SPATIAL-TEMPORAL RECOVERY FOR HIERARCHICAL FRAME BASED VIDEO COMPRESSED SENSING

Wenbin Che, Xinwei Gao, Xiaopeng Fan, Feng Jiang, Debin Zhao

Dept. of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{chewenbin, xwgao.cs, fxp, fjiang, dbzhao}@hit.edu.cn

ABSTRACT

In this paper, the hierarchical frame based video compressed sensing (CS) framework is proposed, which 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. By considering the spatial and temporal correlations of the video sequence, a spatial-temporal sparse representation based recovery is proposed for this framework. The similar blocks in both the current frame and these recovered reference frames are composed as a spatial-temporal group, which is defined as the unit of the sparse representation. By exploiting the low dimensional subspace description of each group, the video CS recovery is converted as a low-rank matrix approximation problem, which can be solved by exploiting the hard thresholding and the gradient descent. Experimental results show that the proposed method achieves better performance against both the state-of-art still-image CS recovery algorithms and the existing residual domain based video CS reconstruction approaches.

Index Terms— Video compressed sensing, hierarchical structure framework, spatial-temporal sparse representation

1. INTRODUCTION

As a new methodology of signal-sampling and recovery, compressed sensing(CS) has been extensively studied in recent years. As applied to video frames, this theory makes the sampling process faster than traditional sampling methods. Significant process in video CS has been made with a single-pixel cameras[1], based on representing a video in the Fourier domain or the wavelet domain. However, video CS faces challenges including high recovery quality at a relatively low subrate[2]. Low subrate which makes it easier to capture video sequences at a high speed by camera will result in a poor recovery performance using the still-image CS recovery algorithms. By considering the spatial and temporal correlations, it is possible to achieve a high-quality even employing a low subrate[3]. Mun et al.[4] proposed a residual recovery based on Motion Compensation(MC), which utilized the temporal redundancy and residual sparse property in video sequence. Two subtrates are used in sampling stage of the residual recovery, where high subrate is adopted for key frames and low subrate for non-key frames.

In the CS theory, the signal can be well recovered if it is sparse enough in some domain. Mun et al.[5] cast the CS reconstruction in the base of contourlet transform or complex-valued dual-tree wavelet transform(DWT), 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.[6, 7]). In [6], 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 [6], enforcing local smoothness and non-local sparsity simultaneously. A group sparse representation (GSR) modeling was further developed in [7], using the non-local grouping technique as well. In essence, this modeling efficiently utilized the intrinsic low-rank property 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.

In this paper, we consider the Block Compressed Sensing(BCS) recovery of video sequences in which the hierarchical structure and group sparse representation based method are used to aid the recovery process. We employ different subrates for different layers. The 3D patch matching modeling, the hard thresholding and the gradient descent are also adopted to the recovery stage. It can be found in experimental simulations that the proposed CS recovery based on hierarchical structure outperforms the state-of-art still-image recovery method. Additionally, the proposed technique exceeds the quality of residual domain based reconstruction by a large margin.

This work has been supported in part by the Major State Basic Research Development Program of China (973 Program 2015CB351804), the National Science Foundation of China under Grant No. 61272386.

2. BACKGROUND

In image compressed sensing theory, we focus on the situation in which the image is sampled by BCS[8]. Consider the classical CS problem:

$$\min_x \|\alpha\|_0,\ s.t\ y = \Phi x, x = \Psi\alpha, \tag{1}$$

where x is real valued signal to recover with length N, y represents an sampled vector with length M ($M \ll N$), and Φ is an $M \times N$ CS measurement matrix. α represents the sparse code vector of x over Ψ. The ideal recovery procedure searches for the α with the smallest ℓ_0 norm consistent with the observed y. In BCS, the input image is divided into small non-overlapping blocks and sampled with operator Φ_B which is related to the subrate. Let x_i represent the vectorized signal of the $i\text{-}th$ block, the output vector is

$$y_i = \Phi_B x_i, \tag{2}$$

where y_i represents the measurements of the $i\text{-}th$ block. For simplicity, the sampling process can be rewritten as

$$Y = \Phi_B X, \tag{3}$$

where $X = (x_1, ..., x_i, ...)$ covers all the pixels of the input image and the corresponding output is $Y = (y_1, ..., y_i, ...)$.

