J. Vis. Commun. Image R. 24 (2013) 1232–1242

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Dictionary learning based reconstruction for distributed compressed video sensing

Haixiao Liu *, Bin Song, Hao Qin, Zhiliang Qiu

State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 29 August 2012
Accepted 15 August 2013
Available online 28 August 2013

Keywords:
Compressed sensing
Distributed video coding
Dictionary learning
Undersampled correlation noise model
Distributed compressed video sensing
Sparse recovery
Maximum likelihood
Energy minimization

ABSTRACT

Distributed compressed video sensing (DCVS) is a framework that integrates both compressed sensing and distributed video coding characteristics to achieve a low-complexity video coding. However, how to design an efficient reconstruction by leveraging more realistic signal models that go beyond simple sparsity is still an open challenge. In this paper, we propose a novel "undersampled" correlation noise model to describe compressively sampled video signals, and present a maximum-likelihood dictionary learning based reconstruction algorithm for DCVS, in which both the correlation and sparsity constraints are included in a new probabilistic model. Moreover, the signal recovery in our algorithm is performed during the process of dictionary learning, instead of being employed as an independent task. Experimental results show that our proposal compares favorably with other existing methods, with 0.1–3.5 dB improvements in the average PSNR, and a 2–9 dB gain for non-key frames when key frames are subsampled at an increased rate.

1. Introduction

Distributed video coding (DVC) [1] refers to a special video coding paradigm that encodes frames of a video sequence independently and decodes them jointly. As the temporal redundancies are exploited by the decoder exclusively, the computational burden is shifted from the encoder to the decoder, which makes DVC potentially applicable to many fields, e.g., wireless multimedia sensor networks (WMSN), video conferencing with mobile devices and surveillance systems. However, it still requires enormous data collection followed by data compression and thus, wastes valuable resources. Compressed sensing (CS) [2–4] is an innovative concept that has attracted considerable research interest in the signal processing community. It provides a new way to collect data incorporating both acquisition and compression, and consequently helps reduce the required number of measurements and transcend hardware limitations. Hence, the advantage of CS makes it a natural fit for DVC, due to the great reduction of sampling rate, power consumption and computational complexity.

Benefit from CS and DVC, distributed compressed video sensing (DCVS) [5–11] has recently emerged as a new way to directly capture video data via random projections at a low-complexity encoder, while performing joint reconstruction at a more complex decoder. The main challenge of DCVS is how to utilize the spatial/temporal redundancy in video at the decoder to achieve sparse representation and efficient reconstruction. One of the earlier works addressing DCVS was presented by Prades-Nebot et al. [5], in which a video sequence is divided into key frames and non-key (NK) frames. Key frames are intra encoded and decoded using traditional video compression standards; while NK frames are projected and recovered using CS techniques, with an adaptive redundant dictionary built by picking blocks from previously reconstructed frames. A similar method was proposed in [6], introduced as an inter-frame sparsity model. However, in these schemes, it is still required to capture huge amounts of raw video data for key frames, which are encoded using conventional compression algorithms.

Another DCVS framework was proposed in [7,8], wherein the dictionary learning algorithm K-SVD [12] is directly employed by extracting samples from previous recovered frames together with the side information. As soon as the trained dictionary is obtained, NK frames are reconstructed by using the conventional sparse recovery algorithms. In this method, sparse representation and reconstruction are designed as independent tasks. However, this has a negative impact in terms of consuming resources, as the sparse coefficient calculation has already been included in the process of dictionary learning. Besides, a scalable framework of DCVS was presented in [9] to achieve optimal quality of service. In [10], an initialization and several stopping criteria were proposed for NK frames to speed up the convex optimization, and in [11] a measurement compression scheme by using the channel coding was proposed. Note that there also exist other literatures about

* Corresponding author. Fax: +86 29 88204409.
E-mail addresses: hxliu@stu.xidian.edu.cn (H. Liu), bsong@mail.xidian.edu.cn (B. Song), hqin@mail.xidian.edu.cn (H. Qin), zlqiu@mail.xidian.edu.cn (Z. Qiu).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.08.007

CS-based video coding [13–17], e.g., a new dictionary generation scheme using an iterative fashion between reconstructing and filtering [15] and an adaptive-ADMM algorithm for CS with partial known support and signal value information [17] were proposed in our previous work. Nevertheless, most of these techniques, which are aimed to explore temporal/spatial redundancy at the encoder and achieve higher sampling efficiency, are not suited for DVC as far as limited resource is concerned.

In this paper, we propose a dictionary learning based reconstruction algorithm for DCVS. Our goal is to improve the reconstruction performance by leveraging more realistic signal models that go beyond simple sparsity and compressibility (by including the video signal structure), while retaining very low computation complexity at the encoder. One of our contributions is to introduce a novel correlation noise model (CNM) between the original video frame and its side information (SI) when video sequences are compressively sampled at a rate that is far below the Nyquist rate. To distinguish from the conventional notation in standard DVC, we denote our model as the "undersampled" CNM. To be specific, a new statistical model is presented in this work to characterize the error pattern of the correlation noise, and then offers an efficient way to describe the temporal correlation in undersampled videos. Another main contribution of this paper is that we propose a dictionary learning based reconstruction scheme, wherein we try to learn a dictionary that efficiently describes the content of video frames, and simultaneously permits to capture the correlation in sequences by including the CNM constraint. In this respect, we concentrate on the problem of two views and develop a maximum likelihood (ML) method. In our algorithm, the ML optimization is cast as an energy minimization problem, which can then be solved by iterating reconstruction and dictionary update. Consequently, our recovery method can achieve an efficient sparse representation for DCVS, and at the same time obtain the corresponding coefficients to recover video signals. In other words, both the dictionary learning and reconstruction are performed under the correlation constraint in order to achieve a good visual quality. To the best of our knowledge, there is no literature available to analyze CNM when the video sequence is compressively sampled, or to formulate the dictionary learning for DCVS with the prior on CNM.

Lastly, it is worth noting that in this paper we mainly focus on developing a dictionary learning based reconstruction algorithm for DCVS, which provides a novel fully low-complexity video compression paradigm and an alternative scheme adaptive to the environment where raw video data is not available, instead of competing compression performance against the current compression standards or DVC schemes, which need raw data available for encoding.

The rest of this paper is organized as follows. The overview of background is given in Section 2. The proposed ML dictionary learning method is described in Section 3. Section 4 presents the DCVS reconstruction with dictionary learning. Simulation results are described in Section 5, followed by conclusions in Section 6.

2. Background

2.1. Compressed sensing

Suppose that f is a discrete signal of length n, and let x be its coefficients in some orthonormal basis $\Psi \in R^{n\times n}$. Signal f is said to be k-sparse with respect to Ψ if only its $k$coefficients are nonzero. According to the CS theory, a k-sparse signal can be acquired through the linear random projections $y = \Phi f$, where $y \in R^m$ is the sampled vector with $m < n$ and Φ is an $m \times n$ measurement matrix that is incoherent with Ψ. Here we define the measurement rate (MR) for the signal as

$$MR = m/n. \tag{1}$$

More specifically, the measurement y is a random linear combination of the entries of f, which can be viewed as the compressed version of f. Although the recovery of the signal from the measurement is an ill-posed problem since $m < n$, the CS theory states that the reconstruction can be achieved by solving the following l_1 minimization problem [2,3]

$$\hat{x} = \arg\min ||x||_1, \quad \text{subject to } y = \Phi\Psi x. \tag{2}$$

This convex optimization problem, namely basis pursuit (BP), can be recast as a linear program to be efficiently solved. Many other algorithms, such as matching pursuit, can also be employed to recover the coefficients. Note that, signals of interest in practice are often not sparse but approximately sparse, i.e., their coefficients are generally different to zero, although only a small number of them have significant amplitude values. It has been proven that, under certain conditions [4], the solution to (2) can still recover the most significant coefficients, and hence, provides a good approximation of the signal.

2.2. Distributed video coding

In a typical DVC solution [1], the video frames are categorized into key frames and Wyner-Ziv (WZ) frames. Key frames are intra-coded by traditional video compression standards such as H.264/AVC, while WZ frames are intra-frame encoded but inter-frame decoded. At the encoder side, without performing motion estimation, the compression of a WZ frame f_{WZ} is achieved by transmitting only part of the parity bits derived from the channel-encoded version of f_{WZ}. At the decoder, the side information f_{SI} is first generated by motion-compensated interpolation, thus could be viewed as a noisy version of the original WZ frame. Then the decoder uses the received parity bits and f_{SI} to recover f_{WZ}. By exploiting the source statistics at the decoder side, the major computation complexity of DVC is shifted from the encoder to the decoder.

