is worth noting that each frame in the video sequences is projected in a block-by-block fashion Eq. (5). The variable $f$ is then separable, which means that the $f$-minimization step can be carried out via a number of scalar minimizations. Similarly, the variables $d$ and $u$ have also the component separability property. In other words, the subproblems Eq. (12) - Eq. (14) can be easily achieved by solving the corresponding component of each block ($i = 1$ to $B$), while the dictionary $D$ is learned in a frame level to provide sparse representation for all blocks. Therefore, we could obtain

$$\{D^{k+1}, x_i^{k+1}\} = \arg\min_{D,x_i} \frac{\lambda_1}{2}\sum_i \|R_i f^k - Dx_i\|_2^2 + \mu\sum_i \|x_i\|_0 \quad (15)$$

*for* $i = 1$ *to* $B$

$$\begin{aligned}f_i^{k+1} &= \arg\min_{f_i} \frac{1}{2}\|\Phi f_i - y\|_2^2 + \frac{\lambda_1}{2}\|f_i - D^{k+1}x_i^{k+1}\|_2^2 \\ &\quad + \frac{\rho}{2}\|d_i^k - \Psi^T(f_i - f_{SI,i}) + u_i^k\|_2^2 \\ d_i^{k+1} &= \arg\min_{d_i} \lambda_2\|d_i\|_1 + \frac{\rho}{2}\|d_i - \Psi^T(f_i^{k+1} - f_{SI,i}) + u_i^k\|_2^2 \\ u_i^{k+1} &= u_i^k + d_i^{k+1} - \Psi^T(f_i^{k+1} - f_{SI,i})\end{aligned}$$

*end*

In the following, we discuss the way of solving each subproblem.

1) $\{D, x\}$-subproblem

First of all, the problem Eq. (11) could be efficiently solved by the K-SVD algorithm. It can be worked out by the iteration between two procedures [22], [23]. In the first procedure, $D$ is fixed and the sparsest coefficients are found to agree with the data fidelity. This process is the CS recovery in essence. The second procedure is dictionary learning, which maintains the coefficients fixed while updating $D$ column by column through a rank-one approximation of a residual matrix by singular value decomposition (SVD). Although the K-SVD algorithm cannot guarantee to reach a global minimum, it still shows excellent performances in many applications [23].

In the DCVS framework, the temporal redundancy existed in video signals is mainly revealed by means of the correlation between the frame and its SI, which is generated by motion-compensated interpolation at the decoder. Consequently, we first extract $N$ training patches from $f_{SI}$ (along the temporal direction) in our scheme, where SI is divided into several overlapping blocks (along the spatial direction) and each block is vectorized as a column. Note that $f_{SI}$ is regarded as the initial value of $f_{NK}$ (to be shown in the initialization procedure of Algorithm 1). Then, we apply K-SVD algorithm to these patches to learn the dictionary $D$ for the current NK frame. Once $D$ is obtained from the training patches, we apply OMP to compute the sparse coefficient vector $x_i$ for each block.

2) $f$-subproblem

In the $f_i$-minimization step, as $\Phi^T\Phi + \lambda_1 I + \rho\Psi\Psi^T$ is a positive definite coefficient matrix and invertible, $f_i^{k+1}$ is an affine function given by

$$\begin{aligned}f_i^{k+1} = (\Phi^T\Phi + \lambda_1 I + \rho\Psi\Psi^T)^{-1}(\Phi^T y + \lambda_1 D^{k+1}x_i^{k+1} \\ +\rho\Psi(d_i^k + \Psi^T f_{SI,i} + u_i^k)) \qquad (16)\end{aligned}$$

Furthermore, the Eq. (16) can be easily solved by employing a direct method [19]. This method for solving a linear system $Fx = b$ are based on first factoring $F = F_1F_2\cdots F_k$ into a product of simpler matrices, and then computing $x = F^{-1}b$ by solving a sequence of problems of the form $F_i z_i = z_{i-1}$ where $z_1 = F_1^{-1}b$ and $x = z_k$. The solving step is sometimes also called a back-solve. The computational cost of factorization and back-solve operations depends on the sparsity pattern and other properties of $F$. The cost of solving $Fx = b$ is the sum of the cost of factoring $F$ and the cost of the back-solve. In our case, we use the Cholesky factorization to obtain $\Phi^T\Phi + \lambda_1 I + \rho\Psi\Psi^T = L * L^T$ with a lower triangular matrix $L$, and then $f_i^{k+1}$ could be computed as

$$f_i^{k+1} = (L^T)^{-1}L^{-1}(\Phi^T y + \lambda_1 D^{k+1} x_i^{k+1} + \rho\Psi(d_i^k + \Psi^T f_{SI,i} + u_i^k)) \tag{17}$$

3) $d$-subproblem

In the $d_i$-minimization step, even though $d_i$ in (15) is not differentiable, we can easily compute a simple closed-form solution to this problem by using subdifferential calculus. Explicitly, $d_i^{k+1}$ can be solved by the soft thresholding operator

$$d_i^{k+1} = Threshold\left(\Psi^T(f_i^{k+1} - f_{SI,i}) - u_i^k, \frac{\lambda_2}{\rho}\right) \tag{18}$$

wherein $Threshold(m,n) := \max\{m-n, 0\} - \max\{-m-n, 0\}$.