Because the length of x_i is longer than that of y_i, it is impossible to recover X from measurements Y in general; however, if x is sufficiently sparse, exact recovery is possible[9]. In this case, the key to CS recovery is the production of a sparse set of significant transform coefficients, $X = \Psi\alpha$. The recovery performance depends on the sparsity of α. As [7] has concluded, the sparsity of image signal can be well exhibited by group sparse representation. Using GSR,we can find a solution to (3) as follow:

$$\min \|\alpha_i\|_0\ , s.t\ Y = \Phi_B X, G_i = \Psi_i \alpha_i, \tag{4}$$

where G_i consists of similar patches from images[7].

Since the matched image patches exploits the similarity of overlapping patches, the effective rank of each group is low. Let $G_i = U_i \Sigma_i V_i^T$ be the singular value decomposition (SVD) for G_i and n be the number of singular values. Suppose the number of non-zero largest singular value is k_i, we have $k_i \ll n$ and other singular values in G_i tend to be near-zero, the matrix $G_i = U_i \Sigma_i V_i^T$ is really sparse or nearly sparse. Under this condition, Ψ_i can be defined as

$$\Psi_i = [d_{G_i \otimes 1}, d_{G_i \otimes 2} ... d_{G_i \otimes n}], \tag{5}$$

where $d_{G_i \otimes j} = u_{G_i \otimes j} {v^T}_{G_i \otimes j}$, $u_{G_i \otimes j}$ and $v_{G_i \otimes j}$ are column vectors of U_i and V_i.

[7] utilized the aid of the particular dictionary learning method by SVD to derive the group sparsity model. Hence, the lower the rank of G_i is, the sparser the coefficients of G_i will be. Hence, (4) is equivalent to the following minimization problem

$$\min \|rank(G_i)\|, s.t\ Y = \Phi_B X, \tag{6}$$

which is solved iteratively incorporated with gradient descent in this paper.

3. VIDEO RECOVERY BASED ON HIERARCHICAL FRAME STRUCTURE

3.1. Hierarchical Structure

The image Compressed Sensing theory can be extended to video signal. The straightforward consideration is that video frames in the sequence are sampled in a 2D fashion at the same subrate and recovered independently. However, obviously spatial and temporal correlations can be utilized in video CS framework. One efficient form, MC-BCS-SPL, is proposed in [4] where two subrates are used at the sampling stage. Each Group of Pictures(GOP) contains a key frame which is sampled at a relatively high subrate and other non-key frames sampled at a low subrate. At the recovery stage, reference frames are used to achieve a better performance than independent recovery method. As the GOP size increases, the effect of key frame will be lessened.

In this paper, we proposed the CS hierarchical structure of video sequences. Suppose each GOP contains 2^n frames, the hierarchical structure when $n = 3$ is illustrated in Figure 1. Take the first GOP for example, we define $n + 1$ different layers. At the sampling stage,different subrates are used as:

$$S_{layer\_i} > S_{layer\_j},\ s.t.\ i < j, \tag{7}$$

S_{layer_i} represents the subrate of higher layer than $layer\_j$. The measurement matrix should also be adjusted by which layer the current frame belongs to. The recovery process begins from the highest layer frames, i.e., layer_1, which is sampled at the highest subrate. We will discuss the detail about the recovery algorithm in subsection 3.2.

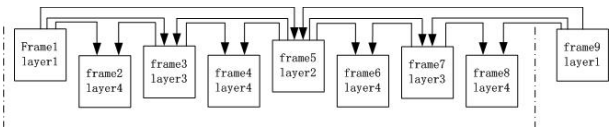

Fig. 1. Hierarchical Structure in a GOP. Each GOP contains 8 frames and every frame is recovered with two reference frames except layer_1.

3.2. Recovery algorithms

Given a frame-by-frame acquisition, the most straightforward recovery would be to recover the individual frames independently using BCS-SPL[8]. However, such a method ignores

the fact that consecutive video frames are usually highly correlated.