In this paper, we mainly focus on DCVS that combines the advantages of both CS and DVC. We propose to learn dictionaries that can efficiently describe the content of NK frames and simultaneously permit to capture the correlation structure of video signals. In this respect, we present a dictionary learning based reconstruction algorithm for DCVS with a maximum likelihood method, in which a new probabilistic model is formulated by including a sparsity prior and a novel undersampled CNM constraint.

3. ML learning of dictionaries for DCVS

3.1. Problem formulation

The conventional DCVS structure is employed in our paper (to be shown in Fig. 1), wherein the key frame f_K is projected and reconstructed using the orthonormal basis Ψ and the traditional CS recovery algorithm. For the NK frame f_{NK}, it is first split into several non-overlapping $b \times b$ blocks. Each block is vectorized as $f_{NK,b} \in R^n$ $(n = b^2)$ and projected using the random measurement matrix $\Phi \in R^{m\times n}$, i.e., $y_{NK,b} = \Phi f_{NK,b}$. Then the measurement $y_{NK,b} \in R^m$ is transmitted to the decoder.

Now, we begin to formulate the probabilistic framework for the ML learning of a redundant dictionary $D \in R^{n\times n}$ that is used to recover the NK frame block $f_{NK,b}$ (if the dictionary is overcomplete, $D \in R^{n\times l}$ and $l > n$). To be specific, our ultimate goal is to improve the reconstruction quality through learning a sparse representation dictionary D, i.e., $\min ||f_{NK,b} - \hat{f}_{NK,b}||_2$, where $\hat{f}_{NK,b} = D\hat{x}$ is the recovered block and its coefficient $\hat{x}$ is extremely sparse. Although

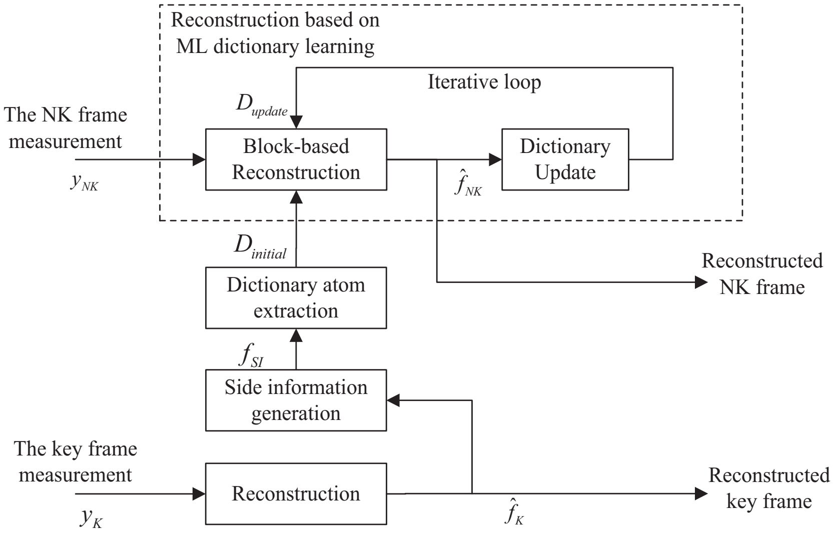

Fig. 1. Proposed DCVS reconstruction with ML dictionary learning.

the original $f_{NK,b}$ is unavailable at the decoder, this optimization can be easily achieved through $\min\|y_{NK,b} - \Phi\hat{f}_{NK,b}\|_2$ by using the random measurement matrix as implied in the CS theory.

Inspired by the work of Tošić and Frossard for stereo images [18], we define the likelihood that the measurement $y_{NK,b}$ is well approximated by the projection of $\hat{f}_{NK,b}$ under a sparsity prior and an undersampled CNM constraint. In other words, we want to learn a dictionary that can efficiently describe the content of frames and simultaneously permit to capture the correlation in video sequences, by including the signal structure in the form of CNM. Therefore, we need to maximize the probability that the observed $y_{NK,b}$ is well approximated by the projection of the recovered $\hat{f}_{NK,b}$, where $\hat{f}_{NK,b}$ can be sparsely represented by D and the undersampled CNM constraint between $\hat{f}_{NK,b}$ and f_{SI} is also satisfied. Hence, the goal of learning is to find the redundant dictionary D^* that is the solution of the following optimization problem:

$$D^* = \underset{D}{\operatorname{argmax}}\{\log P(y_{NK,b}, f_{SI}|\Phi, D)\} \tag{3}$$

in which the SI f_{SI} is introduced to characterize the undersampled CNM constraint. To solve this optimization problem, we employ the prior knowledge that the block has a sparse representation with respect to the dictionary. Then applying the chain rule, we can approximate the probability in (3) as

$$P(y_{NK,b}, f_{SI}|\Phi, D) = P(y_{NK,b}, f_{SI}|\Phi, D, x)P(x|\Phi, D) \tag{4}$$

Considering the fact that $P(y_{NK,b}, f_{SI}|\Phi, D, x) \leqslant \min\{P(y_{NK,b}|\Phi, D, x), P(f_{SI}|\Phi, D, x)\}$, the problem in (3) can be solved through $\max P(y_{NK,b}|\Phi, D, x)$ and $\max P(f_{SI}|\Phi, D, x)$, which are implemented by using $\max\{P(y_{NK,b}|\Phi, D, x)P(f_{SI}|\Phi, D, x)\}$ in our algorithm, since f_{SI} does not bring more information to $y_{NK,b}$ than Φ, D, x. Then we can deduce the approximation of D^* as

$$\begin{aligned} D^* &= \underset{D}{\operatorname{argmax}}\{\max_x[\log P(y_{NK,b}, f_{SI}|\Phi, D, x)P(x|\Phi, D)]\} \\ &\approx \underset{D}{\operatorname{argmax}}\{\max_x[\log P(y_{NK,b}|\Phi, D, x)P(f_{SI}|\Phi, D, x)P(x|\Phi, D)]\}, \end{aligned} \tag{5}$$

where the objective function consists of three components: (1) the measurement likelihood $P(y_{NK,b}|\Phi, D, x)$; (2) the prior on the coefficients $P(x|\Phi, D)$, and (3) the CNM constraint $P(f_{SI}|\Phi, D, x)$. In the following, each of the three terms is evaluated.

3.2. Measurement likelihood

The measurement likelihood $P(y_{NK,b}|\Phi, D, x)$ actually depicts the projection of the approximation error e, since

$$\begin{aligned} P(y_{NK,b}|\Phi, D, x) &= P(y_{NK,b} - \Phi Dx) = P(\Phi f_{NK,b} - a\Phi Dx) \\ &= P[\Phi(f_{NK,b} - Dx)] = P(\Phi e), \end{aligned} \tag{6}$$

and it can be modeled as a zero-mean Gaussian noise as a consequence of the Central Limit Theorem [19]. The reason is that the projection is indeed a linear combination of the approximation error e, in which the i-th component e_i $(i = 1, 2, \ldots, n)$ is assumed to have independent and identical distribution. Then the measurement likelihood can be rewritten as

$$\begin{aligned} P(y_{NK,b}|\Phi, D, x) &= P(y_{NK,b} - \Phi Dx) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\|y_{NK,b} - \Phi Dx\|_2^2\right), \end{aligned} \tag{7}$$

where σ^2 is the variance of the Gaussian noise.

3.3. Prior on coefficients

The prior on the sparse coefficients $P(x|\Phi, D)$ is the conditional probability of x, given the measurement matrix Φ and the dictionary D. Since x is the sparse representation coefficient of $f_{NK,b}$, it is only correlated with the dictionary D. Therefore, we have $P(x|\Phi, D) = P(x|, D)$, which can be further modeled as (8), an effective and widely applicable method in literatures. See, e.g., [20,21].

$$P(x|D) = \prod_{i=1}^{n} P(x_i|D) = \left(\frac{1}{C_\beta}\right)^n \exp\left(-\beta\sum_{i=1}^{n} S(x_i)\right) \tag{8}$$

in which β is a parameter that controls its steepness, C_β is the normalizing constant, x_i $(i = 1, 2, \ldots, n)$ represents the i-th coefficient and S determines the distribution shape of the sparse coefficients. Here we defer the formulation of the function S to Section 4, but emphasize that the distribution of coefficients S depends critically on the dictionary D, which is generated in DCVS by using the temporal/spatial redundancies in videos. It should be noted that we have used in (8) the assumption that the statistic independence of x_i is incorporated, which results in a factorial distribution of $P(x|D)$.

3.4. Undersampled CNM

$P(f_{SI}|\Phi, D, x)$ is the conditional probability of the side information f_{SI}, given Φ, D and x. Since SI is generated from motion-compensated interpolation in pixel-domain, f_{SI} is uncorrelated with the measurement matrix Φ. Thus we have $P(f_{SI}|\Phi, D, x) = P(f_{SI}|D, x)$.