In summary, our proposed DL based reconstruction algorithm is outlined as follows.

**Algorithm 1**

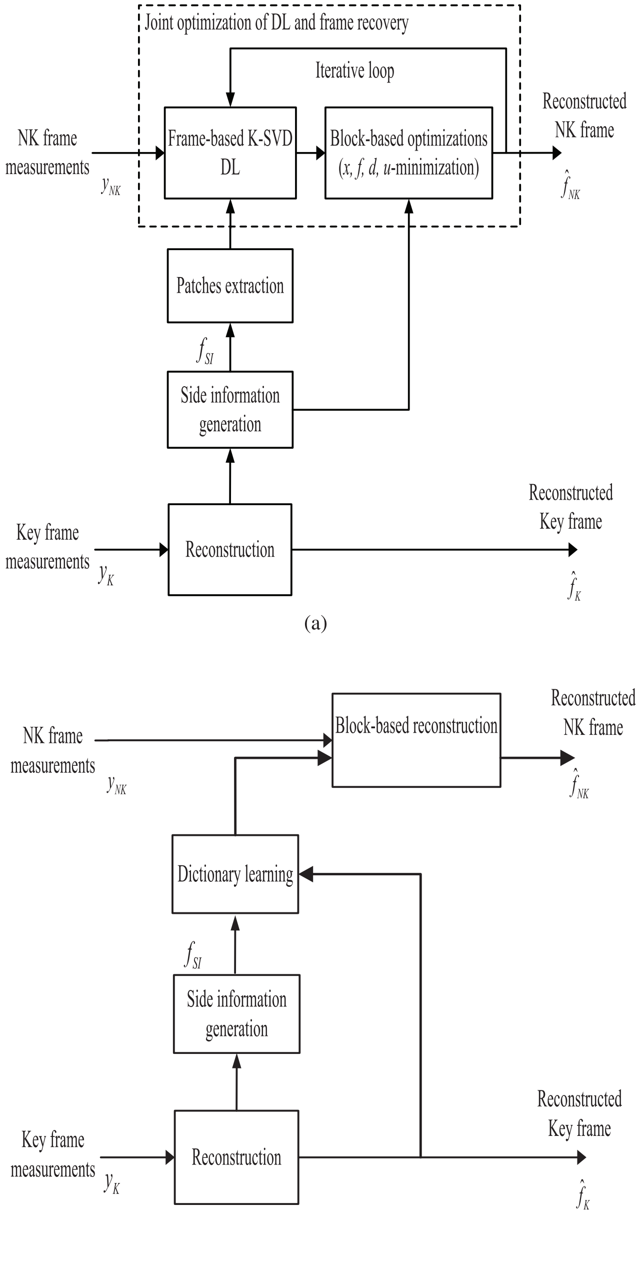
```
Input: $y$-received measurements which can be separated into blocks
    $y = [y_1^T, y_2^T, \dots, y_B^T]^T$
    $\lambda_1, \lambda_2, \rho$-regularization parameters
    $f_{SI}$-side information, $\Psi$-DCT transform basis, $\Phi$-(block-based) measurement matrix
    $tol$- stopping threshold, $num$-maximum iterations
Output: $f$-reconstructed NK frame
1: Initialization: $f^0 = f_{SI}$, $d^0 = 0$, $u^0 = 0$
   Cholesky factorization $\Phi^T\Phi + \lambda_1 I + \rho\Psi\Psi^T = L * L^T$
2: While $k < num$ and $\|f^{k+1} - f^k\|_2/\|f^k\|_2 < tol$ do
     Learn dictionary $D^{k+1}$ from a subset of $f^k$ blocks patches using K-SVD algorithm.
     For each block ($i = 1, 2, \dots, B$)
       a) Compute $x_i^{k+1}$ using OMP.
       b) Compute $f_i^{k+1}$ using Eq. (17).
       c) Compute $d_i^{k+1}$ using Eq. (18).
       d) Update $u_i^{k+1}$.
     Increase $k$.
3: End while
```

## 4. DCVS with Dictionary Learning Based Reconstruction

As shown in Fig. 1 (a), the architecture of DCVS with DL based reconstruction is presented. First of all, the video sequence is divided into groups of pictures (GOP), wherein the first frame in each GOP (i.e., the key frame) is "INTRA" recovered in order to avoid error accumulation and the other NK frames are "INTER" reconstructed using SI. It can then be implied that a significant low-complexity video encoder can be easily available, for the reason that only the projection process is required at the encoder, wherein all frames are compressively (and independently) projected and transmitted. The reconstruction process of frames at the decoder of DCVS is represented below. Besides, in Fig. 1 (b), We also show the DL-REC algorithm which performs the dictionary learning and reconstruction separately.

