Neurocomputing 174 (2016) 404–412

Contents lists available at ScienceDirect

Neurocomputing

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

Hierarchical frame based spatial–temporal recovery for video compressive sensing coding

Xinwei Gao *, Feng Jiang, Shaohui Liu, Wenbin Che, Xiaopeng Fan, Debin Zhao

Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 31 January 2015
Received in revised form
18 June 2015
Accepted 14 July 2015
Available online 11 September 2015

Keywords:
Video compressed sensing
Hierarchical structure framework
Spatial–temporal sparse representation

ABSTRACT

In this paper, the divide-and-conquer based hierarchical video compressive sensing (CS) coding framework is proposed, in which the whole video is independently divided into non-overlapped blocks of the hierarchical frames. The proposed framework outperforms the traditional framework through the better exploitation of frames correlation with reference frames, the unequal sample subrates setting among frames in different layers and the reduction of the error propagation. At the encoder, compared with the video/frame based CS, the proposed hierarchical block based CS matrix can be easily implemented and stored in hardware. Each measurement of the block in a different hierarchical frame is obtained with the different sample subrate. At the decoder, by considering the spatial and temporal correlations of the video sequence, a spatial–temporal sparse representation based recovery is proposed, in which the similar blocks in the current frame and these recovered reference frames are organized as a spatial–temporal group unit to be represented sparsely. Finally, the recovery problem of video compressive sensing coding can be solved by adopting the split Bregman iteration. Experimental results show that the proposed method achieves better performance against many state-of-the-art still-image CS and video CS recovery algorithms.

1. Introduction

In recent years, Compressive Sensing (CS) has been extensively studied, whose purpose is to reconstruct the signal from its observed measurements

$$y_v = \Phi_v v, \tag{1}$$

where $v \in R^{kN}$ is lexicographically stacked representations of the original video sequence (k is the frame number of the video and N is the pixel number of each frame) and $y_v \in R^{kM}$ is the CS measurements observed by a random $kM \times kN$ measurement matrix Φ_v, ($M \ll N$). The sample subrate $r = M/N$. It is noticed that, the size of the video measurement matrix is too big to be implemented and stored in hardware. In order to relieve the problem, by the idea of divide-and-conquer, the video is divided into many frames and the measurement of each frame f_i is linearly projected by a frame based measurement matrix Φ_f

$$y_{f_i} = \Phi_f f_i, \tag{2}$$

where $f_i \in R^N$ is lexicographically stacked representations of the ith frame and $y_{f_i} \in R^M$ is the CS measurements observed by a random $M \times N$ measurement matrix Φ_f. However, with the standard-definition video or high-definition (HD) video, the implementation and storage problems still exist. The size of a random measurement matrix for a block is much smaller than that for the whole frame, and the problem of large storage cost of the whole image measurement matrix is avoided by employing the block-based random measurement matrix instead. For this season, block-based CS [1] is proposed, in which each frame is divided into many non-overlapped blocks, each block is linearly projected by the same random measurement matrix.

$$y_{b_i} = \Phi_b b_i. \tag{3}$$

The block-based measurement matrix design can be seen as a special case of Eq. (1), if the whole matrix can be written as a block diagonal with the block matrix along the diagonal [2].

By the hardware implementation, the process in video CS has been made with a single-pixel camera [3], based on representing a video in the Fourier domain or the Wavelet domain. And then, more complicated cameras [4,5] are proposed by considering the correlation in the spatial or temporal domain. It can be seen that the coherence between the Gaussian random matrix and the recovery dictionary is low making the recovery of video compressive sensing

* Corresponding author.
E-mail addresses: xwgao.cs@hit.edu.cn (X. Gao), fjiang@hit.edu.cn (F. Jiang), shliu@hit.edu.cn (S. Liu), chewenbin@hit.edu.cn (W. Che), fxp@hit.edu.cn (X. Fan), dbzhao@hit.edu.cn (D. Zhao).

http://dx.doi.org/10.1016/j.neucom.2015.07.110
0925-2312/

effective. So the Gaussian random matrix is widely used on the hardware to generate linear measurements in compressive sensing of the signal.

Although video CS measurement process can be regarded as a combination of acquisition and compression, this process is not a real video compression in the strict information theoretic sense, because it cannot directly produce a bitstream from the sensing device hardware, which can be only seen as a technology of dimensionality reduction in essence [2]. As a very important technology of image/video coding, quantization is introduced into the CS image/video coding model [6], which is applied for the CS measurements of each frame. However, due to the random characteristic of generated frame measurements by the random matrix Φ, isometric scalar quantization does not perform well in rate-distortion performance. Inspired by the success of the block-based hybrid video coding, such as HEVC [7] and H.264, the inter-prediction coding technology can be used in the CS measurement process of each video frame. Some works [2,8,9] on the block-based CS (BCS) hybrid coding framework are presented. Mun and Fowler proposed the block-based quantized compressed sensing with differential pulse-code modulation (DPCM) [2] and uniform scalar quantization. In [2], the previous decoded measurement is taken as the candidate of the current measurement. Zhang et al. extended the DPCM based CS measurement coding and proposed the spatially directional predictive coding (SDPC) [8], in which the intrinsic spatial correlation between neighboring measurements of natural images is further explored. In the BCS [1] measurement coding, Khanh et al. [9] point out that, the spatial correlation among neighboring blocks becomes higher as block size decreases and the CS recovery of a small block is less efficient than that of a large block. In order to balance the conflict between compressed ratio and reconstructed quality, a structural measurement matrix (SMM) is proposed [9] to achieve a better RD performance, in which the image is sampled by some small blocks, and reconstructed with large blocks spliced by the small block.

Since each CS measurement of the frame can be coded, the traditional framework of the whole video CS coding in implementation is that, each frame of the video sequence is recovered independently with the same sample subrate. Mun and Fowler [10] proposed a video CS framework with key frames and non-key frames, which utilized the temporal redundancy in video sequence to improve the recovery quality of the non-key frames. Two sample subrates are used in sampling stage, where the high sample subrate is adopted for the key frames and the low sample subrate for the non-key frames. By considering the factors of the hardware implementation, the spatial–temporal correlation and the reconstructed quality, the hierarchical video compressive sensing (CS) coding framework is proposed in this paper, in which the whole video is independently divided into non-overlapped blocks of the hierarchical frames. The proposed framework outperforms the traditional framework through the better exploitation of frames correlation with reference frames, the unequal sample subrates setting among frames in different layers and the reduction of the error propagation. At the encoder, compared with the video/frame based CS, the proposed hierarchical block based CS matrix can be easily implemented and stored in hardware. Each measurement of the block in different hierarchical frame is obtained with the different sample subrate. Finally, these measurements are coded into bitsteams by the prediction, the quantization and the entropy coding.