Based on the Wyner-Ziv theorem [1], the statistical dependency between a frame f and its estimation f_{SI} could be modeled as a virtual correlation channel, where f_{SI} is the so-called SI and can be viewed as a noisy version of f. Moreover, the correlation between f and f_{SI} is often formulated as a CNM that follows the Laplacian distribution [22],

$$P(f_{SI}|D, x) = P(f_{SI} - Dx) = \left(\frac{\alpha_0}{2}\right)^n \exp(-\alpha_0 ||f_{SI} - Dx||_1), \tag{9}$$

where α_0 is the Laplacian distribution parameter which constitutes a good tradeoff between model accuracy and complexity, D denotes the sparsifying dictionary and x is its coefficient vector.

However, in the DCVS framework, the measurement paradigm consists of linear projections of video frames or blocks, into a small set of measurement vectors. When the key frames are subsampled at a quite low rate MR_K (e.g., smaller than a threshold MR_{Th}), these frames and their interpolated SI at the decoder would endure significantly poor-quality reconstruction as the insufficient number of measurements. In such case, the CNM based on Laplacian distribution (9) cannot accurately describe the error pattern any more. Here, we propose a new distribution as defined in (10) to model the "undersampled" correlation noise.

$$\begin{aligned} P(f_{SI}|D, x) &= P(f_{SI} - Dx) \\ &= \left(\frac{1}{\alpha_1}\right)^n \exp\left(-\frac{1}{\alpha_2^2} ||f_{SI} - Dx - \alpha_3 I||_2^2\right), \end{aligned} \tag{10}$$

in which α_i $(i = 1, 2, 3)$ are the non-zero constants and I is the unit vector.

For example, we take the Foreman QCIF sequence with $MR_K = 0.1$ to evaluate the presented undersampled CNM and set the group of pictures (GOP) to be 2. As shown in Fig. 2, the actual histogram of correlation noise and the distribution (10) (with $\alpha_1 = 0.21$, $\alpha_2 = 30.5$, and $\alpha_3 = 13.5$) are depicted, respectively. Additionally, the Laplacian distributions (9) with $\alpha_0 = 0.05$, 0.1, 0.3, 0.5 are also illustrated. It can be easily seen that the presented distribution model could fit the actual correlation noise $(f_{NK} - f_{SI})$ more accurately than (9). Moreover, to further verify the above hypothesis, we use Curve Fitting Tool of MATLAB to test its goodness-of-fit (GOF). Here the GOF of a function $f(x)$ is defined as

$$R = 1 - \frac{\sum_{i=1}^{n} (f(x_i) - p_i)^2}{\sum_{i=1}^{n} (p_i - \bar{p})^2}, \tag{11}$$

where p_i is the actual probability density at x_i, and $\bar{p}$ is its average value. It should be noted that the closer R; approaches to 1, the better the goodness.

A summary of the results for various QCIF video sequences is presented in Table 1, wherein the parameters α_i $(i = 1, 2, 3)$ are estimated from data fitting. For a typical significance GOF level of 0.98–0.99, those values imply accepting the hypothesis that the correlation noise indeed follows the model (10) when the sampling rate of key frames MR_K is comparatively small. In a word, in order to accurately characterize the correlation noise in DCVS, a more realistic undersampled CNM is presented especially when video signals are compressively sampled, i.e., different statistic distributions, (9), (10), should be employed for different MR.

3.5. Energy minimization optimization

In the above discussion, we have defined all three components of the objective function in (5). To solve the ML optimization problem of (5), an equivalent energy minimization problem is introduced as:

$$D* = \underset{D}{\operatorname{argmax}}\{\min_x E(y_{NK,b}, f_{SI}, x, \Phi, D)\}, \tag{12}$$

in which E denotes the energy function as defined in (13). Here the normalization constants are omitted since they do not influence the optimization problem.

$$E(y_{NK,b}, f_{SI}, x, \Phi, D) = \begin{cases} \frac{1}{2\sigma^2} ||y_{NK,b} - \Phi Dx||_2^2 + \beta \sum_{i=1}^{n} S(x_i) + \alpha_0 ||f_{SI} \\ -Dx||_1 \quad MR_K > MR_{Th}, \\ \frac{1}{2\sigma^2} |y_{NK,b} - \Phi Dx||_2^2 + \beta \sum_{i=1}^{n} S(x_i) + \frac{1}{\alpha_2^2} ||f_{SI} - Dx \\ -\alpha_3 I||_2^2 \quad MR_K \leqslant MR_{Th}. \end{cases} \tag{13}$$

As depicted in (13), the energy function thus consists in the sum of three main terms:

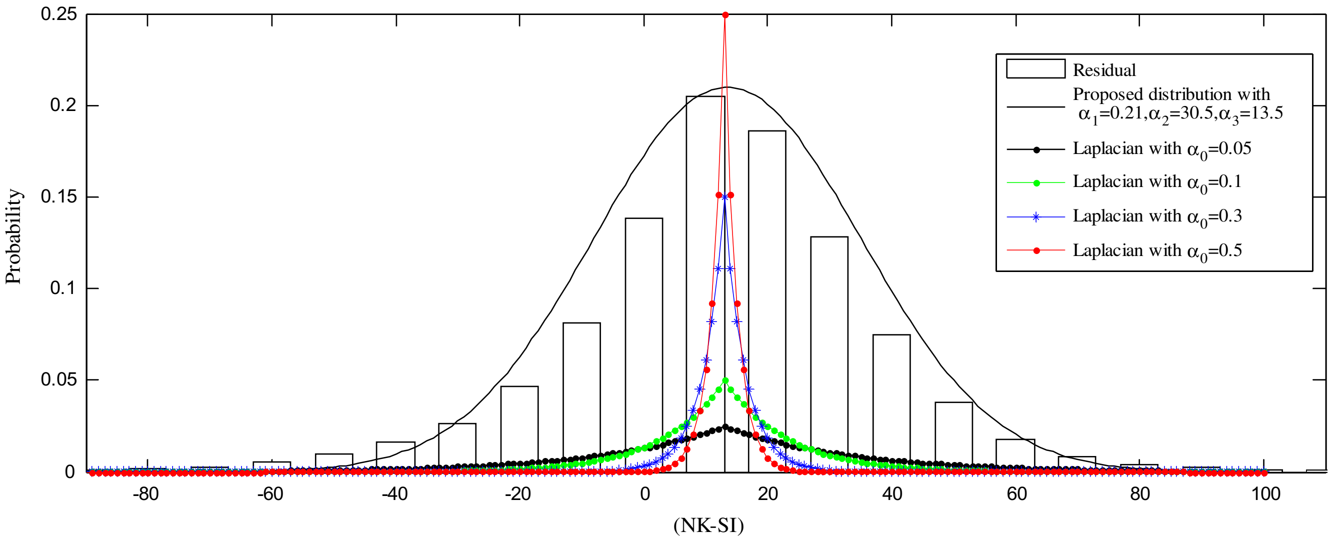

Fig. 2. Histogram of the correlation noise for Foreman (QCIF) with $MR_K = 0.1$.

Table 1
The GOF test results of correlation noise for QCIF sequences with $MR_K = 0.1$.

Sequences	Parameters α_i $(i = 1, 2, 3)$	Goodness-of-fit (GOF)
Foreman	0.194, 27.27, 13.88	0.997
Mobile	0.098, 55.83, 18.31	0.9975
News	0.1701, 29.42, 13.15	0.9849
Carphone	0.1999, 25,82, 12.18	0.9852
Coastguard	0.1996, 25.81, 12.51	0.9878

(1) the data fidelity term $||y_{NK,b} - \Phi Dx||_2^2$, expressed by the energy of the projection of approximation errors. Based on this, our algorithm can improve the reconstruction video quality (in the pixel domain) through the measurement domain, in the sense that the measurement term $||y_{NK,b} - \Phi Dx||_2^2$ is actually introduced to achieve the minimization of $||f_{NK,b} - \hat{f}_{NK,b}||_2$.
(2) the sparsity term $\sum S(x_i)$, characterizing the sparsity degree of the coefficients with respect to the given dictionary;
(3) the undersampled CNM constraint term $||f_{SI} - Dx||_1$ or $||f_{SI} - Dx - \alpha_3 I||_2^2$, expressing the correlation between the frame and its side information, which follows the statistical distribution (9), (10) for different MR_K.

The energy minimization problem can then be solved by iterating between two steps [20,21]. In the first step, D is kept constant and the energy function is minimized with respect to a set of coefficients x. This step is essentially the CS reconstruction, also called sparse coding. The second step is dictionary learning. It keeps the coefficients constant, while performing the gradient descent on D to minimize the energy. Therefore, this alternating optimization process iterates between reconstruction and dictionary update steps until convergence. The detail implementation of the optimization process for DCVS is described in the following section.