### 4.1 Key Frames

Similarly to the traditional DVC scheme, the first frames in each GOP are often encoded and decoded independently. In other words, the key frames $f_K$ are projected and reconstructed with the conventional CS scheme, by exploiting the signal sparsity property and solving the traditional $l_1$ minimization problem.



(a)

(b)

**Fig. 1** DCVS algorithms, (a) proposed DCVS decoder architecture with DL based reconstruction, (b) DL-REC

$$\min_x \frac{1}{2}\|y_K - \Phi_K\Psi x\|_2^2 + \mu\|x\|_1 \tag{19}$$

where $\Psi$ is the fixed sparsifying basis, e.g., DWT or DCT, $x$ is the sparse coefficient vector under the basis of $\Psi$, $\Phi_K$ is the measurement matrix for the measurement $y_K$ and $\mu$ is a non-negative parameter. Consequently, the recovered frame can be obtained via $\hat{f}_K = \Psi\hat{x}_K$, in which $\hat{x}_K$ is the solution of Eq. (19).

### 4.2 Non-Key Frames

In this algorithm, NK frames $f_{NK}$ are projected block by block so as to retain local information as much as possible and further improve the recovery effect, namely, $y_{NK} = \Phi_{NK} f_{NK}$, where $\Phi_{NK}$ is the diagonal measurement matrix as defined in Eq. (5).

At the decoder, for each $f_{NK}$, its SI is obtained by motion-compensated interpolation (MCI) of its previous and next reconstructed key frames. MCI technique has been successfully used for side information generation in DVC [1]. Then, we use the SI frame $f_{SI}$ to learn the dictionary for the current frame $f_{NK}$ as follows. i) $f_{SI}$ is divided into several non-overlapping blocks. For each block in the SI frame, we extract the five training patches, including the nearest four blocks overlapping this block and this block itself, where each extracted patch can be viewed as a column vector $D_i \in R^N$ where $N$ represents the size of the block. ii) The proposed DL based reconstruction method is then applied to recover the NK frame from limited numbers of measurements, wherein the sparse representation and signal recovery are jointly optimized. At each step of our presented iterative algorithm, we firstly apply the K-SVD algorithm [11] to the training patches $D_i$ to learn the frame-level dictionary $D \in R^{N \times dsize}$ along both the temporal and spatial directions, in which $D$ is a redundant dictionary containing $dsize$ atoms. Using the learned dictionary can usually provide sparser representation for $f_{NK}$ than using a fixed DCT or DWT basis. Second, the subproblems, $\{x, f, d, u\}$-minimizations, are solved respectively. As revealed in Algorithm 1, the dictionary is learned in a frame level, while the variables $\{x, f, d, u\}$ are updated by solving the corresponding component of each block. In other words, once the dictionary is learned using K-SVD, the current frame is then reconstructed in a block-by-block fashion.

## 5. Simulation Results

We test the performance of our proposed framework on two video sequences: Foreman and Mobile. The frame size is set to 176×144 (QCIF) and 352×288 (CIF), and only the luminance components are processed in our simulations. At the encoder, the size of GOP is set to 2, and both frames are split into non-overlapping blocks of size 16×16. (Note that the sizes of non-overlapping blocks and that of dictionaries also correspond to $N$ and $N \times dsize$ respectively as mentioned in the former Section. Here we set both of $N$ and $dsize$ to be 256). Each block is projected using the random Gaussian matrix with various sampling rates. At the decoder, the recovery of key frame in Eq. (19) is performed by employing the conventional ADMM algorithm, and the recovery of NK frame is solved by utilizing Algorithm 1, wherein the empirical parameters, $\lambda_1$, $\lambda_2$, $\rho$ and $\mu$, are set to 1 for simplicity. Note that the study of the values of the parameters is beyond the scope of this paper.

The proposed method is compared with two alternatives, i) the traditional basis pursuit (BP) method Eq. (4) and ii) the reconstruction method with K-SVD dictionary learning [8], [9]. Compared to our proposal, the conventional ADMM algorithm is utilized to achieve sparse recovery in the BP scheme wherein the DCT transform provides the sparsifying basis. In the Refs. [8], [9], the K-SVD algorithm is firstly employed to train the dictionary using blocks from the two adjacent key frames and SI, and then the frame recovery is achieved via sparse reconstruction Eq. (4). Thus, this method [8], [9] is denoted as the dictionary-learning-and-then-reconstruction scheme (DL-REC) in our work, since it requires the trained dictionary ahead of the reconstruction process. For a fair comparison, the same DL method is used in our proposal and DL-REC, i.e., only the patches obtained from SI are utilized to train the dictionary. Besides, the same test conditions are used for key frames, including the same Gaussian measurement matrix, the DCT sparsifying basis and the default ADMM sparse recovery algorithm.