At the decoder, the measurements of the frames are decoded by the inverse process of the prediction, the quantization and the entropy coding. From many fewer acquired measurements than suggested by the Nyquist sampling theory, the CS theory demonstrates that a signal x can be reconstructed with high probability when it exhibits sparsity in some domain Ψ, which has greatly changed the way engineers think of data acquisition,

$$x = \Psi\theta. \tag{4}$$

If θ is a sparse coefficient vector, the signal x is sparse under the domain Ψ. The performance will be poor, using the still-image CS recovery algorithms to the video CS measurement. By considering the spatial and temporal correlations of the video, it is possible to achieve a high-quality recovered video even employing a low sample subrate. A motion compensation based residual recovery was proposed [10], which utilized the temporal redundancy in video sequence. Two subtrates are used in a sampling stage, where high subrate is adopted for key frames and low subrate for non-key frames. Then, not only the temporal redundancy, but also the multi-images redundancy and the multiview redundancy are taken into account in [11]. Mun et al. [12] cast the CS reconstruction in the base of contourlet transform or complex-valued dual-tree wavelet transform, resulting in better performance compared to the conventional fixed domain based recovery methods. However, it is almost impossible to find a universal domain in which all kinds of signals are sparse. As an alternative to the CS reconstruction scheme, the iterative algorithms based on non-local patches have been proposed recently (e.g. [13,14]). In [13], the number of nonzeros 3-D transformation coefficients of a group, which is stacked by the non-local patches, was used to measure the non-local sparsity. Additionally, the collaborative sparsity measure was established in [13], enforcing local smoothness and non-local sparsity simultaneously. A group sparse representation (GSR) modeling was further developed in [14], using the non-local grouping technique as well. In essence, this modeling efficiently utilized the intrinsic self-similarity in the spatial domain of natural images, which also exhibits the patch similarity among patch group. Also, GSR modeling improves the performance of recovery over conventional fixed domain based recovery methods.

Inspired by the idea of GSR, at the decoder of the proposed framework, by considering the spatial and temporal correlations of the video sequence, a spatial–temporal sparse representation based recovery is proposed to improve the recovered quality, in which the similar blocks in both the current frame and these recovered reference frames are grouped as a spatial–temporal group unit to be sparse represented. These reference frames are selected by the optimal decision of the hierarchical based framework. At the decoder, by considering the spatial and temporal correlations of the video sequence, a spatial–temporal sparse representation based recovery is proposed, in which the similar blocks in the current frame and these recovered reference frames are organized as a spatial–temporal group unit to be represented sparsely. Experimental results show that the proposed method achieves better performance against many state-of-the-art still-image CS and video CS recovery algorithms.

2. Proposed hierarchical frame based video CS coding framework

2.1. The key frame based video CS framework

The straightforward consideration of the video CS measurement sampling is to design a video-based measurement matrix. However, the size of the video measurement matrix is too big to be implemented and stored in hardware. With the standard-definition video or the high-definition (HD) video, the implementation and storage problems of frame-based measurement matrix still exist. Then, the block-based CS sampling [1] is proposed, in which each frame of the video is divided into the non-overlapped blocks, and each block is independently and linearly projected by the

same block-based measurement matrix. Because the size of a block-based measurement matrix for a block is much smaller than the frame-based and video-based matrices, the problem of large storage cost of the whole image measurement matrix is avoided by employing the block-based random measurement matrix. Like the technology of distributed video coding (DVC), the block-based video CS framework has a more lightweight encoder and pushes the computational complexity to the decoder, which is suitable for the video sensing applications with the constrain of low-complexity and low-power. The traditional framework in implementation is that, each frame of the video sequence is recovered independently with the same sample subrate. Mun and Fowler [10] proposed a video CS framework with key frames and non-key frames, which utilized the temporal redundancy in video sequence to improve the recovery quality of the non-key frames. Two sample subrates are used in sampling stage, where the high sample subrate is adopted for the key frames and the low sample subrate for the non-key frames.

2.2. The proposed hierarchical frame based video CS framework

In this paper, we propose a divide-and-conquer based hierarchical video compressive sensing (CS) coding framework, which can be shown in Fig. 1. In the proposed framework, frames are divided into many groups of pictures (GOP), in which the 2^{k-1} frames are defined by k layers $l_1, l_2, \ldots, l_k$ with different sample subrates for different layers. At the sampling stage, the video is defined as $v = \{f_1, f_2, \ldots, f_n\}$, the constrain of the sample subrates r is

$$r_{l_i} > r_{l_j}, \quad \text{s.t. } i < j, \tag{5}$$

r_{l_i} represents the sample subrate of the qth frame f_q on the ith layer ($q \in l_i$). The block-based measurement matrix should also be adjusted by which layer the current frame belongs to.

Compared with the temporal-independent framework and the key-frame based framework, our proposed hierarchical frame based video compressive sensing (CS) coding framework has some advantages.

First, better temporal correlation with reference frames is exploited in the proposed framework, in which the nearest forward and backward recovered frames in the higher layer are considered as the reference frames to improve the recovery quality of the current frame. However, only the forward reference frames are taken part in the temporal-independent framework and the key-frame based framework. In the proposed framework, the unequal sample subrates are used in the hierarchical frames in different layers. It can be noticed in Fig. 2 that, performance of the frame recovery with one nearby reference frame is better than that without reference by the same sample subrate. In addition, the nearby reference with better quality with higher sample subrate can help to recover a better current frame. Through the observation, We select the forward and backward recovered frames with higher sample subrates as reference frames at the recovery stage of the current frame. But in the temporal-independent framework, the sample subrate of the reference frame is equal to that of the current frame.

Second, the selection strategy of the recovered reference frames is proposed in this framework. Another observation points out that, the temporal distance between the current frame and the reference frame also influences the recovery performance (shown in Fig. 3). With the same sample subrate, the larger temporal distance will produce the worse performance of the current frame. When the distance is big enough ($d > 7$), the temporal correlation becomes weak and the performance is worse than that of temporal-independent framework. So, in our proposed framework the nearest recovered frames in the higher layer are selected as the reference of the current frame. Moreover, better temporal correlation with reference frames is exploited in the proposed framework, in which the forward and backward recovered frames in the higher layer are both considered as the reference frames to improve the recovery quality of the current frame. However, only the forward reference frames take part in the temporal-independent framework and the key-frame based framework.

Third, our proposed hierarchical structure will efficiently employ the correlation among frames and reduce the recovery error propagation. Suppose the GOP size is 2^{k-1}. Using the temporal-