4. DCVS reconstruction with ML dictionary learning

We are now ready to present the DCVS reconstruction architecture based on ML dictionary learning. As shown in Fig. 1, the general structure of DCVS is employed. The measurements of key frames and NK frames are transmitted independently, wherein the quantization and entropy coding of measurements are not considered, since they are beyond the scope of this paper. It can be easily implied that, by emerging the CS and DVC technologies, a significant low-complexity video coding will be easily achieved, since both key frames and NK frames are compressively projected. The reconstruction of key frames and NK frames at the DCVS decoder are described, respectively as follows.

4.1. Key frames

At the DCVS decoder, each key frame f_K is reconstructed via solving the minimization problem as:

$$\min_x \frac{1}{2}||y_K - \Phi_K \Psi x||_2^2 + \tau||x||_1, \tag{14}$$

where Ψ is the discrete wavelet transform (DWT) basis, x is the sparse coefficient vector with respect to Ψ, Φ_K is the measurement operator for f_K and τ is a non-negative parameter. In order to achieve the independence of key frames coding, DWT is employed as the sparse representation basis, and the frame-based projection is applied for key frames in our scheme (as the frame-based sensing matrix is denser, that implies more incoherent in the sparsifying domain). The recovered key frame $\hat{f}_K = \Psi\hat{x}_K$, in which $\hat{x}_K$ is the solution of (14). Finally, the SI frame f_{SI} is generated from motion-compensated interpolation, using the previous and next recovered key frames.

4.2. Non-key frames

In our DCVS framework, block-based projection is employed for NK frames f_{NK} in order to preserve more local information that can improve the reconstruction performance, i.e., the measurement $y_{NK,b}$ is obtained via $y_{NK,b} = \Phi f_{NK,b}$, where $f_{NK,b}$ is the vectorized block in f_{NK} and Φ is the random measurement matrix. Before we proceed further, we would like to first discuss the dictionary generation method and its sparse coefficient distribution in DCVS.

4.2.1. Dictionary initialization

As shown in Fig. 1, an initial dictionary $D_{initial}$ is firstly built for each block $f_{NK,b}$ by using its SI f_{SI}. As we know, if the dictionary can be learned based on training samples/atoms extracted from the

Table 2
The GOF test results of $S(x_i) = |x_i|$ for QCIF sequences.

Sequences	Parameters (C_β, β)	Goodness-of-fit
Foreman	1.603, 36.83	0.9905
Mobile	1.602, 35.97	0.9904
News	1.447, 38.32	0.9973
Carphone	1.54, 37.29	0.9942
Coastguard	1.621, 36.11	0.9892

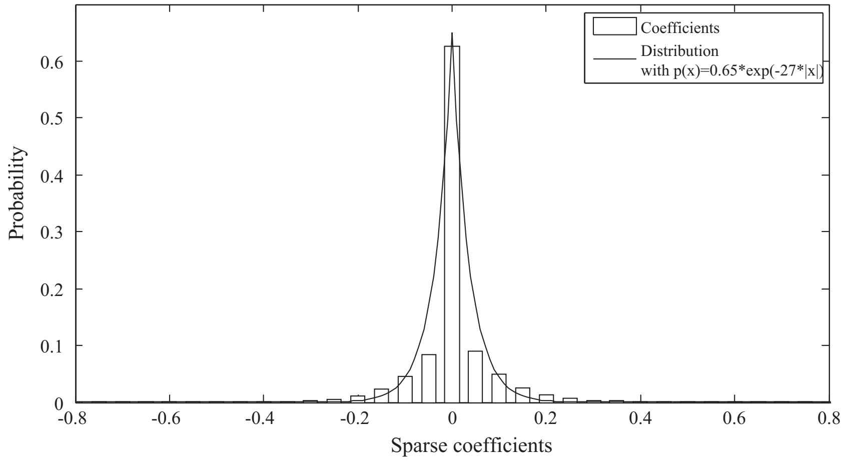

Fig. 3. The sparse coefficient distribution of NK frames for Foreman (QCIF).

Table 3
The performance comparison of QCIF sequences.

MR_K	MR_{NK}	Foreman			Mobile			News		
		DL-REC	Baseline	Proposed	DL-REC	Baseline	Proposed	DL-REC	Baseline	Proposed
Average PSNR of all frames (dB)										
0.1	0.1	20.44	20.71	20.92	15.34	15.29	15.38	18.14	18.19	18.32
0.2	0.2	23.70	24.25	24.47	17.05	17.04	17.38	21.35	21.35	21.59
0.3	0.3	26.37	27.28	27.46	18.70	18.83	19.01	24.24	24.56	24.60
0.4	0.4	28.61	29.91	30.17	20.35	20.65	21.17	26.67	27.24	27.69
0.5	0.5	30.71	32.23	32.66	21.85	22.38	23.19	27.68	30.01	30.99
Average PSNR of all NK frames (dB)										
0.5	0.1	25.79	29.76	30.79	17.78	20.24	21.70	23.27	26.72	28.57
0.5	0.2	27.56	31.33	32.03	19.46	21.24	22.43	26.21	28.14	29.73
0.5	0.3	28.73	32.38	33.12	20.40	21.91	23.16	27.50	29.09	30.83
0.5	0.4	30.25	33.22	34.07	21.25	22.46	23.86	28.32	30.02	31.77
0.5	0.5	30.87	33.92	34.80	21.90	22.96	24.60	25.97	30.65	32.60

frame itself, this redundant dictionary could provide much sparser representation. Although the original frame is not available at the decoder, a good approximation can still be obtained by using the statistical dependency in DVC, i.e., the side information. Following the method in our previous work [15], we generate the initial dictionary $D_{initial}$ for each block $f_{NK,b}$ by extracting neighboring blocks in f_{SI} as the atoms.

4.2.2. Sparse coefficient distribution

As mentioned in Section 3.3, the prior on sparse coefficients can be modeled as (8) and the function S determines the distribution shape of the coefficients. However, in our DCVS framework, the dictionary $D_{initial}$ is first initialized by picking up the neighboring blocks in SI. Therefore, the recovered NK frame coefficients with respect to $D_{initial}$ will be extremely sparse. In this paper, we present the model $S(x_i) = |x_i|$ to characterize the sparse coefficients for DCVS. For example, we plot the actual distribution of recovered coefficients for all NK frames in the Foreman QCIF sequence, as shown in Fig. 3, where key frames are recovered using (14) with $MR_K = 0.5$ and GOP = 2. The size of $D_{initial}$ and $f_{NK,b}$ are set to 256×256 and 16×16, respectively. Besides, the distribution (8) with $S(x_i) = |x_i|$ and $C_\beta = 1/0.65$, $\beta = 27$ has also been plotted in Fig. 3. It is quite obvious that the distribution with $S(x_i) = |x_i|$ is basically in accordance with the actual sparse coefficient histogram.

Similarly, in order to further verify the proposed model, we also use Curve Fitting Tool of MATLAB to test GOF of (8) (here $f(x_i) = |x_i|$). The results for various QCIF video sequences are presented in Table 2, which imply that the proposed distribution $p(x_i) = (1/C_\beta) * \exp(-\beta|x_i|)$ (with GOF values of 0.98–0.99) can accurately characterize the actual sparse coefficients with respect to $D_{initial}$. As a result, $\sum S(x_i) = \sum |x_i| = ||x||_1$, and the energy function (13) can then be rewritten as:

$$E(y_{NK,b}, f_{SI}, x, \Phi, D) = \begin{cases} \frac{1}{2\sigma^2}||y_{NK,b} - \Phi Dx||_2^2 + \beta||x||_1 & \\ \quad +\alpha_0||f_{SI} - Dx||_1 & MR_K > MR_{Th}, \\ \frac{1}{2\sigma^2}||y_{NK,b} - \Phi Dx||_2^2 + \beta||x||_1 & \\ \quad +\frac{1}{\alpha_2^2}||f_{SI} - Dx - \alpha_3 I||_2^2 & MR_K \leqslant MR_{Th}. \end{cases} \tag{15}$$

4.2.3. Block-based reconstruction and dictionary update

Based on the above discussion, the energy minimization can be solved iteratively between reconstruction and dictionary update steps. In the framework of DCVS (as shown in Fig. 1), the block-based reconstruction, i.e., the sparse coding step in the energy minimization problem (12), is indeed performed by (16). Here, the non-negative parameters λ_i $(i = 1, 2, 3)$ are introduced for tradeoff.

$$\min_x \begin{cases} \frac{1}{2}||y_{NK,b} - \Phi Dx||_2^2 + \lambda_1\beta||x||_1 & \\ \quad +\lambda_2\alpha_0||f_{SI} - Dx||_1 & MR_K > MR_{Th}, \\ \frac{1}{2}||y_{NK,b} - \Phi Dx||_2^2 + \lambda_1\beta||x||_1 & \\ \quad +\lambda_3\frac{1}{\alpha_2^2}||f_{SI} - Dx - \alpha_3 I||_2^2 & MR_K \leqslant MR_{Th}. \end{cases} \tag{16}$$

It can be seen that when $\lambda_2, \lambda_3 = 0$ the problem (16) is directly reduced to the conventional CS reconstruction problem (14). In other words, our proposed reconstruction method incorporates both the sparse prior and the correlation constraint in video sequences. This operation indeed imposes prior knowledge of video signals to CS reconstruction, since sparsity alone is essentially not sufficient for

Table 4
The performance comparison of CIF sequences.