In our simulations, the peak signal-to-noise ratio (PSNR) (as well as the perceptual quality) of the reconstructed video sequences are utilized to evaluate the reconstruction performance. Specially, different measurement rates ($MR$ varying from 0.1 to 0.5) are used for key frames and NK ones. Here, it should be noted that, in most papers related to video coding based on CS, frames are undersampled with identical $MR$, that is $MR$ of key frames ($MR_K$) is equivalent to that of NK frames ($MR_{NK}$). However, from the perspective of distributed CS [27], the performance of joint reconstruction can be achieved better by using smaller $MR_{NK}$ and larger $MR_K$, if the temporal redundancy between frames could be exploited at the decoder. Consequently, in the DCVS framework, key frames can be undersampled at an increased $MR_K$ in relation to $MR_{NK}$ [8], [11], [17]. Thus, we use the similar conditions described in [15], [16] in our experiment, taking the following two cases $MR_K = MR_{NK}$ and $MR_K > MR_{NK}$ into consideration. In the case of the latter situation, the average PSNR results are calculated only for NK frames, since the DL based recovery method we proposed is designed for NK frames (as demonstrated in Fig. 1 (a)). The total results of the two cases are demonstrated in Table 1-2, for QCIF and CIF video sequences separately.

As shown in Table 1, when $MR_K = MR_{NK}$ (varying from 0.1 to 0.5), our proposal increases about 0.92-4.14 dB over the BP scheme for QCIF sequences in the average PSNR, approximately 1.55-4.14 dB for Foreman, 1.07-2.3 dB for Mobile. When $MR_K > MR_{NK}$, $MR_K$ is set to 0.5 and $MR_{NK}$ changes with the range of 0.1 to 0.5. Under such circumstances, we could have a 2.33-15.31 dB gain in PSNR of NK frames in average (3.26-15.31 dB for Foreman, 2.33-7.58 dB for Mobile). For CIF sequences, we could obtain a 0.7-4.74 dB increase over the BP scheme when, and a 1.39-15.65 dB improvement when $MR_K = MR_{NK}$, as illustrated in Table 2. Furthermore, in comparison to the DL-REC

**Table 1** The performance comparison of QCIF sequences

| $MR_K$ | $MR_{NK}$ | Average PSNR of all frames(dB) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Foreman | | | Mobile | | |
| | | BP | DL-REC | Proposed | BP | DL-REC | Proposed |
| 0.1 | 0.1 | 10.34 | 14.22 | 14.48 | 10.91 | 12.59 | 13.21 |
| 0.2 | 0.2 | 19.74 | 21.30 | 21.51 | 15.88 | 16.15 | 16.98 |
| 0.3 | 0.3 | 24.21 | 25.06 | 25.76 | 17.75 | 17.69 | 18.82 |
| 0.4 | 0.4 | 27.09 | 27.71 | 28.74 | 19.54 | 19.27 | 20.68 |
| 0.5 | 0.5 | 29.37 | 29.74 | 30.99 | 21.10 | 20.61 | 22.26 |
| $MR_K$ | $MR_{NK}$ | Average PSNR of all NK frames(dB) | | | | | |
| | | Foreman | | | Mobile | | |
| | | BP | DL-REC | Proposed | BP | DL-REC | Proposed |
| 0.5 | 0.1 | 10.35 | 22.52 | 25.65 | 10.97 | 17.30 | 18.55 |
| 0.5 | 0.2 | 20.35 | 27.03 | 29.25 | 16.10 | 18.20 | 20.86 |
| 0.5 | 0.3 | 24.47 | 28.09 | 30.39 | 17.93 | 18.94 | 21.78 |
| 0.5 | 0.4 | 27.06 | 29.17 | 31.54 | 19.60 | 19.55 | 22.66 |
| 0.5 | 0.5 | 29.40 | 30.13 | 30.66 | 21.18 | 20.20 | 23.51 |