In MC-BCS-SPL[4], the recovery performance is determined by accuracy of recovered reference frames. Because all the non-key frames are sampled at a low subrate and they would be used as reference frame once recovered, the current frame can not always use a high-quality recovered frame as its reference.

Fig. 2 shows that the higher the subrate of reference frame is, the better the performance would be. The PSNR would be lower without reference. For this reason, We will select recovered frames with higher subrate as reference frames at the recovery stage. In Fig. 3, it is quite obvious that the recovery performance will be improved if the recovered frames with nearer distance from current frame are selected as reference frames. As illustrated in Fig. 1, in our proposed hierarchical structure in the Video CS, any frame will always be recovered using nearest higher layer reference frame.

Our proposed hierarchical structure will efficiently employ the correlation among frames and reduce the recovery error propagation. Suppose the GOP size is 2^n. The recovery error will propagate with 2^n-1 times in[4]. In our proposed framework, it will be reduced to $\log_2 2^n = n$ times. Hence, the proposed method can achieve better performance than traditional methods, such as BCS-SPL and MC-BCS-SPL.

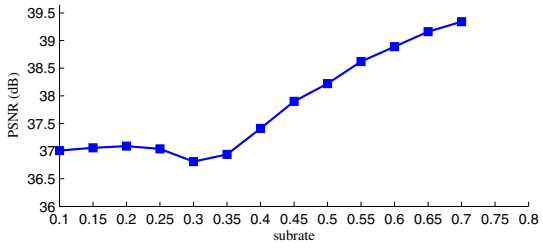

Fig. 2. The PSNR of recovered frame for sequence $foreman$ when subrate of referene frame changes. The PSNR is 36dB when recovered independently, worse than that of reference frame based method. The subrate of current frame is 0.3.

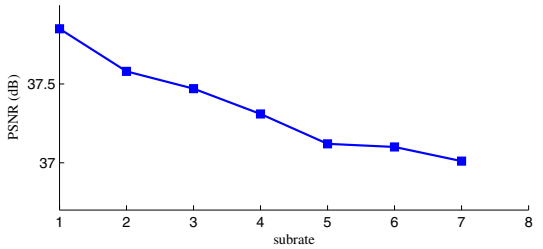

Fig. 3. The PSNR of recovered frame for sequence $foreman$ when distance between current and reference frame changes. The subrate of current frame is 0.3.

Our algorithm will utilize the techniques of 3D patch matching, hard thresholding and gradient descent. Spatial-temporal patch matching method is proposed in this paper for group construction: First, divide every frame into m overlapping patches of size $\sqrt{B} \times \sqrt{B}$ and each patch is denoted by x_k, i.e., $k = 1, 2, ..., m$. Suppose the current frame is X_c and its forward reference frame and backward reference frame are X_f and X_b. Then, define the group G_k corresponding to x_k, whose columns are the vectorized matched patches, i.e., $G_k = \{G_{k\otimes c}, G_{k\otimes f}, G_{k\otimes b}\}$, where $G_{k\otimes c}, G_{k\otimes f}, G_{k\otimes b}$ are groups of similar patches from current, forward and backward reference frames. Hard thresholding is used in keeping the low-rank property of G_k. Suppose $G_k = U\Sigma V^T$ is the SVD for G_k and σ_j denote the j-th largest singular value. The hard thresholding process can be described as Algorithm 1.

Algorithm 1: $\bar{G}_k = H(G_k, \tau),\ s.t\ \min \|rank(G_k)\|$

$G_k = U\Sigma V^T, \sigma_j = \Sigma(j, j)$;

$$\bar{\sigma_j} = \begin{cases} \sigma_j\,, & \sigma_j \geq \tau \\ 0\,, & \sigma_j < \tau \end{cases}$$

$\bar{\sigma} = (\bar{\sigma_1}, ..., \bar{\sigma_j}, ...),\ \Sigma_\tau = diag(\bar{\sigma})$;

