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 $\Phi_v$, ($M \ll N$). 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 $\Phi_f$

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

where $f_i \in R^N$ is lexicographically stacked representations of the $i$th frame and $y_{f_i} \in R^M$ is the CS measurements observed by a random $M \times N$ measurement matrix $\Phi_f$. 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 $\Phi$, 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 $\Psi$, which has greatly changed the way engineers think of data acquisition,

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

If $\theta$ is a sparse coefficient vector, the signal $x$ is sparse under the domain $\Psi$. 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