**Table 2** The performance comparison of CIF sequences

| $MR_K$ | $MR_{NK}$ | Average PSNR of all frames(dB) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Foreman | | | Mobile | | |
| | | BP | DL-REC | Proposed | BP | DL-REC | Proposed |
| 0.1 | 0.1 | 10.59 | 15.30 | 15.33 | 10.76 | 12.70 | 13.26 |
| 0.2 | 0.2 | 19.74 | 21.63 | 22.82 | 23.15 | 15.81 | 16.40 |
| 0.3 | 0.3 | 25.99 | 26.59 | 27.06 | 17.83 | 18.23 | 19.22 |
| 0.4 | 0.4 | 28.61 | 28.89 | 29.52 | 19.73 | 20.03 | 21.30 |
| 0.5 | 0.5 | 30.70 | 30.64 | 31.40 | 21.48 | 21.64 | 23.24 |
| $MR_K$ | $MR_{NK}$ | Average PSNR of all NK frames(dB) | | | | | |
| | | Foreman | | | Mobile | | |
| | | BP | DL-REC | Proposed | BP | DL-REC | Proposed |
| 0.5 | 0.1 | 10.59 | 25.32 | 26.24 | 10.8 | 17.39 | 19.21 |
| 0.5 | 0.2 | 22.36 | 27.10 | 29.06 | 16.06 | 18.70 | 21.27 |
| 0.5 | 0.3 | 26.29 | 28.28 | 30.10 | 17.99 | 19.78 | 22.45 |
| 0.5 | 0.4 | 28.57 | 29.46 | 31.02 | 19.79 | 20.83 | 23.74 |
| 0.5 | 0.5 | 30.68 | 30.56 | 32.07 | 21.55 | 21.88 | 25.10 |

scheme, our proposed scheme with the joint optimization of DL and signal recovery offers a PSNR gain of around 0.03-2.16 dB and 0.13-3.45 dB, in the following two cases $MR_K = MR_{NK}$ and $MR_K > MR_{NK}$ respectively. As indicated in the tables, a large PSNR gain could be easily got through the DL-REC mechanism than BP, in that more signal inter-frame correlation structures are exploited at the decoder. Besides, superior performance can be achieved using our presented method, since the sparse representation and frame recovery are jointly optimized in our proposal while the two processes are independently considered in DL-REC. Besides, it can also be seen that when $MR_K > MR_{NK}$, better PSNR results of NK frames could be obtained with key frames being undersampled at an increased sampling rate, since the more information could be extracted from SI with higher quality. For example, when $MR_K = 0.5$ and $MR_{NK} = 0.1$, PSNR of NK frames in the video sequence of Foremen QCIF is about 25.65dB using our proposal, while the PSNR result is only 18.63dB for NK frames in the case that both $MR_K$ and $MR_K$ are set to 0.1.

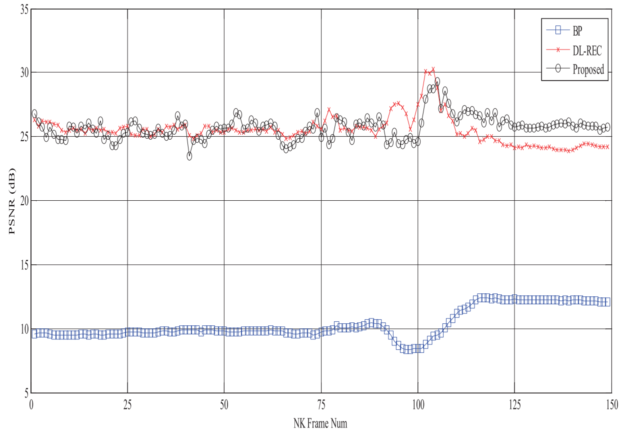


**Fig. 2** The performance comparison of each NK frame for Foreman (QCIF) when $MR_K = 0.5$ and $MR_{NK} = 0.1$

As shown in Figs. 2-3, the PSNR results of each NK frame when $MR_K = 0.5$ and $MR_{NK} = 0.1$ are demonstrated for the Foreman, Mobile sequences respectively. We can conclude that almost each frame is recovered with a PSNR gain by using our proposal than the BP and DL-REC scheme. It should be noted the quality outcomes of some frames experience a dramatic decrease, e.g., the 95th and 102th NK frames in the Foreman QCIF. The leading cause lies that the context transformations of these frames are relatively bigger, which results in the side information of inferior quality which fails to offer sound atoms to learn dictionaries. The performance contrasts for CIF video sequences