MR_K	MR_{NK}	Foreman			Mobile			News		
		DL-REC	Baseline	Proposed	DL-REC	Baseline	Proposed	DL-REC	Baseline	Proposed
Average PSNR of all frames (dB)										
0.1	0.1	21.94	23.01	23.26	15.05	15.13	15.28	19.43	19.84	19.98
0.2	0.2	25.74	26.99	27.20	17.11	17.37	17.64	23.71	24.87	25.11
0.3	0.3	28.00	29.76	29.99	18.65	19.64	19.76	26.76	28.91	29.10
0.4	0.4	29.71	31.95	32.51	19.86	21.67	22.12	29.26	32.32	32.85
0.5	0.5	31.13	34.00	34.68	20.98	23.62	24.41	33.67	35.16	36.17
Average PSNR of all NK frames (dB)										
0.5	0.1	24.73	30.42	31.16	16.78	21.24	22.72	23.15	30.94	32.92
0.5	0.2	27.28	31.80	32.78	18.25	22.46	23.63	25.66	32.76	34.63
0.5	0.3	27.88	32.72	33.98	18.73	23.27	24.49	26.81	33.94	35.93
0.5	0.4	28.25	33.51	34.91	19.08	24.00	25.37	27.56	35.04	36.87
0.5	0.5	28.47	34.24	35.59	19.35	24.67	26.24	32.88	35.87	37.90

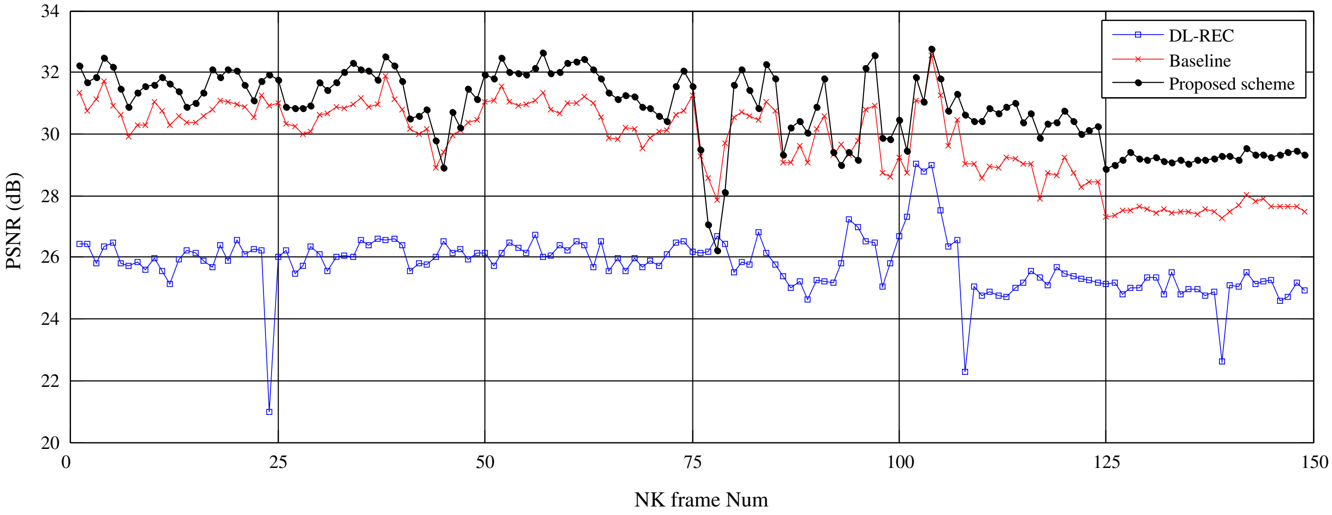

Fig. 4. The performance comparison of each NK frame for Foreman QCIF when $MR_K = 0.5$ and $MR_{NK} = 0.1$.

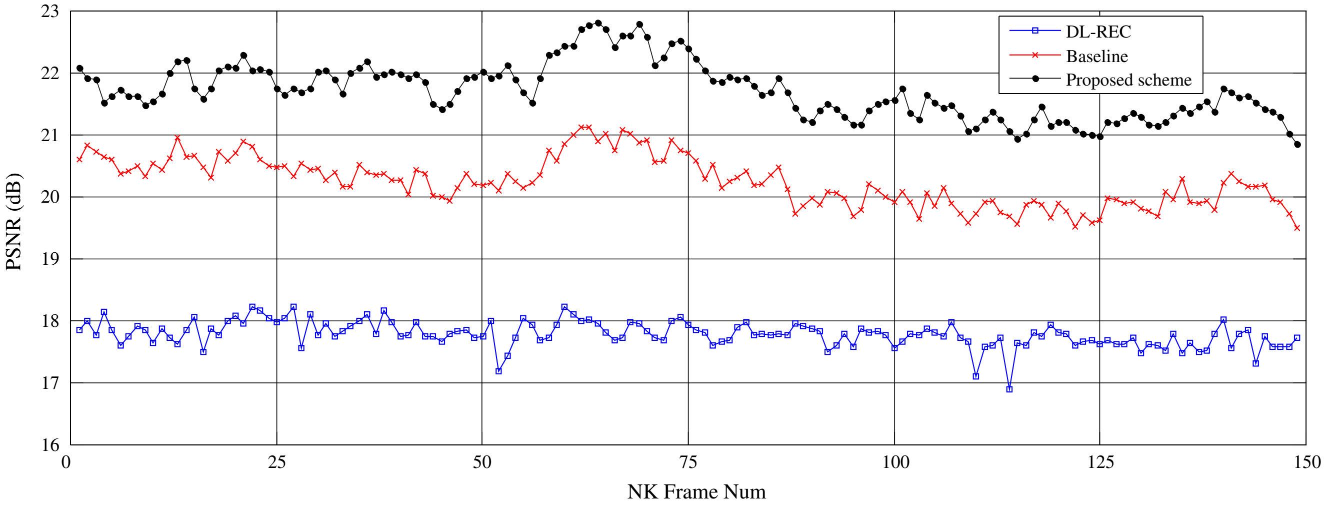

Fig. 5. The performance comparison of each NK frame for Mobile QCIF when $MR_K = 0.5$ and $MR_{NK} = 0.1$.

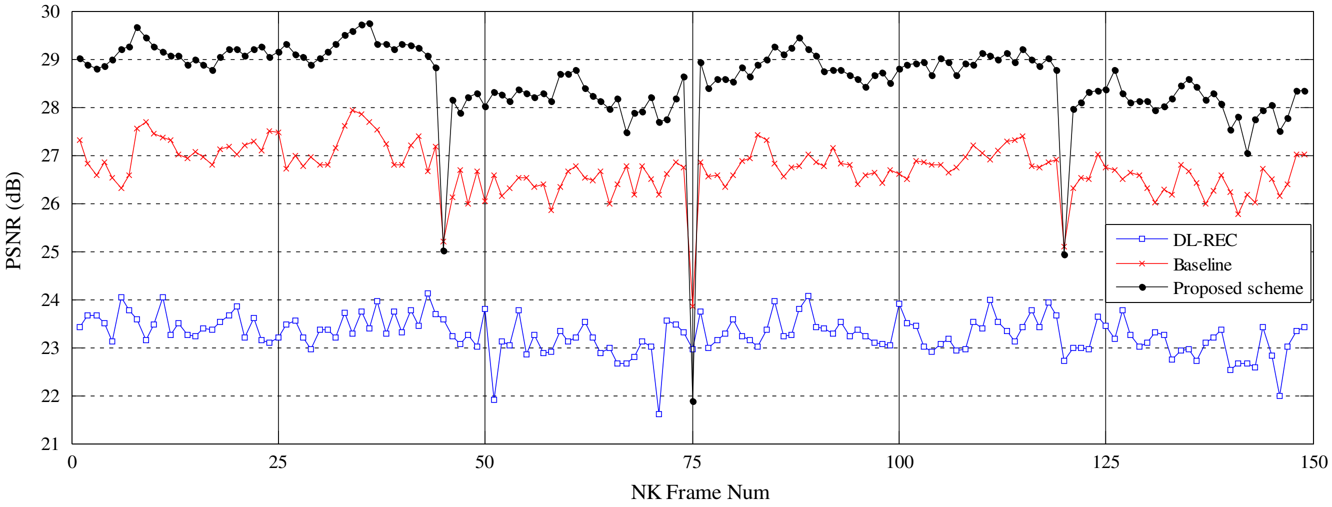

Fig. 6. The performance comparison of each NK frame for News QCIF when $MR_K = 0.5$ and $MR_{NK} = 0.1$.

image/video reconstruction with good visual quality. It is worth noting that our reconstruction is included in the process of dictionary learning. On one hand, the block-based reconstruction is one step of optimizing E, to prepare the coefficients for dictionary update. On the other hand, this operation directly generates the recovered $\hat{f}_{NK} = D\hat{x}_{NK}$ as the output of DCVS, in which $\hat{x}_{NK}$ is the solution of (16).