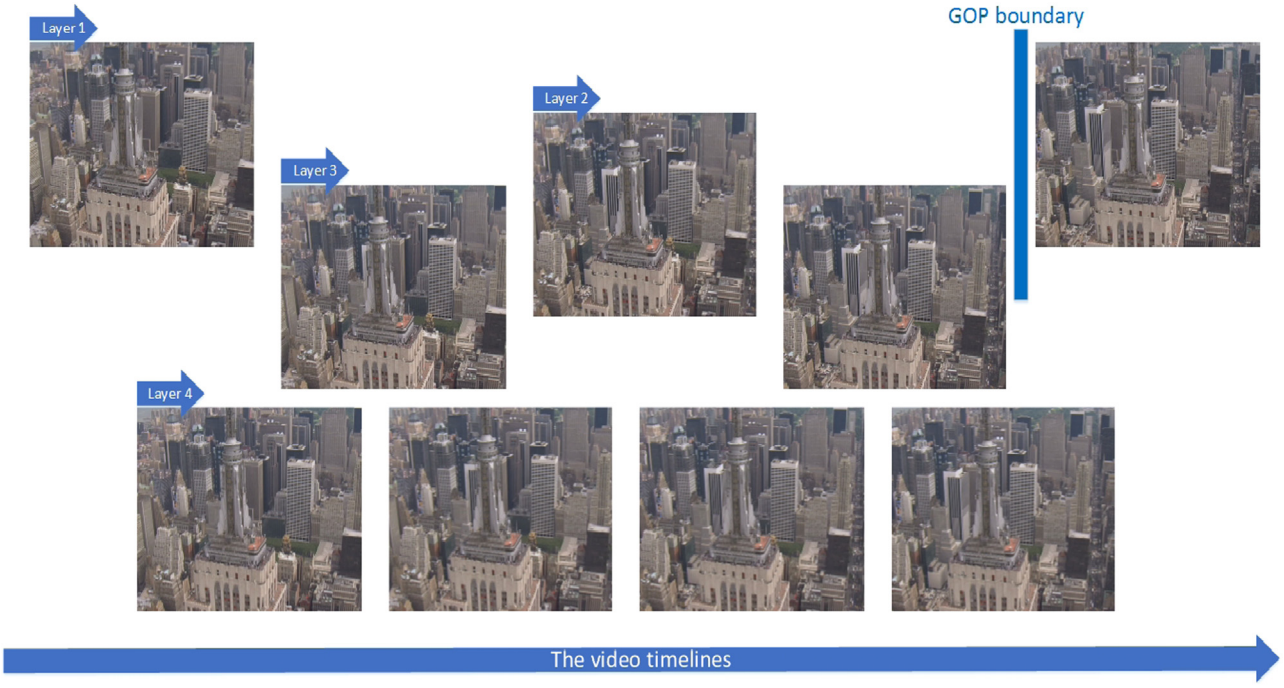


Fig. 1. Hierarchical structure based video CS framework with the $GOP = 2^{k-1}$ (k=4 for example).

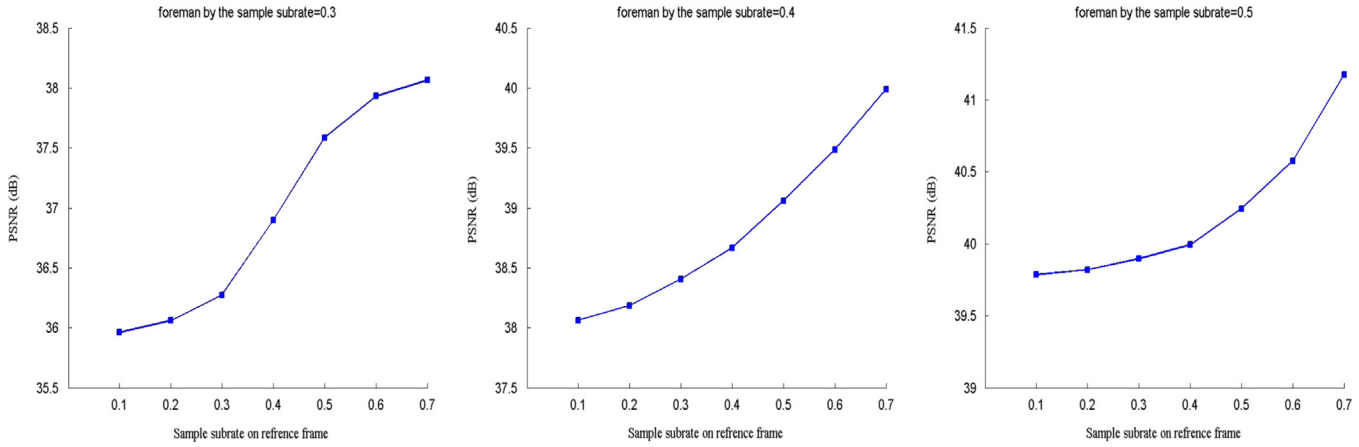


Fig. 2. The PSNR of recovered the 2nd frame (foreman) by using the 1st frame as a reference with different sample subrate (0.1–0.7). The subrate of current recovered frame is 0.3, 0.4 and 0.5 respectively.

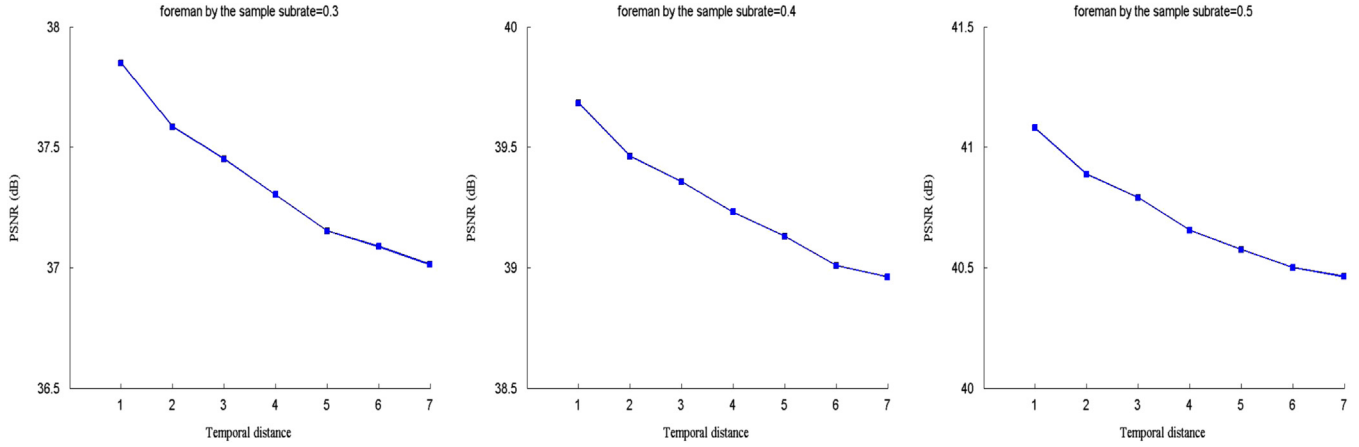


Fig. 3. The PSNR of recovered the 8th frame (foreman) by using the original reference frame with different temporal distance (1st–7th). The subrate of current recovered frame is 0.3, 0.4 and 0.5 respectively.

independent framework and the key-frame based framework, the error will both propagate with $2^{k-1}-1$ times. In our proposed framework, it will be reduced to $k-1$ times with k hierarchical layers.

In conclusion, the proposed hierarchical frame based video compressive sensing (CS) coding framework is more effective than the temporal-independent framework and the key-frame based framework. Next section, the details of the proposed method will be introduced.

3. Spatial–temporal video recovery with hierarchical frame structure

3.1. The DPCM based CS measurement coding

Although the video CS measurement process can be regarded as a combination of the acquisition and the compression, this process is not a real video compression in the strict information theoretic sense, because it cannot directly produce a bitstream from the sensing device hardware, and it can be only seen as a technology of dimensionality reduction in essence. In the proposed hierarchical frame based video CS coding framework, the hierarchical frames of the video are divided into non-overlapped blocks and these blocks are linear projected by the Gaussian random matrix (GRM) with the corresponding sample subrates. At the CS measurement y_v encoder, each current measurement of a block is coded by the prediction measurement. In DPCM based CS measurement coding [2], the prediction measurement is the decoded measurement of the previous block; In SDPC [8], the optimal prediction measurement is selected from a set of candidates that are generated by four designed directional predictive modes. Then, the prediction residuals are uniform scalar quantized and entropy encoded into bitstreams. The decoder process is the inverse of encoder, in which by using the DPCM-based CS coding (shown in Fig. 4) the reconstructed video CS measurement $\tilde{y}_v$.