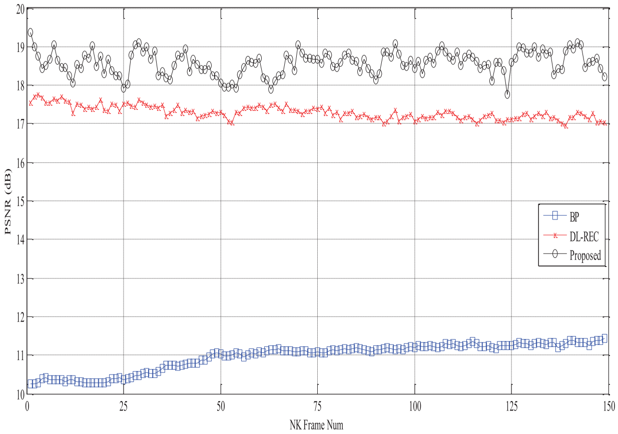


**Fig. 3** The performance comparison of each NK frame for Mobile (QCIF) when $MR_K = 0.5$ and $MR_{NK} = 0.1$

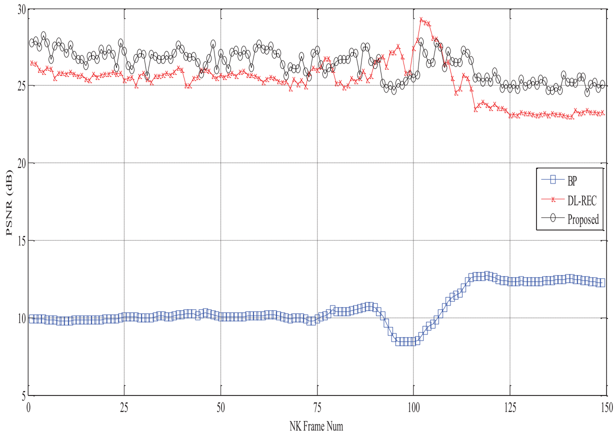


**Fig. 4** The performance comparison of each NK frame for Foreman (CIF) when $MR_K = 0.5$ and $MR_{NK} = 0.1$

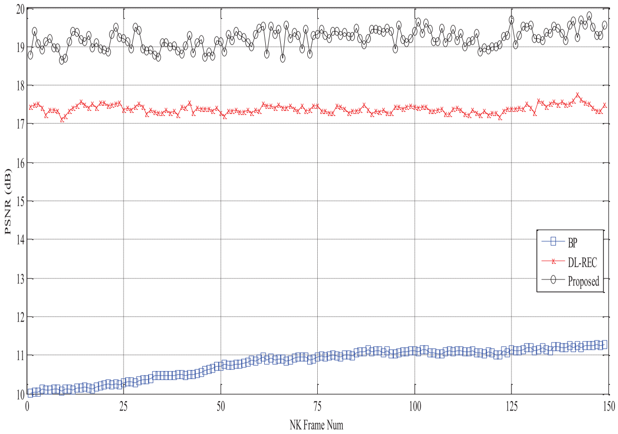


**Fig. 5** The performance comparison of each NK frame for Mobile (CIF) when $MR_K = 0.5$ and $MR_{NK} = 0.1$

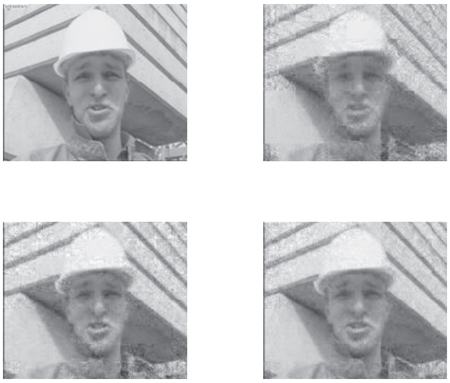


**Fig. 6** Recovered 52nd frame of Foreman (QCIF) when $MR_K = 0.3$ and $MR_{NK} = 0.3$. (Top left) The original frame. (Top right) BP, 24.01dB; (Bottom left) DL-REC, 25.6dB; (Bottom right) our proposal, 27.33 dB.

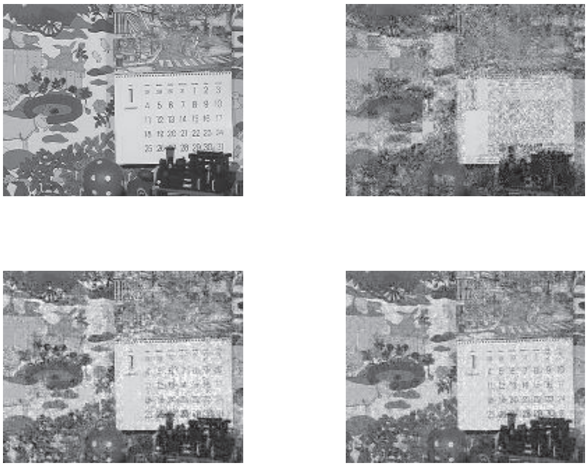


**Fig. 7** Recovered 52nd frame of Mobile (CIF) when $MR_K = 0.5$ and $MR_{NK} = 0.2$. (Top left) The original frame. (Top right) BP, 16.07dB; (Bottom left) DL-REC, 18.87dB; (Bottom right) our proposal, 21.4dB.

are shown in Figs. 4-5. We can conclude that PSNR results of nearly every NK frame recovered employing our method are better than those using DL-REC and BP, with the average PSNR improvement of 9.21-15.65 dB and 0.92-1.82 dB respectively. Additionally, Fig. 6 shows an instance of the reconstructed 52nd NK frames in the Foreman QCIF employing the three algorithms when both $MR_K$ and $MR_K$ are equal to 0.3. The recovered 52nd NK frames in the video sequence of Mobile CIF are also presented in Fig. 7 when $MR_K = 0.5$ and $MR_{NK} = 0.2$. It could be seen that better subjective image quality could be achieved by using our presented DL based recovery method.