As soon as the recovered $\hat{x}_{NK}$ is obtained, the dictionary D_{update} is updated at the gradient direction to minimize the energy. Then the optimization is performed by iterating between reconstruction and dictionary update until convergence is achieved. Note that the complexity of the dictionary learning based reconstruction is highly dependent on the specific number of iterations. Compared to the conventional method, the extra cost is only generated from dictionary update in each iteration. However, for many applications of DCVS, an efficient low-complexity encoder is the major concern. The computational complexity needs to be shifted to the joint decoder, which is often assumed without constraints. For example, in WMSN, the video captured by one or several sensors are encoded independently but decoding jointly by a PC server or workstation.

5. Simulation results

In this paper, several video sequences (Y frames for each) with QCIF (176 × 144) and CIF (352 × 288) resolutions are employed to evaluate the proposed ML dictionary learning based reconstruction algorithm. Processing is carried out only on the luminance component. In our simulations, the DCVS structure described in Section 4 is used with the GOP size of 2. At the encoder, each NK frame f_{NK} is split into several non-overlapping 16 × 16 blocks, and all blocks are projected independently using the Gaussian matrix, while key frames f_K are projected using the SBHE matrix [23]. The measurement rate threshold MR_{Th} is set to 0.2 that determines the block-based reconstruction formulation in (16). Specifically, the key frame reconstruction of f_K in (14) is solved via the SpaRSA [24] algorithm (with its default parameter settings), and the block-based NK frame reconstruction (16) is solved using the CVX toolbox [25], wherein the parameters α_i $(i = 0, 2, 3)$ and β are taken based on Tables 1 and 2, and the tradeoff parameters are set from experience as follows, $\lambda_1 = 0.028$, $\lambda_2 = 5.7$ and $\lambda_3 = 1$.

We compare our proposal to two alternatives, (i) the NK frame reconstruction based on K-SVD dictionary learning [7,8] and (ii)

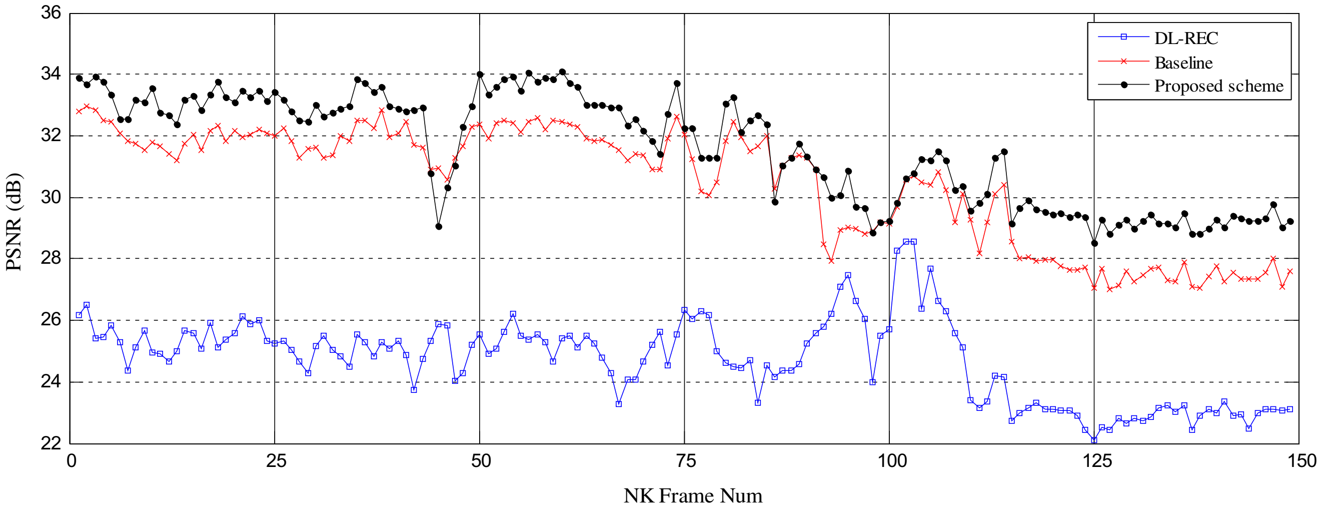

Fig. 7. The performance comparison of each NK frame for Foreman CIF when $MR_K = 0.5$ and $MR_{NK} = 0.1$.

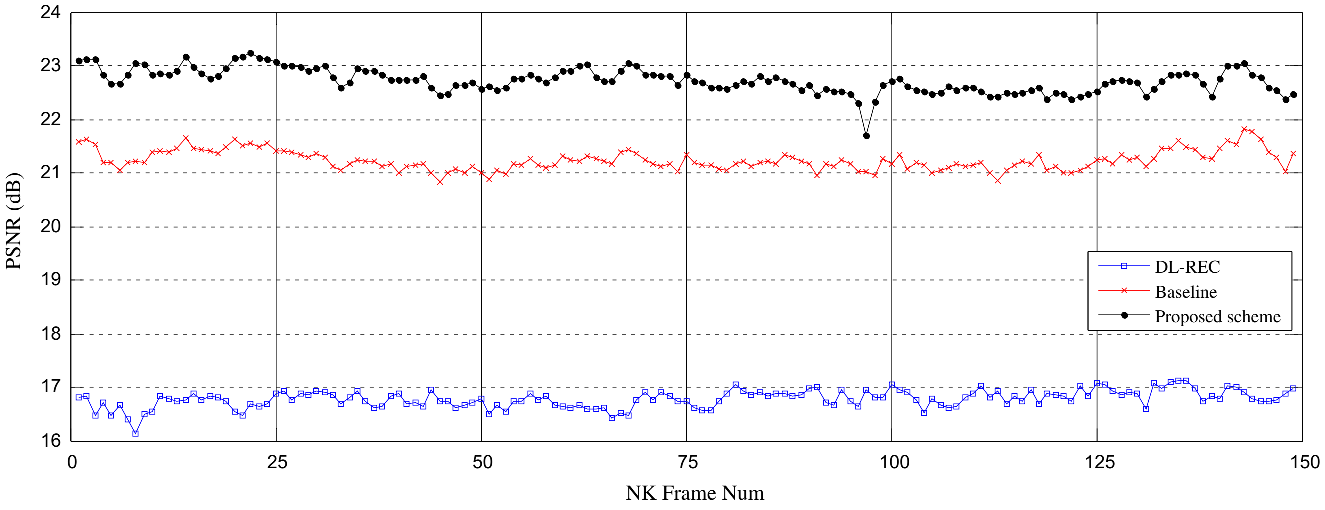

Fig. 8. The performance comparison of each NK frame for Mobile CIF when $MR_K = 0.5$ and $MR_{NK} = 0.1$.

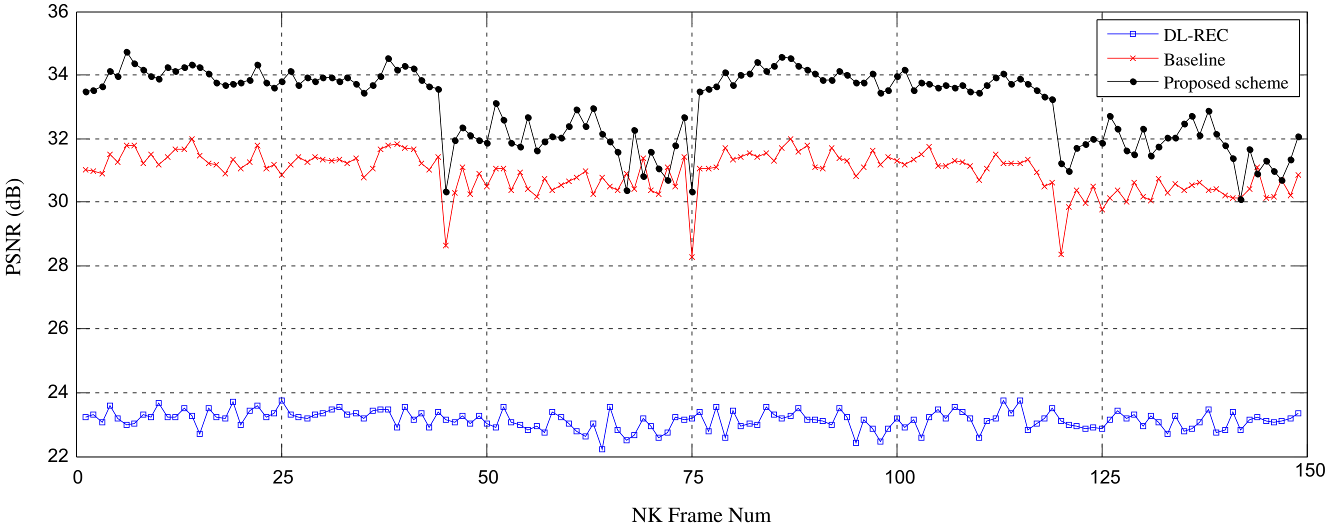

Fig. 9. The performance comparison of each NK frame for News CIF when $MR_K = 0.5$ and $MR_{NK} = 0.1$.