3.2. The spatial–temporal sparse representation modeling

At the decoder, by the de-quantization on quantizer indexes from the bitstream the quantized residuals $\tilde{r}_b$ can be obtained, which is then added by the prediction $\hat{y}_b$, producing the reconstructed CS measurements group $\tilde{y}_b = \tilde{r}_b + \hat{y}_b$, ready for further prediction coding. At last, all the reconstructed measurements $\tilde{y}_b$ are obtained to compose the frame measurement $\tilde{y}_f = \{\tilde{y}_{b_1}, \tilde{y}_{b_2}, \ldots, \tilde{y}_{b_m}\}$ sequentially, which are then utilized for the current frame reconstruction by CS recovery algorithms. Compressive sensing theory allows that a current frame f_i can be exactly recovered from its space measurements y_{f_i} acquired by linear projection with the sampling subrate r_i if f_i has advantage of being sparse in a domain, e.g. FFT domain [15], DCT domain, DWT domain, or some incoherent domains. Different from other signals, natural image as a two-dimensional signal has its own prior

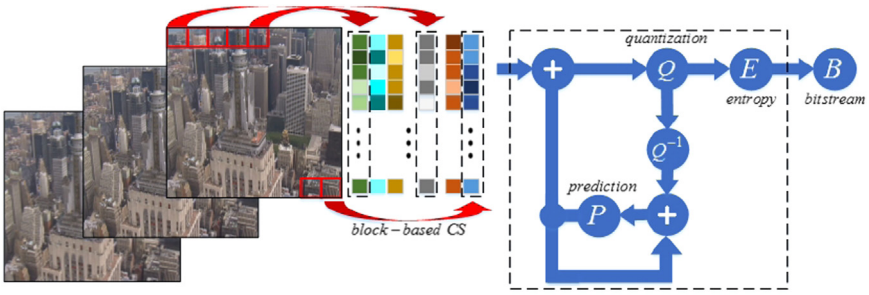


Fig. 4. The diagrammatic of the block-based DPCM video CS coding.

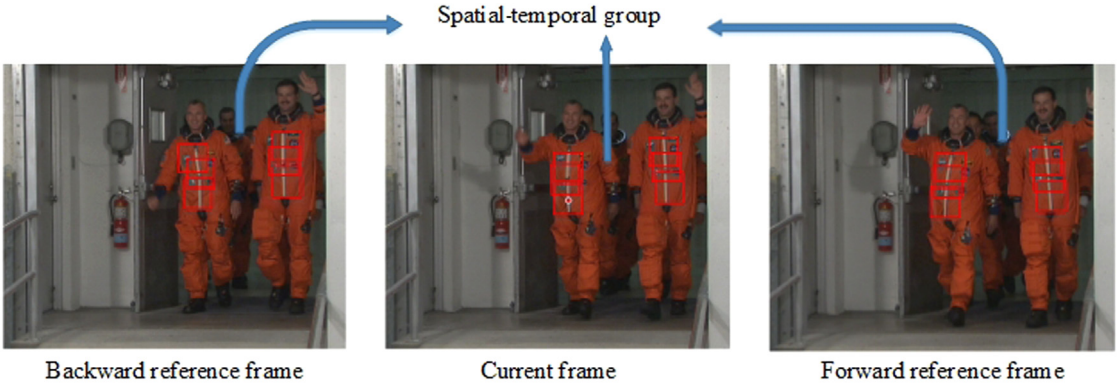


Fig. 5. The diagrammatic of the $G_{i,j}$ composition.

knowledge, such as local smoothing model and group sparse model. To cope with the ill-posed problem of single-frame CS recovery, these traditional methods employ various image prior knowledge for regularizing the solution to the following minimization problem:

$$\arg\min_f \frac{1}{2}\|y_f - \Phi f\|_2^2 + \lambda\Gamma(f), \tag{6}$$

where y_f is the observed frame measurement value at the encoder, $\frac{1}{2}\|y_f - \Phi f\|_2^2$ is the l_2-norm data-fidelity term, $\Gamma(f)$ is called the regularization term denoting image prior and λ is the regularization parameter. Due to that image prior knowledge plays an important role in the performance of the uncompressed single-frame CS restoration algorithms, designing effective regularization terms $\Gamma(f)$ to reflect the image priors is at the core of image restoration. Some image prior models are usually used as the regularization term, such as the local smoothing based total variation (TV) model [16,17], low-rank matrix model [18], nonlocal self-similarity [19] based model and these dictionary based models: DCT model, DWT model, KSVD [20] model and GSR [21] model. However, different from the single-frame, the video sequence not only has the characteristics in the spatial domain but also has the correlation in the temporal domain, which can help to get the better recovered performance of the frame with references in a video than the single-frame without reference frames. The motion compensation based recovery method for video CS measurement was proposed [10], in which the current frame is estimated by the reference of nearest recovered frame. This recovery performance is determined by accuracy of the recovered reference frame and the estimated errors are propagated with the decoding processing. Recently, the group-based sparse representation [21] for image CS recovery is an effective and widely used model, which outperforms the traditional patch-based sparse representation method [20] by considering the correlations between patches in the sparse representation and reducing the computational complexity in the dictionary learning. However, only the self-similarity in spatial domain is utilized. Unlike [10,20,21], in our proposed spatial–temporal sparse representation based video CS recovery method, all the similar blocks in the current frame, the forward reference frame and backward reference frame are exploited. To rectify the above problems of the sparse representation in video CS recovery, we propose a spatial–temporal sparse representation modeling in the unit of spatial–temporal group instead of patch, aiming to exploit the nonlocal similarity in both the spatial domain and the temporal domain of the video sequence simultaneously in a unified framework. In the proposed spatial–temporal group design, the current frame f_i is divided into some overlapped blocks $b_{i,j}$ of size $L_b \times L_b$, where j is the index of block in the ith frame. for each block $b_{i,j}$, its n_b most similar blocks in the current frame f_i, and the corresponding forward and backward reference frames: f_{for_i}, f_{back_i}, which comprise the set $S_{b_{i,j}}$.

$$S_{b_{i,j}} = \{b_k \mid \|b_k - b_{i,j}\|_2^2 < T\}. \tag{7}$$

Note that, the corresponding forward and backward reference frames: f_{for_i} and f_{back_i} are the nearest frames of the current frame f_i with higher layers, each block in $S_{b_{i,j}}$ is represented as a vector, while each spatial–temporal group $G_{i,j}$. The composition of $G_{i,j}$ is shown in Fig. 5. Each spatial–temporal group $G_{i,j}$ is represented by the form of matrix, which is composed of nonlocal blocks in the spatial–temporal domain with similar contents. So in the proposed method, the $\Gamma(f_i)$ is defined by

$$\Gamma(f_i) = \sum_j \|\alpha_{G_{i,j}}\|_0 \quad \text{s.t. } G_{i,j} = D_{i,j}\alpha_{G_{i,j}}, \tag{8}$$

where $D_{i,j}$ is the dictionary of the spatial–temporal group $G_{i,j}$ and $\alpha_{G_{i,j}}$ is the sparse coefficients. The minimization problem of video CS recovery for our proposed spatial–temporal group based sparse representation can be formalized as