$\bar{G}_k = U\Sigma_\tau V^T$

Algorithm 2: VCS recovery for the current frame based on Hierarchical Structure

Input: $Y = (y_1, ..., y_k, ...), \phi_B$ and τ
Output: Recovered current frame $\hat{X}$

for *iteration number* $iter = 1, 2, ..., Max\_iter$ **do**
  for *each block* k **do**
    $\tilde{x}_k = x_k + \phi_B^T(y_k - \phi_B x_k)$;
  $\tilde{X} = (\tilde{x}_1, ..., \tilde{x}_k, ...)$;
  create groups G_k by spatial-temporal patch-matching from current and reference frames for $\tilde{x}_k$;
  for *each group* G_k **do**
    $\bar{G}_k = H(G_k, \tau)$
  update the current frame by weighted averaging all the reconstructed groups $\bar{G}_k$;

By the function of $\bar{G}_k = H(G_k, \tau)$, the rank of G_k becomes lower. As discussed in section 2, nonlocal self-similarity of natural images can be effectively characterized and signal sparsity is better exhibited to improve reconstruction performance in the recovery process described below.

For any arbitrary vector x, to find its approximation $\tilde{x}$ on the hyper-plane $\{z : y = \phi_B z\}$, the matrix size of ϕ_B is decided by the subrate in different layers. The gradient descent technique uses the following formula

$$\tilde{x} = x + \phi_B^T(\phi_B\phi_B^T)^{-1}(y - \phi_B x) \tag{8}$$

As ϕ_B is intercepted from a random orthonormal matrix in our experimental simulations, that is, $\phi_B \phi_B^T = I$, (8) can be simplified into

$$\tilde{x} = x + \phi_B^T (y - \phi_B x) \tag{9}$$

The detailed recovery algorithm for one frame is described in Algorithm 2, and recoveries for the other frames follow the same rule.

4. EXPERIMENT RESULT

In this section, we will examine the performance of the proposed reconstruction method. All of the video sequences are sampled by BCS with block size of 32×32. In our recovery implementation, all parameters are set empirically. Concretely, the size of each patch, i.e, $\sqrt{B} \times \sqrt{B}$, is set to be 8×8.

Then the recovery performance of proposed method will be evaluated. To avoid the inaccuracy due to violent movements, the size of training window in two reference frames is modified as:

$$L_f = L_b = L_c + (ref\_dist - 1) \times 8 \tag{10}$$

where ref_dist represents the distance between current frame and its two references.

The numbers of best matched patches in the three frames are also set as the same values,i.e., $l_f = l_c = l_b = 20$. τ is set to be 67.5. We denote the subrate of layer_1 frame, layer_2 frame, layer_3 frame and layer_4 frame as $S_{layer\_1}, S_{layer\_2}, S_{layer\_3}, S_{layer\_4}$, respectively. We set $\bar{S} = 0.3,\ 0.4\ or\ 0.5$ as the average subrate in a GOP and let $S_{layer\_4} = n_4 \bar{S}, S_{layer\_3} = S_{layer\_4} + n_3 \bar{S}$, $S_{layer\_2} = S_{layer\_3} + n_2 \bar{S}$, $S_{layer\_1} = S_{layer\_2} + n_1 \bar{S}$. n_1,n_2,n_3,n_4 are object subrate factors for each layer. Obviously, S_{layer_1} is the highest subrate value.

The proposed method is compared with three alternatives, i.e., DWT, SGSR and MC-BCS-SPL on testing former 80 frames of common video sequences. DWT is a recovery method which recovers video sequences frame by frame independently. SGSR is known as the state-of-art algorithm for image CS recovery. MC-BCS-SPL is a typical residual domain based recovery scheme, the critical process of which is the residual recovery based on motion compensation(MC). We implement it with full-search ME using quarter-pixel accuracy and a search window of ± 30 pixels. To demonstrate the fact that the unequal sample subrates setting among frames in different layers and hierarchical structure make the contribution to better performance of recovery in our proposed method, two cases of experiment are set. In case 1, we use the proposed recovery method with No Hierarchical Structure and only the Previous frame is used as a reference frame for each current frame(NHSP). All the frames are sampled at the same subrate. In case 2, Hierarchical Structure is used but all the frames are sampled at the Same subrate(HSS).