the scheme using our method but without dictionary update (denoted as the baseline scheme). Compared to our algorithm, the dictionary in the baseline scheme is built by directly picking blocks from SI without update step and iterative loop, while the block-based reconstruction is still performed using (16). The reconstruction scheme in [7,8] totally relies on the K-SVD dictionary learning method, in which training patches from two neighboring key frames and SI are used to train the dictionary. Then the frame-based reconstruction is performed using (14). This method needs to first obtain the learned dictionary before recovering frames. Here, we denote this algorithm [7,8] as the dictionary-learning-and-then-reconstruction scheme (DL-REC). For a fair comparison, the same test conditions (the SBHE measurement matrix and DWT sparsifying basis) are used for key frames, while for NK frames the size of all dictionaries/basis is set to 256 × 256 and used the CVX toolbox to perform sparse recovery in all simulations.

As a primary measure of reconstruction quality, we calculate the peak signal-to-noise ratio (PSNR) averaged over all (or only NK) frames under consideration. In our experiment, various *MRs* are employed for key frames as well as NK frames. In most literatures, frames are usually subsampled at the same *MR*, i.e., *MR* of key frames (MR_K) equals that of NK frames (MR_{NK}). However, from the view point of joint sparsity model in distributed compressive sensing (DCS) [26], MR_{NK} can be smaller than MR_K if the correlation between frames is exploited in the process of reconstruction. Consequently, key frames could be subsampled at an increased MR_K with respect to MR_{NK}; see, e.g., [7,9,14,15]. Thus, the same conditions used in [14,15] are employed in our simulations, considering

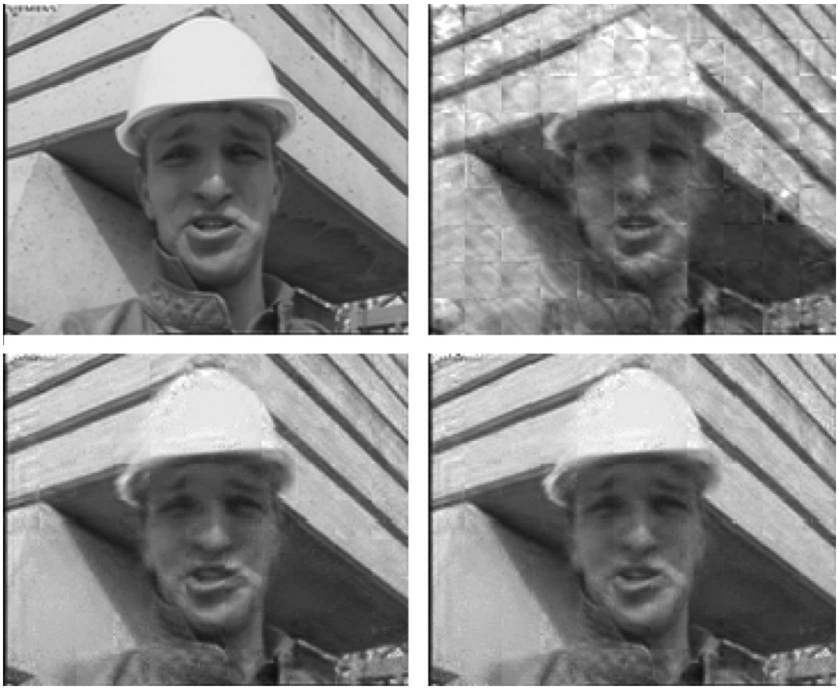

Fig. 10. Recovered 52nd frame of the Foreman QCIF sequence when $MR_K = 0.5$ and $MR_{NK} = 0.5$. (Top left) The original frame. (Top right) The DL-REC scheme. (Bottom left) The baseline scheme. (Bottom right) Our proposed scheme.

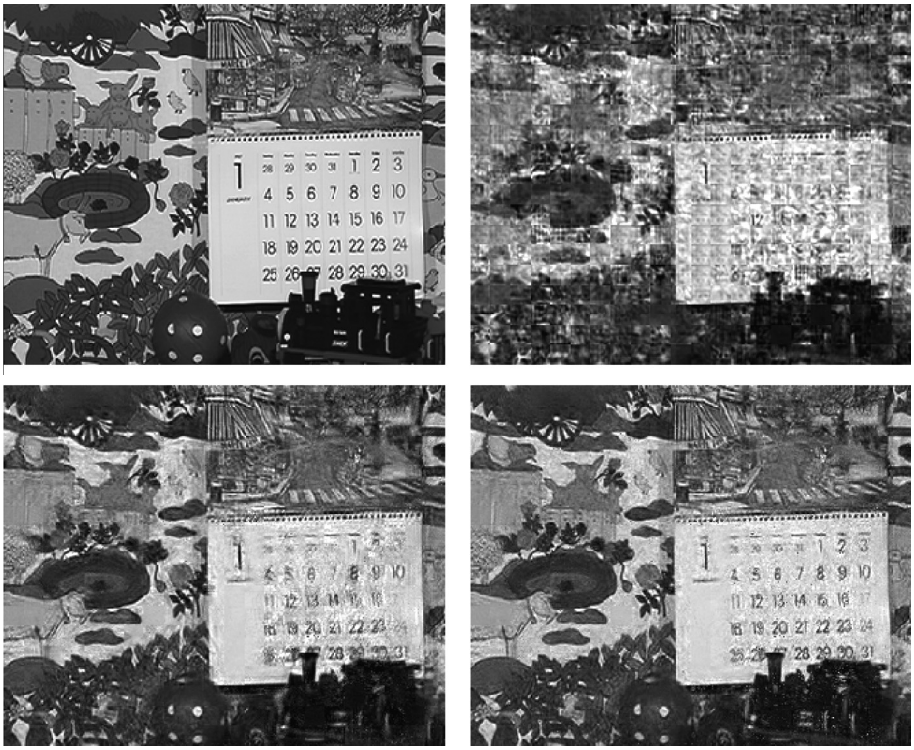

Fig. 11. Recovered 52nd frame of the Mobile CIF sequence when $MR_K = 0.5$ and $MR_{NK} = 0.5$. (Top left) The original frame. (Top right) The DL-REC scheme. (Bottom left) The baseline scheme. (Bottom right) Our proposed scheme.

the following two cases $MR_K = MR_{NK}$ and $MR_K > MR_{NK}$. In the latter situation, PSNR results are averaged over NK frames, since our proposed ML dictionary learning based reconstruction algorithm is employed only for NK frames as shown in Fig. 1. A summary of the results from both cases is presented in Tables 3 and 4, for QCIF and CIF sequences, respectively.

From Table 3, it can be seen that for QCIF sequences, when $MR_K = MR_{NK}$ our proposal increases about 0.04–3.36 dB over the DL-REC scheme in the average PSNR (0.48–1.95 dB for Foreman, 0.04–1.34 dB for Mobile, and 0.18–3.31 dB for News) as *MR* increases from 0.1 to 0.5. When $MR_K > MR_{NK}$, MR_K is set to 0.5 while MR_{NK} varies between 0.1 and 0.5. In this case, we also observe a 2.7–6.63 dB gain in the average PSNR of NK frames (3.93–5 dB for Foreman, 2.6–3.92 dB for Mobile, and 3.33–6.63 dB for News). For CIF sequences, as depicted in Table 4, we observe a 0.23–3.59 dB and 5–9 dB increase over the DL-REC scheme when $MR_K = MR_{NK}$ and $MR_K > MR_{NK}$, respectively. Additionally, compared with the baseline scheme, our proposed scheme using dictionary update offers a PSNR gain (for both QCIF and CIF sequences) of around 0.1–1.01 dB when $MR_K = MR_{NK}$, and 0.7–2.03 dB when $MR_K > MR_{NK}$.

As implied in the simulation results, a significant PSNR gain can be obtained by using the baseline scheme in most cases, since more correlation information between video frames is exploited in the process of reconstruction (16) compared with DL-REC. Furthermore, better performance can be achieved through dictionary update and iterative loop as shown in the tables. It can also be seen that when key frames are subsampled at an increased MR_K with respect to MR_{NK}, the higher reconstruction quality of key frames will result in better PSNR results, since the more accurate dictionary can be initialized from SI. For example, when $MR_K = 0.5$ and $MR_{NK} = 0.1$, the average PSNR of the Foremen QCIF sequence can be increased approximately 5 dB.