$$\arg\min_{f_i} \frac{1}{2}\|y_{f_i} - \Phi f_i\|_2^2 + \sum_j \|\alpha_{G_{i,j}}\|_0 \quad \text{s.t. } G_{i,j} = D_{i,j}\alpha_{G_{i,j}}, \tag{9}$$

where the $D_{i,j}$ is a self-adaptive dictionary for $G_{i,j}$, which is trained directly from the estimate $\tilde{G}_{i,j}$ of $G_{i,j}$. We exploited the singular value decomposition (SVD) to $\tilde{G}_{i,j}$:

$$\tilde{G}_{i,j} = U_{i,j} A_{i,j} V_{i,j}^T, \tag{10}$$

where $u_{i,j\otimes k}$ and $v_{i,j\otimes k}$ are the kth columns of $U_{i,j}$ and $V_{i,j}$ respectively. The dictionary $D_{i,j}$ can be represented by $u_{i,j\otimes k}$ and $v_{i,j\otimes k}$, that is $D_{i,j} = \{u_{i,j\otimes k} v_{i,j\otimes k}^T\}_{k=1}^{n_k}$. Therefore, the ultimate adaptively dictionary for $G_{i,j}$ is learned. Moreover, the self-adaptive dictionary learning for each group with low complexity is more efficient and effective than the dictionary learning from the off-line video sequences.

3.3. The proposed algorithm

In the case of video CS restoration with compressed measurements, the true measurements y_v do not exist in the decoder, and the instead l_2 data-fidelity term $\frac{1}{2}\|\tilde{y}_{f_i} - \Phi f_i\|_2^2$ cannot accurately reflect the accuracy of the actual measurements value. Because the decoded measurements value $\tilde{y}$ reconstructed at the decoder by the sum of the predicted measurements and the quantized residual is not equal to the observed measurement value y at the encoder. To deal with this problem, the $\Gamma(f)$ regularization term is set a bigger regularization parameter λ in the problem with inaccurate l_2-norm data-fidelity term $\frac{1}{2}\|\tilde{y}_{f_i} - \Phi f_i\|_2^2$ than that with the accurate l_2-norm data-fidelity term $\frac{1}{2}\|y_{f_i} - \Phi f_i\|_2^2$.

Finally, we formulate our problem as the following minimization problem:

$$\arg\min_{f_i} \frac{1}{2}\|\tilde{y}_{f_i} - \Phi f_i\|_2^2 + \lambda \sum_j \|\alpha_{G_{i,j}}\|_0, \quad \text{s.t. } G_{i,j} = D_{i,j}\alpha_{G_{i,j}}. \tag{11}$$

Then the general solution to this problem is given by adopting the framework of split Bregman iteration [22] (SBI). This minimization problem can be translated into an equivalent constrained optimization problem by introducing variables u and v:

$$\arg\min_{u,v} = \frac{1}{2}\|\tilde{y}_{f_i} - \Phi u\|_2^2 + \lambda \sum_j \|v_j\|_0, \quad \text{s.t. } u = \{D_{i,j} v_j\}_{j=1}^{n_j}. \tag{12}$$

Algorithm 1. Generalized solution for Eq. (12) by split Bregman iteration.

Input:

Set μ, initialize $u^{(0)}$, $v^{(0)}$, $D^{(0)}$ and $z^{(0)}$, $t=0$;

1: **while** (stopping criterion is not satisfied $t < t_{max}$)

2: $u^{(t+1)} = \arg\min_u \frac{1}{2}\|\tilde{y}_{f_i} - \Phi u\|_2^2 + \frac{\mu}{2}\|u - D^{(t)} v^{(t)} - z^{(t)}\|_2^2$;

3: $v^{(t+1)}, D^{(t+1)} = \arg\min_{v,D} \lambda \sum_j \|v_j\|_0 + \frac{\mu}{2}\|u^{(t+1)} - Dv - z^{(t)}\|_2^2$;

4: $z^{(t+1)} = z^{(t)} - \left(u^{(t+1)} - D^{(t+1)} v^{(t+1)}\right)$;

5: $t = t+1$;

6: **end while**

So in this case the SBI addresses the minimization problem Eq. (12) into u sub-problem and v sub-problem as shown in Algorithm 1. Given $z^{(t)}$, $v^{(t)}$ and $D^{(t)}$, the $u^{(t+1)}$ sub-problem is essentially a minimization problem of strictly convex quadratic function, that is

$$u^{(t+1)} = \arg\min_u \frac{1}{2}\|\tilde{y}_{f_i} - \Phi u\|_2^2 + \frac{\mu}{2}\|u - v^{(t)} - z^{(t)}\|_2^2. \tag{13}$$

The steepest gradient descent method is utilized to solve Eq. (13):

$$\tilde{u}^{(t+1)} = u^{(t)} - \gamma \frac{\partial \frac{1}{2}\|\tilde{y}_{f_i} - \Phi u\|_2^2 + \frac{\mu}{2}\|u^{(t)} - v^{(t)} - z^{(t)}\|_2^2}{\partial u^{(t)}}, \tag{14}$$

where γ represents the optimal step. Therefore, solving u sub-problem only requires computing the following equation iteratively:

$$\tilde{u}^{(t+1)} = u^{(t)} - \gamma\left(\Phi^T \Phi u^{(t)} - \Phi^T \tilde{y}_{f_i} + \mu\left(u^{(t)} - v^{(t)} - z^{(t)}\right)\right). \tag{15}$$

And then, the solution of the v sub-problem is dependent on the regularization term $\Gamma(v)$. If $\Gamma(v)$ is l_2-norm regularization term, it has the close form solution by a least square method, else if $\Gamma(v)$ is l_1-norm or l_0-norm regularization term, the above minimization problem is large-scale and highly non-convex. Some approximation approaches, such as TV [16], HQ [23], K-SVD [20] and GSR [21], have been proposed to solve this l_1-norm or l_0-norm problem. In this paper, $\Gamma(f_i)$ is defined as $\Gamma(f_i) = \sum_j \|\alpha_{G_{i,j}}\|_0$ to present the spatial–temporal group sparse representation. Given $z^{(t)}$ and $u^{(t+1)}$, the $v^{(t+1)}$ and $D^{(t+1)}$ sub-problem is

$$v^{(t+1)}, D^{(t+1)} = \arg\min_{v,D} \lambda \sum_j \|v_j\|_0 + \frac{\mu}{2}\|u^{(t+1)} - Dv - z^{(t)}\|_2^2. \tag{16}$$

For each group $G_{i,j}$ in the ith frame,

$$v_j^{(t+1)}, D_j^{(t+1)} = \arg\min_{v_j, D_j} \frac{\lambda}{\mu}\|v_j\|_0 + \frac{1}{2}\|D_j v_j - e_j^{(t+1)}\|_2^2 \tag{17}$$

where $e_j^{(t+1)} = u_j^{(t+1)} - z_j^{(t)}$. $D^{(t+1)}$ is the self-adaptive learned dictionary from $e_j^{(t+1)}$ using the scheme described above. Due to the unitary property of $D^{(t+1)}$ by using the scheme [21], the above mathematical expression is equivalent to

$$v_j^{(t+1)} = \arg\min_{v_j} \frac{\lambda}{\mu}\|v_j\|_0 + \frac{1}{2}\|v_j - r_j\|_2^2 \tag{18}$$

where $e_j^{(t+1)} = D_j r_j$ and $\|D_j v_j - D_j r_j\|_2^2 = \|v_j - r_j\|_2^2$. Therefore, the closed-form solution is expressed as the processing of hard thresholding on the r_j with the thresholding value $\sqrt{2\lambda/\mu}$.