Lastly, the convergence of our method is shown as an example by using a seven-time-iteration. To make it clear, we use the first 100 frames of the Foreman QCIF to perform the simulation. In specific, the key frames are reconstructed using DCT with different $MR$; NK frames are reconstructed using our proposal as the number of iterations varies from 1 to 7. The recovery results in terms of PSNR of NK frames are demonstrated in Fig. 8. It can be easily illustrated from the figure that more numbers of iterations could lead to higher-quality reconstruction. A favourable recovery effect can be reached via three-four iterations.

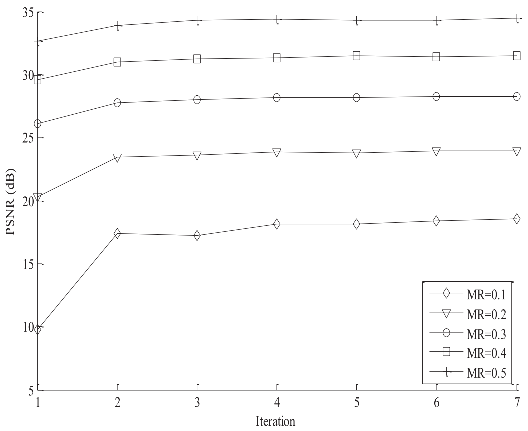


**Fig. 8** Variation of average PSNR value with the number of iterations for Foreman (QCIF).

## 6. Conclusion

We present a spatiotemporal dictionary learning based reconstruction method for DCVS, in which we attempt to enhance the reconstruction effect by jointly optimizing sparse reconstruction and signal recovery. In our work, we develop a novel joint optimization model, wherein the dictionary learning is included into the iterative algorithm as one single step and the frame recovery is achieved in an -analysis fashion under the correlation constraint. Furthermore, an ADMM based algorithm is also outlined to solve the underlying optimization problem. Experimental results show that the proposed method offers improved MR-distortion tradeoff and better reconstruction quality compared with the traditional methods. In order to further enhance the recovery performance using fewer measurements, some significant problems should be studied in the future: (1) efficient side information extraction methods; (2) more accurate correlation noise model; and (3) fast dictionary learning algorithms.

## Acknowledgements

This paper has been supported by the National Natural Science Foundation of China (Nos. 61271173 and 61372068), the Research Fund for the Doctoral Program of Higher Education of China (No. 20130203110005), the Fundamental Research Funds for the Central Universities (No. K5051301033), the 111 Project (No. B08038), and also supported by the ISN State Key Laboratory.

### References

[1] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, “Distributed video coding,” Proceedings of the IEEE, vol.93, no.1, pp.71–83, 2005.

[2] D. Slepian and J.K. Wolf, “Noiseless coding of correlated information sources,” IEEE Trans. Inf. Theory, vol.19, no.4, pp.471–480, 1973.

[3] A.D. Wyner and J. Ziv, “The rate-distortion function for source coding with side-information at the decoder,” IEEE Trans. Inf. Theory, vol.22, no.1, pp.1–10, 1976.

[4] E. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, vol.52, no.2, pp.489–509, 2006.

[5] D.L. Donoho, “Compressed sensing,” IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, 2006.

[6] J. Prades-Nebot, Y. Ma, and T. Huang, “Distributed video coding using compressive sampling,” Proc. Picture Coding Symposium, pp.1–4, 2009.

[7] T.T. Do, Y. Chen, D.T. Nguyen, N. Nguyen, L. Gan, and T.D. Tran, “Distributed compressed video sensing,” Proc. IEEE Int. Conf. on Image Process., pp.1393–1396, 2009.

[8] H.-W. Chen, L.-W. Kang, C.-S. Lu, P. Frossard, H. Li, F. Wu, B. Girod, S. Li, and G. Wei, “Dynamic measurement rate allocation for distributed compressive video sensing,” Proc. SPIE Visual Communications and Image Process., pp.774401–774410, 2010.

[9] H.-W. Chen, L.-W. Kang, and C.-S. Lu, “Dictionary Learning-based distributed compressive video sensing,” Proc. Picture Coding Symposium, pp.210–213, 2010.

[10] H. Liu, B. Song, H. Qin, and Z. Qiu, “Dictionary learning based reconstruction for distributed compressed video sensing,” J. Vis. Commun. Image R., vol.24, no.8, pp.1232–1242, 2013.

[11] W. Xu, Z. He, K. Niu, and J. Lin, “Sub-Sampling framework of distributed video coding,” Proc. IEEE Int. Symposium on Circuits and Systems, pp.1145–1148, 2010.