The PSNR comparisons for all the test frames in three video sequences under consideration in cases of 30% to 50% average frame subrates are provided in Table 1. The proposed method provides quite promising results, achieving the highest PSNR among all the comparative algorithms over all the cases.

Then we take $\bar{S} = 0.3$ as an example to show the gain of recovery performance by proposed method over other alternatives. Compared with DWT, SGSR, MC-BCS-SPL, the proposed method improved about 4dB, 1dB and 4dB. We now focus on the result of NHSP. When using hierarchical structure at the stage of recovery, the proposed method can improve more than 1dB. By comparing the proposed method wtih HSS, the performance is about 0.6dB higher. The result reveals that the hierarchical structure and unequal subrate setting provides en effective contribution to the recovery performance.

Table 1. Average PSNR in dB for several video sequences

$\bar{S}$	method	Foreman	Football	Bus
0.3	DWT	33.84	28.50	25.04
	SGSR	36.04	31.96	28.10
	MC-BCS-SPL	33.66	30.15	27.73
	NHSP	35.78	32.09	28.20
	HSS	36.93	32.95	29.09
	proposed	**37.54**	**33.94**	**30.32**
0.4	DWT	35.72	30.29	26.65
	SGSR	37.90	34.70	31.12
	MC-BCS-SPL	35.80	32.66	29.58
	NHSP	37.68	34.82	31.27
	HSS	39.02	35.75	32.83
	proposed	**39.68**	**36.55**	**33.99**
0.5	DWT	37.40	32.10	28.27
	SGSR	39.70	36.96	34.02
	MC-BCS-SPL	38.05	33.65	31.12
	NHSP	39.66	37.11	34.27
	HSS	40.88	38.04	36.28
	proposed	**41.46**	**38.83**	**37.24**

5. CONCLUSION

In this paper, we proposed a hierarchical frame framework to address video recovery problem based on video compressed sensing theory. By employing unequal subrate in different layers, we can effectively utilize the spatial and temporal correlations of the video sequence. Incorporating the techniques of 3D patch matching, hard thresholding and gradient descent, our scheme is implemented in an iterative fashion. Experimental results reveal that the developed video CS recovery strategy is able to increase the recovery by a large margin compared with the current existing methods.

6. REFERENCES

[1] A. C. Sankaranarayanan, C. Studer, and R. G. Baraniuk, "CS-MUVI: Video compressive sensing for spatial-multiplexing cameras," in *IEEE International Conference on Computational Photography*, 2012, pp. 1–10.

[2] J. Yang, X. Yuan, X. Liao, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Video compressive sensing using gaussian mixture models," *IEEE Transations on Image Processing*, vol. 23, pp. 4863–4878, 2014.

[3] C. Chen, E. W. Tramel, and J. E. Fowler, "Compressed-sensing recovery of images and video using multihypothesis predictions," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*. IEEE, 2011, pp. 1193–1198.

[4] S. Mun and J. E. Fowler, "Residual reconstruction for block-based compressed sensing of video," in *Proceedings of the IEEE Data Compression Conference*, March 2011, pp. 183–192.

[5] S. Mun and J. E. Fowler, "Block compressed sensing of images using directional transforms," in *Proceedings of the International Conference on Image Processing*, 2009, pp. 3021–3024.

[6] J. Zhang, D. Zhao, C. Zhao, R. Xiong, S. Ma, and W. Gao, "Compressed sensing recovery via collaborative sparsity," in *Proceedings of the IEEE Data Compression Conference*, April 2012, pp. 287–296.

[7] J. Zhang, D. Zhao, F. Jiang, and W. Gao, "Structural group sparse representation for image compressive sensing recovery," in *Proceedings of the IEEE Data Compression Conference*, March 2013, pp. 331–340.

[8] L. Gan, "Block compressed sensing of natural images," in *Proceedings of the International Conference on Digital Signal Processing*, July 2007, pp. 403–406.

[9] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Februray 2006.