It can be noticed that each sub-problem minimization may be much easier than the original problem Eq. (12). In fact, we acquire the efficient solution for each separated sub-problem, which enables the whole soft decoding algorithm more efficient. By averaging all the groups, each recovery pixel of the current frame is obtained as the average of the corresponding pixels in different groups.

4. Experiment result

In this section, we present experimental results of the divide-and-conquer based hierarchical video compressive sensing (CS) coding. All the experimental test video sequences are shown in Fig. 6, these video sequences *Bus*, *Foreman* and *Football* are of size 352×288 and *crew* and *City* are of size 704×576. The block size of BCS is set to be 32×32. Each video sequence has 96 frames with 12 GOPs. Concretely, the size of each block in the spatial–temporal group is set to be 8×8 and each spatial–temporal group has 60 blocks. The size of training windows for searching matched patches in current frame and its forward and backward reference frames are set to be identical. The numbers of best matched patches in the three frames are also set as the same values. $\sqrt{2\lambda/\mu}$ is set to be 67.5. Here, by taking $k=4$ as examples, we denote the subrate of 1th layer frames, 2nd layer frames, 3rd layer frames and 4th layer frames as S_{layer_1}, S_{layer_2}, S_{layer_3} and S_{layer_4} respectively. We set the sample subrate: $\overline{S}=0.3$, 0.4 and 0.5 as the average subrate in a GOP and let $S_{layer\_4} = n_4\overline{S}$, $S_{layer\_3} = S_{layer\_4} + n_3\overline{S}$, $S_{layer\_2} = S_{layer\_3} + n_2\overline{S}$, $S_{layer\_1} = S_{layer\_2} + n_1\overline{S}$. n_1,n_2,n_3,n_4 are object subrate factors for each layer. Obviously, S_{layer_1} has the highest subrate value of all the layers.

The comparison is conducted on some representative techniques in the literature: SPL-DWT [12], MH [24] and GSR [21] for

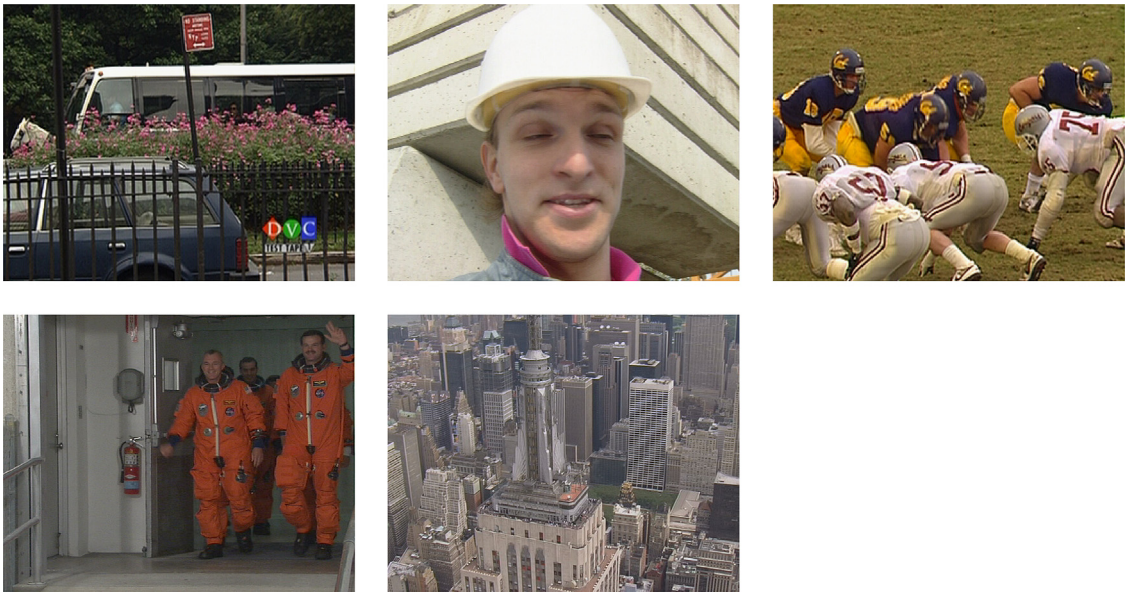


Fig. 6. All experimental test video sequences.

Table 1
Average PSNR in dB for several video sequences.

$\overline{S}$	Method	*Bus*	*Foreman*	*Football*	*Crew*	*City*	*Average*
0.3	SR-SPL-DWT	25.04	33.84	28.50	38.51	28.82	30.94
	SR-MH	26.93	34.23	30.03	38.90	36.36	33.29
	SR-GSR	28.10	36.04	31.96	39.05	33.52	33.72
	MC-BCS-SPL	28.10	34.50	28.60	37.51	32.69	32.28
	MH-BCS-SPL	27.19	36.3	28.86	39.36	33.17	32.97
	Hi-1ref-GSR	28.20	35.78	32.09	38.96	33.91	33.78
	Hi-STGSR	**30.32**	**37.54**	**33.94**	**39.76**	**35.74**	**35.46**
0.4	SR-SPL-DWT	26.65	35.72	32.12	40.02	29.70	32.84
	SR-MH	29.04	35.74	32.14	40.38	38.18	35.09
	SR-GSR	31.12	37.90	34.70	40.53	36.36	36.12
	MC-BCS-SPL	31.16	36.58	30.97	39.45	35.83	34.79
	MH-BCS-SPL	29.27	38.09	31.17	41.12	35.29	34.98
	Hi-1ref-GSR	31.27	37.68	34.82	40.55	36.73	36.21
	Hi-STGSR	**33.99**	**39.68**	**36.55**	**41.10**	**38.30**	**37.92**
0.5	SR-SPL-DWT	28.27	37.40	32.10	41.51	30.70	33.99
	SR-MH	31.32	37.23	34.13	41.85	39.83	36.87
	SR-GSR	34.02	39.70	36.96	41.91	38.39	38.19
	MC-BCS-SPL	33.63	38.78	33.22	40.96	37.56	36.83
	MH-BCS-SPL	31.31	39.69	33.50	42.81	37.28	36.91
	Hi-1ref-GSR	34.27	39.66	37.11	41.96	38.78	38.35
	Hi-STGSR	**37.24**	**41.46**	**38.83**	**42.40**	**40.04**	**39.99**

each frame of the video sequences with the same sample subrate (SR); the motion compensation (MC) based BCS-SPL [10] for the key-frame and non-key-frame CS coding with full-pixel motion search; the multihypothesis based key-frame and non-key-frame video CS coding [25]; GSR with one backward reference frame for the hierarchical based CS coding. In the proposed method, Hi-1ref-GSR means only one reference frame is used, and Hi-STGST means the proposed hierarchical video compressive sensing (CS) recovery with the forward and backward reference frames.