[12] M. Aharon, M. Elad, and A.M. Bruckstein, “The K-SVD: an algorithm for desigining of overcomplete dictionaries for sparse representation,” IEEE Trans. on Signal Processing, vol.54, no.11, pp.4311–4322, 2006.

[13] V. Thirumalai and P. Frossard, “Distributed representation of geometrically correlated images with compressed linear measurements,” IEEE Trans. on Image Processing, vol.21, no.7, pp.3206–3219, 2012.

[14] Y. Liu, M. Li, and D.A. Pados, “Motion-aware decoding of compressed-sensed video,” IEEE Trans. Circuits Syst. Video Technol., vol.23, no.3, pp.438–444, 2013.

[15] Z. Liu, A.Y. Elezzabi, and H.V. Zhao, “Maximum frame rate video acquisition using adaptive compressed sensing,” IEEE Trans. Circuits Syst. Video Technol., vol.21, no.11, pp.1704–1718, 2011.

[16] S. Pudlewski, A. Prasanna, and T. Melodia, “Compressed-sensing-enabled video streaming for wireless multimedia sensor networks,” IEEE Trans. Mobile Computing, vol.11, no.6, pp.1060–1072, 2012.

[17] H. Liu, B. Song, H. Qin, and Z. Qiu, “A Dictionary generation scheme for block-based compressed video sensing,” Proc. IEEE Int. Conf. on Signal Processing, Communication and Computing, pp.1–5, 2011.

[18] H. Liu, B. Song, H. Qin, and Z. Qiu, “An adaptive-ADMM algorithm with support and signal value detection for compressed sensing,” IEEE Signal Process. Lett., vol.20, no.4, pp.315–318, 2013.

[19] H. Liu, B. Song, F. Tian, and H. Qin, “Regularised reweighted BPDN for compressed video sensing,” Electronics Letters, vol.50, no.2, pp.83–84, 2014.

[20] H. Liu, B. Song, F. Tian, and H. Qin, “Joint sampling rate and bit-depth optimization in compressive video sampling,” IEEE Trans. Multimedia, vol.16, no.6, pp.1549–1562, 2014.

[21] C. Deng, W. Lin, B.-S. Lee, and C.T. Lau, “Robust image coding based on compressive sensing,” IEEE Trans. Multimedia, vol.14, no.2, pp.278–290, 2012.

[22] R. Glowinski and A. Marrocco, “Sur l'approximation, par elements finisd'ordre un, et la resolution, par penalisation-dualité d'une classe de problems de dirichlet non lineares,” Rev. Francaise d'Automat. Inf. Recherche Opérationelle, vol.9, pp.41–76, 1975.

[23] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximations,”

Computers and Mathematics with Applications, vol.2, pp.17–40, 1976.

[24] Y. Wang and L. Ying, "Compressed sensing dynamic cardiac cine MRI using learned spatiotemporal dictionary," IEEE Trans. on Biomedical Engineering, vol.61, no.4, pp.1109–1120, 2014.

[25] M.V. Afonso, J.M. Bioucas-Dias, and M.A.T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," IEEE Trans. Image Process., vol.19, no.9, pp.2345–2356, 2010.

[26] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," SIAM J. Imag. Sci., vol.1, no.3, pp.248–272, 2008.

[27] M.F. Duarte, S. Sarvotham, D. Baron, M.B. Wakin, and R.G. Baranuik, "Distributed compressed sensing of jointly sparse signals," Proc. Asilomar Conf. on Signals, Systems and Computers, pp.1537–1541, 2005.

**Fang Tian** received the M.S. degree in communication and information systems from Xidian University, Xi'an, China, in 2003. She is currently working toward the Ph.D. degree from Xidian University, Xi'an, China. Her research interests include video compression, communication, and compressed video sensing.

**Jie Guo** received the B.S. degree in communication engineering from Zhengzhou University, Zhengzhou, China, in 2011. She is currently working toward the Ph.D. degree from Xidian University, Xi'an, China. Her research interests include video compression, transmission, and compressed video sensing.

**Bin Song** received the B.S., M.S., and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1996, 1999, and 2002, respectively. He is currently a Professor at the School of Telecommunications Engineering, Xidian University, Xi'an, China. His research interests and areas of publication include video compression and transmission technologies, video transcoding, error- and packet-loss-resilient video coding, distributed video coding, video signal processing based on compressed sensing, and multimedia communications.

**Haixiao Liu** received the M.S. degree in communication and information systems from Xidian University, Xi'an, China, in 2010. He received the Ph.D. degree from Xidian University, Xi'an, China in Dec. 2014. His research interests include video compression, transmission, and compressed video sensing.

**Hao Qin** received the B.S., M.S., and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1996, 1999, and 2003, respectively. He is currently an Associate Professor at the School of Telecommunications Engineering, Xidian University, Xi'an, China. His research interests include wireless communication, video compression, and communication.