As shown in Figs. 4–6, the PSNR results of each NK frame in the Foreman, Mobile and News QCIF sequences with $MR_K = 0.5$ and $MR_{NK} = 0.1$ are demonstrated. It can be seen that almost each NK frame achieves a PSNR gain over the baseline and DL-REC scheme.

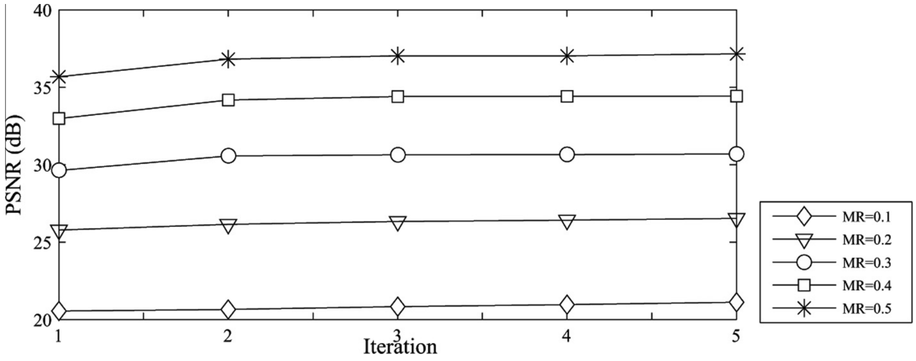

Fig. 12. Average PSNR vs. the number of iterations for Foreman (QCIF).

Note that the PSNR results of some frames undergo a sharp decline, e.g., the 26th and 80th NK frame in Foreman. The major reason is that content changes of these frames in the sequence are somewhat bigger, which leads to the result that the relatively-poor-quality side information cannot afford good atoms to initialize and train an accurate dictionary for NK frame recovery. Meanwhile, the comparisons for each NK frame of CIF sequences are shown in Figs. 7–9. It can be also observed that PSNR of almost each NK frame decoded by using our proposal is higher than that decoded in DL-REC and baseline, with an improvement of 5.95–9.88 dB and 1.11–2.1 dB, respectively on average. Additionally, Fig. 10 shows an example of the recovered 52nd NK frame in the Foreman QCIF sequence by using the three methods when MR_K and MR_{NK} are set to 0.5, and the reconstructed 52nd NK frame in the Mobile CIF sequence is shown in Fig. 11. From the above results, it can be concluded that our proposed scheme provides better performance by using the ML dictionary learning, since the signal structure in addition to the sparsity prior is considered in our probabilistic model.

Finally, a five-time-iteration of our algorithm is illustrated as an example to show its convergence. For simplicity, the first 100 frames of the Foreman QCIF sequence are used in this simulation. Specifically, the GOP is set to 2. The key frames are recovered using DWT with different MR; while the NK frames are tested with the number of iterations varied from one to five. The average PSNR results are shown in Fig. 12. From this figure, one can clearly see that the higher-quality reconstruction can be obtained by increasing the iteration times. Basically, a good reconstruction performance could be efficiently achieved through two–three iterations.

6. Conclusion

In this paper, we propose a dictionary learning based reconstruction algorithm for DCVS. We try to improve the reconstruction performance by leveraging more realistic signal models that go beyond simple sparsity and compressibility by including the video signal structure. In this work, we present a novel undersampled CNM to efficiently describe the correlation structure existed in video, and a maximum likelihood dictionary learning method is proposed, wherein a novel probabilistic model is formulated by including a sparsity prior and the undersampled CNM constraint. In our proposed dictionary learning based reconstruction algorithm, the ML optimization is cast as an energy minimization problem and solved by iterating sparse recovery and dictionary update. In other words, the video reconstruction is performed during the process of dictionary learning, not as an independent task. Therefore, the recovery and dictionary learning can be jointly optimized under the signal structure constraint. Experimental results show that our proposal compares favorably with other existing methods.

To further improve the reconstruction performance with a smaller number of measurements, several important issues need to be investigated for future works: (1) measurement rate allocation: different measurement rate should be adaptively assigned to different frames and regions based on its sparsity degree; (2) fast dictionary learning algorithms; and (3) more efficient algorithms solving the convex optimization problem.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (Nos. 61271173 and 60802032), the Fundamental Research Funds for the Central Universities (No. K5051201045), the 111 Project (No. B08038), and also supported by the ISN State Key Laboratory.

References

[1] B. Girod, A.M. Aaron, S. Rane, D. Rebollo-Monedero, Distributed video coding, Proceedings of the IEEE 93 (1) (2005) 71–83.
[2] E.J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2) (2006) 489–509.
[3] D.L. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306.
[4] E.J. Candes, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics 59 (8) (2006) 1207–1223.
[5] J. Prades-Nebot, Y. Ma, T. Huang, Distributed video coding using compressive sampling, in: Proceedings of the Picture Coding Symposium, 2009, pp. 1–4.
[6] T.T. Do, Yi Chen, D.T. Nguyen, N. Nguyen, Lu Gan, T.D. Tran, Distributed compressed video sensing, in: IEEE International Conference on Image Process, 2009, pp. 1393–1396.
[7] Hung-Wei Chen, Li-Wei Kang, Chun-Shien Lu, Dynamic measurement rate allocation for distributed compressive video sensing, in: SPIE Visual Communications and Image Processing, 2010, pp. 774401–774410.
[8] Hung-Wei Chen, Li-Wei Kang, Chun-Shien Lu, Dictionary learning-based distributed compressive video sensing, in: Picture Coding Symposium, 2010, pp. 210–213.
[9] W. Xu, Z. He, K. Niu, J. Lin, Sub-sampling framework of distributed video coding, in: IEEE International Symposium on Circuits and Systems, 2010, pp. 1145–1148.
[10] Li-Wei Kang, Chun-Shien Lu, Distributed compressive video sensing, in: IEEE International Conference on Acoustics, Speech, and Signal Process, 2009, pp. 1169–1172.
[11] X. Hao, B. Zhuang, A. Cai, Measurement compression in distributed compressive video sensing, in: IEEE International Conference on Broadband Network & Multimedia Technology, 2010, pp. 850–854.
[12] M. Aharon, M. Elad, A.M. Bruckstein, The K-SVD: an algorithm for desigining of overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.
[13] J. Zheng, E.L. Jacobs, Video compressive sensing using spatial domain sparsity, Optical Engineering 48 (8) (2009) 087006-1–087006-10.
[14] S. Mun, J.E. Fowler, Residual reconstruction for block-based compressed sensing of video, in: Data Compression Conference, 2011, pp. 183–192.
[15] Haixiao Liu, Bin Song, Hao Qin, Zhiliang Qiu, A dictionary generation scheme for block-based compressed video sensing, in: IEEE International Conference on Signal Processing, Communication and Computing, 2011, pp. 670–674.
[16] Z. Liu, A. Elezzabi, H. Zhao, Maximum frame rate video acquisition using adaptive compressed sensing, IEEE Transactions on Circuits and Systems for Video Technology 21 (11) (2011) 1704–1718.
[17] Haixiao Liu, Bin Song, Hao Qin, Zhiliang Qiu, An adaptive-ADMM algorithm with support and signal value detection for compressed sensing, IEEE Signal Processing Letters 20 (4) (2013) 315–318.
[18] I. Tošić, P. Frossard, Dictionary learning for stereo image representation, IEEE Transactions on Image Processing 20 (4) (2011) 921–934.
[19] A. Papoulis, S.U. Pillai, Probability, Random Variables and Stochastic Process, fourth ed., Mcgraw-Hill College, New York, 2002.
[20] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1?, Vision Research 37 (23) (1997) 3311–3325
[21] I. Tošić, P. Frossard, Dictionary learning, IEEE Signal Processing Magazine 28 (2) (2011) 27–38.
[22] C. Brites, F. Pereira, Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding, IEEE Transactions on Circuits and Systems for Video Technology 18 (9) (2008) 1177–1190.
[23] L. Gan, T.T. Do, T.D. Tran, Fast compressive imaging using scrambled hadamard ensemble, in: Proceedings of the European Signal Processing Conference, 2008.
[24] S.J. Wright, R.D. Nowak, M.A.T. Figueiredo, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing 57 (7) (2009) 2479–2493.
[25] M. Grant, S.P. Boyd, CVX: Matlab software for disciplined convex programming. <http://cvxr.com/cvx/>.
[26] M.F. Duarte, S. Sarvotham, D. Baron, M.B. Wakin, R.G. Baranuik, Distributed compressed sensing of jointly sparse signals, in: Asilomar Conference on Signals, Systems and Computers, 2005, pp. 1537–1541.