First, the results on the uncompressed block based video CS measurement are shown in Table 1. The PSNR comparisons for all the test frames in three video sequences under consideration in cases of the average frame sample subrates = 0.3, 0,4 and 0.5 are provided in Table 1. The proposed method provides quite promising results, achieving the highest PSNR among all the comparative methods over all the cases. The gains of recovery performance on the sample subrate = 0.3 by the proposed method are more than 4.5 dB, 2.1 dB, 1.7 dB, 3.1 dB and 2.4 dB over other methods: SR-SPL-DWT, SR-MH, SR-GSR, MC-BCS-SPL and MH-BCS-SPL. On the sample subrate = 0.4, the recovery performance gains by the proposed method are about 5 dB, 2.8 dB, 1.8 dB, 3.1 dB and 2.9 dB. When the sample subrate = 0.5, the proposed method outperforms these methods: SR-SPL-DWT, SR-MH, SR-GSR, MC-BCS-SPL and MH-BCS-SPL by 6 dB, 3.1 dB, 1.8 dB, 3.1 dB and 3 dB. Then we focus on the result of Hi-1ref-GSR, when comparing with Hi-1ref-GSR, the gain of the proposed method is about 1.7 dB, 1.7 dB and 1.6 dB on the different sample subrate = 0.3, 0.4 and 0.5.

Then, the results on the compressed block based video CS measurement by DPCM [2] are shown. Following [2,8,9], the actual bitrate is estimated using the zero order entropy of the quantizer indexes, which can be actually produced by a real entropy coder. In all cases, for the experiments, the quantization step is set to be 40 and sampling subrate is set to be 0.3, 0.4 and 0.5. The rate-distortion performance in PSNR in dB and bitrate in bpp is provided in Fig. 7. From the RD performance, the proposed method achieves the highest PSNR over all the cases. The proposed method can improve roughly more than 3.3 dB, 1.7 dB and 0.7 dB on average in comparison with SR-SPL-DWT, SR-MH and SR-GSR. The performances of the proposed method are 0.9 dB and 1 dB higher than MC-BCS-SPL and MH-BCS-SPL. The reconstruction of the proposed method achieves 0.6 dB higher performances on average than Hi-1ref-GSR with the same coded CS measurements.

5. Conclusion

The divide-and-conquer based hierarchical video compressive sensing (CS) coding framework is proposed, in which the whole video is independently divided into non-overlapped blocks of the hierarchical frames. The proposed hierarchical based framework outperforms the traditional framework through the better exploitation of frames correlation with reference frames, the unequal sample subrates setting among frames in different layers and the reduction of the error propagation. At the encoder, compare with the video/frame based CS, the proposed hierarchical block based CS matrix can be easily implemented and stored in hardware. Each measurement of the block in different hierarchical frame is obtained with the different sample subrates. At the

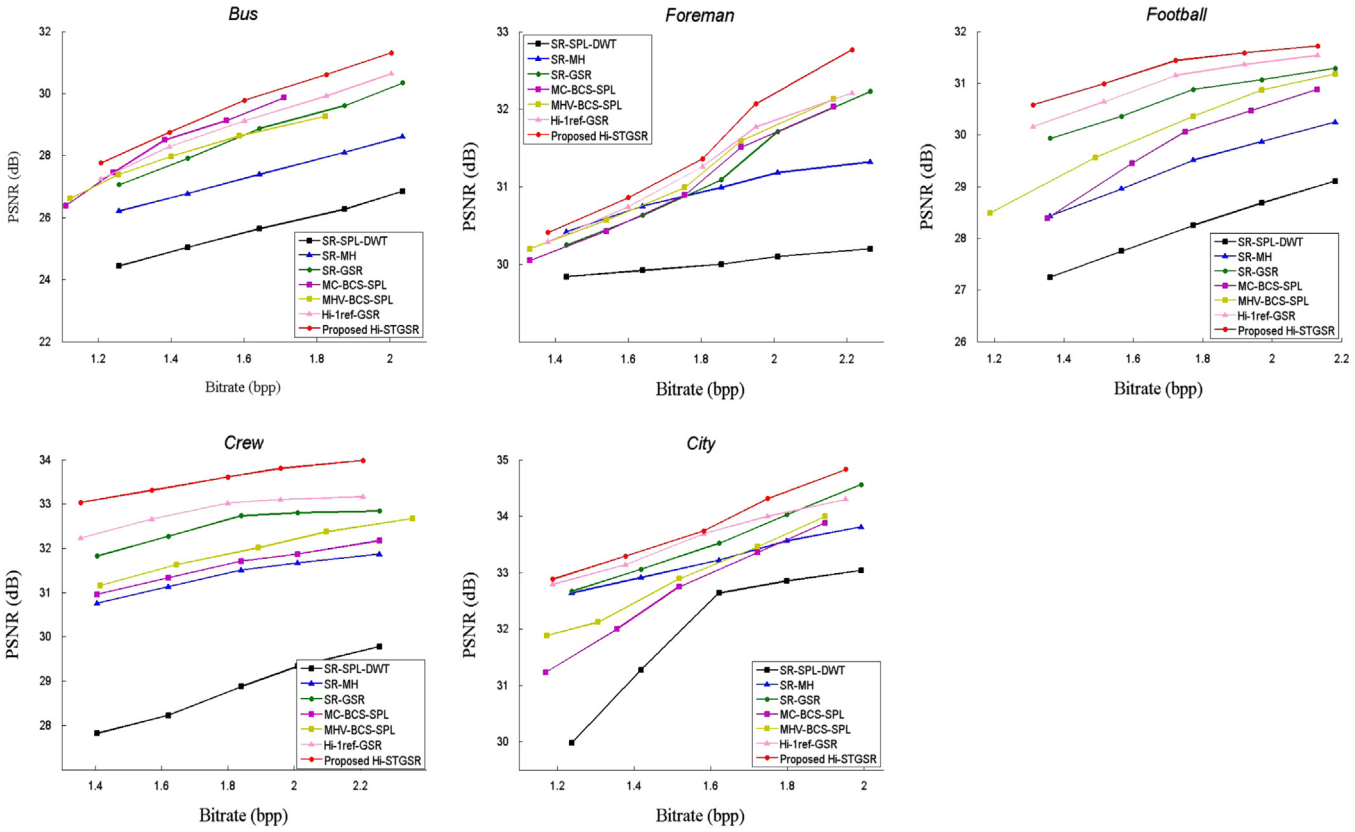


Fig. 7. The rate-distortion performance on test videos.

decoder, by considering the spatial and temporal correlations of the video sequence, a spatial–temporal sparse representation based recovery is proposed, in which the similar blocks in both the current frame and these recovered reference frames are composed as a spatial–temporal group unit to be sparse represented. At the decoder, by considering the spatial and temporal correlations of the video sequence, a spatial–temporal sparse representation based recovery is proposed, in which the similar blocks in the current frame and these recovered reference frames are organized as a spatial–temporal group unit to be represented sparsely. Finally, the recovery problem of video compressive sensing coding can be solved by adopting the split Bregman iteration. Experimental results show that the proposed method achieves better performance against SR-SPL-DWT, SR-MH, SR-GSR, MC-BCS-SPL and MH-BCS-SPL.

Acknowledgements

We would like to acknowledge the editors and reviewers, whose valuable comments greatly improved the manuscript. This work was supported in part by the Major State Basic Research Development Program of China (973 Program 2015CB351804) and the National Natural Science Foundation of China (NSFC) under Grant Nos. 61272386, 61390513, 61472101 and 61572155.

References

[1] Lu Gan, Block compressed sensing of natural images, in: 2007 15th International Conference on Digital Signal Processing, IEEE, 2007, pp. 403–406.

[2] Sungkwang Mun, James E. Fowler, DPCM for quantized blockbased compressed sensing of images, in: Proceedings of the European Signal Processing Conference, 2012, pp. 1424–1428.

[3] Aswin C. Sankaranarayanan, Christoph Studer, Richard G. Baraniuk, CS-MUVI: video compressive sensing for spatial-multiplexing cameras, in: IEEE International Conference on Computational Photography (ICCP), IEEE, 2012, pp. 1–10.

[4] Dikpal Reddy, Ashok Veeraraghavan, Rama Chellappa, P2c2: programmable pixel compressive camera for high speed imaging, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 329–336.

[5] Jason Holloway, Aswin C. Sankaranarayanan, Ashok Veeraraghavan, Salil Tambe, Flutter shutter video camera for compressive sensing of videos, in: IEEE International Conference on Computational Photography (ICCP), 2012, pp. 1–9.

[6] Vivek K. Goyal, Alyson K. Fletcher, Sundeep Rangan, Compressive sampling and lossy compression, IEEE Signal Process. Mag. 25 (2) (2008) 48–56.

[7] Gary J. Sullivan, Jens Ohm, Woo Jin Han, Thomas Wiegand, Overview of the high efficiency video coding (HEVC) standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1649–1668.

[8] Jian Zhang, Debin Zhao, Feng Jiang, Spatially directional predictive coding for block-based compressive sensing of natural images, in: IEEE International Conference on Image Processing, 2013, pp. 1021–1025.

[9] Khanh Quoc Dinh, Hiuk Jae Shim, Byeungwoo Jeon, Measurement coding for compressive imaging using a structural measurement matrix, in: IEEE International Conference on Image Processing, 2013, pp. 10–13.

[10] Sungkwang Mun, James E. Fowler, Residual reconstruction for block-based compressed sensing of video, in: Data Compression Conference (DCC), IEEE, 2011, pp. 183–192.

[11] Maria Trocan, Eric W. Tramel, James E. Fowler, Beatrice Pesquet, Compressed-sensing recovery of multiview image and video sequences using signal prediction, Multimed. Tools Appl. (2013) 1–27.

[12] Sungkwang Mun, James E. Fowler, Block compressed sensing of images using directional transforms, in: 2009 16th IEEE International Conference on Image Processing (ICIP), IEEE, 2009, pp. 3021–3024.

[13] Jian Zhang, Debin Zhao, Chen Zhao, Ruiqin Xiong, Siwei Ma, Wen Gao, Compressed sensing recovery via collaborative sparsity, in: Data Compression Conference (DCC), IEEE, 2012, pp. 287–296.

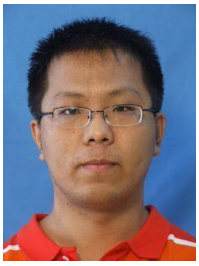
[14] Jian Zhang, Debin Zhao, Feng Jiang, Wen Gao, Structural group sparse representation for image compressive sensing recovery, in: Data Compression Conference (DCC), IEEE, 2013, pp. 331–340.

[15] Liang Xiao, Jun Shao, Lili Huang, Zhihui Wei, A novel compound regularization and fast algorithm for compressive sensing deconvolution, Neurocomputing 119 (2013) 131–138.
[16] Antonin Chambolle, An algorithm for total variation minimization and applications, J. Math. Imaging Vis. 20 (12) (2004) 89–97.
[17] M. Afonso, J. Miguel Sanches, Image reconstruction under multiplicative speckle noise using total variation, Neurocomputing 150 (2015) 200–213.
[18] Yi-Gang Cen, Rui-Zhen Zhao, Li-Hui Cui, Li-Hong Cen, Zhen-Jiang Miao, Zhe Wei, Defect inspection for tft-lcd images based on the low-rank matrix reconstruction, Neurocomputing 149 (2015) 1206–1215.
[19] Miyoun Jung, Xavier Bresson, Tony F. Chan, Luminita A. Vese, Nonlocal Mumford–Shah regularizers for color image restoration, IEEE Trans. Image Process. 20 (6) (2011) 1583–1598.
[20] Michal Aharon, Michael Elad, Alfred Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[21] Jian Zhang, Debin Zhao, Wen Gao, Group-based sparse representation for image restoration, IEEE Trans. Image Process. 23 (8) (2014) 3336–3351.
[22] Tom Goldstein, Stanley Osher, The split Bregman method for L1-regularized problems, SIAM J. Imaging Sci. 2 (2) (2009) 323–343.
[23] Ran He, Xiaotong Yuan, Wei-Shi Zheng, A fast convex conjugated algorithm for sparse recovery, Neurocomputing 115 (2013) 178–185.
[24] Chen Chen, Eric W. Tramel, James E. Fowler, Compressed-sensing recovery of images and video using multihypothesis predictions, in: Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), IEEE, 2011, 1193–1198.
[25] Eric W. Tramel, James E. Fowler, Video compressed sensing with multihypothesis, in: Data Compression Conference (DCC), IEEE, 2011, pp. 193–202.

Xinwei Gao received the B.S. and M.S. degrees in computer science from Harbin Institute of Technology (HIT), Harbin, China, in 2009 and 2011, respectively. From 2011 to 2012, he was with National Engineering Lab for Video Technology, Peking University, Beijing, as a research assistant. He is now working towards the Ph.D. degree at School of Computer Science and Technology, HIT. His current research interests are in image/video coding and processing.

Feng Jiang received the B.S. M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology (HIT), Harbin, China, in 2001, 2003, and 2008, respectively. He is now an Associated Professor in the Department of Computer Science, HIT and a visiting scholar in the School of Electrical Engineering, Princeton University. His research interests include computer vision, pattern recognition and image and video processing.

Shaohui Liu received the B.S. M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology (HIT), Harbin, China, in 2000, 2002, and 2007, respectively. He is now an Associated Professor in the Department of Computer Science, HIT and his research interests include data compression, pattern recognition and image and video processing.



Wenbin Che received the B.S. and M.S. degrees in astronautics from Harbin Institute of Technology (HIT), Harbin, China, in 2011 and 2013, respectively. He is now working towards the Ph.D. degree at School of Computer Science and Technology, HIT. His current research interests are in image/video coding and processing.

Xiaopeng Fan received the B.S. and M.S. degrees from the Harbin Institute of Technology (HIT), China, in 2001 and 2003 respectively, and the Ph.D. degree from Hong Kong University of Science and Technology (HKUST) in 2009. In 2009, he joined the School of Computer Science and Technology, Harbin Institute of Technology (HIT), where he is currently a Professor. From 2003 to 2005, he worked with Intel China Software Laboratory (ICSL) as Software Engineer. His current research interests are in image/video coding and processing, video streaming and wireless communication.

Debin Zhao received the B.S. M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology (HIT), Harbin, China, in 1985, 1988, and 1998, respectively. He is now a Professor in the Department of Computer Science, HIT. He has published over 200 technical articles in refereed journals and conference proceedings in the areas of image and video coding, video processing, video streaming and transmission, and pattern recognition.