Acta Automatica Sinica
Vol. 40, No. 10 October, 2014

Residual Distributed Compressive Video Sensing Based on Double Side Information

CHEN Jian[1] SU Kai-Xiong[1] WANG Wei-Xing[1] LAN Cheng-Dong[1]

Abstract Compressed sensing (CS) is a novel technology to acquire and reconstruct sparse signals below the Nyquist rate. It has great potential in image and video acquisition and processing. To effectively improve the sparsity of signal being measured and reconstructing efficiency, an encoding and decoding model of residual distributed compressive video sensing based on double side information (RDCVS-DSI) is proposed in this paper. Exploiting the characteristics of image itself in the frequency domain and the correlation between successive frames, the model regards the video frame in low quality as the first side information in the process of coding, and generates the second side information for the non-key frames using motion estimation and compensation technology at its decoding end. Performance analysis and simulation experiments show that the RDCVS-DSI model can rebuild the video sequence with high fidelity in the consumption of quite low complexity. About $1 \sim 5$ dB gain in the average peak signal-to-noise ratio of the reconstructed frames is observed, and the speed is close to the least complex DCVS, when compared with prior works on compressive video sensing.

Key words Video coding, compressed sensing, distributed compressive video sensing, residual coding

Citation Chen Jian, Su Kai-Xiong, Wang Wei-Xing, Lan Cheng-Dong. Residual distributed compressive video sensing based on double side information. *Acta Automatica Sinica*, 2014, **40**(10): 2316−2323

The compressed sensing (CS) theory[1−2] put forward by Donoho and Baraniuk et al. during $2004 \sim 2006$ shows that the high dimensional signal can be projected to a low dimensional space through an observation matrix incoherent with the transform basis, as long as the signal is sparse in a certain transform domain. Using a few observations, the signal can be reconstructed precisely. Recent years, the researches on reconstruction algorithm and measurement scheme based on CS have made significant progress[3−10]. The application of CS theory about video coding is still in an exploratory stage, but it has showed great development prospects[11].

In 2006, Wakin et al. obtained sampling data through a single pixel camera[12], and reconstructed the frames via the sparsity in the 2-D wavelet domain (referred to as 2D-CS) and a group of frames via the sparsity in the 3-D wavelet domain (referred to as 3D-CS)[13]. In order to reduce the computing burden of image or video compression, Lu put forward block compressed sensing of natural image (Block-CS) in 2007[14]. Then, some scholars applied the Block-CS into video coding[15−16]. It reduced the computational complexity significantly, however, its reconstructing performance was not ideal. To make full use of inter-frame correlation between moving pictures for further improving coding efficiency, some scholars made a CS coding model for the residual video (referred to as RVCS)[17−18]. During 2009, Do et al. proposed a kind of distributed compressed video sensing DISCOS architecture[19], Prades-Nebot et al. suggested the distributed video coding based on CS (DVC-CS)[20], while Kang et al. studied another version of distributed compressive video sensing (DCVS)[21]. After 2010, more and more scholars further researched on video CS based on frame or block, inter-frame residuals, as well as distributed video coding[22−26]. While those methods have improved the quality of video reconstruction to some extent, the complexity is also increasing.

To fully utilize the correlation of intra-frame and inter-frame, a coding algorithm called residual distributed compressive video sensing based on double side information (RDCVS-DSI) is proposed in this paper to rebuild video sequence in high fidelity under the conditions of lower complexity.

1 Compressive video sensing

1.1 Compressed sensing

According to the CS theory[1−3], signal $\boldsymbol{x}$ can be sparsely represented under some basis $\boldsymbol{\psi}^{N\times N}$

$$\boldsymbol{x} = \boldsymbol{\psi\theta}, \quad \boldsymbol{x} \in \mathbf{R}^N \tag{1}$$

where $\boldsymbol{\theta}$ is the transform coefficients of $\boldsymbol{x}$ in $\boldsymbol{\psi}$ domain. When $\boldsymbol{\theta}$ has only S ($S \ll N$) nonzero elements, signal $\boldsymbol{x}$ is S-sparse under the basis of $\boldsymbol{\psi}$. Partial Fourier transform, DCT and DWT are commonly used in sparse transform. In compressed sampling, signal $\boldsymbol{x}$ is projected into a set of measurement vectors of $\boldsymbol{\phi}$ to give the measured value $\boldsymbol{y}$, i.e.,

$$\boldsymbol{y} = \boldsymbol{\phi x}, \quad \boldsymbol{y} \in \mathbf{R}^M \tag{2}$$

where $\boldsymbol{y}$ is an $M \times 1$ measured values matrix, $\boldsymbol{\phi}$ is an $M \times N$ measurement matrix ($M \ll N$) incoherent with $\boldsymbol{\psi}$[27]. Gaussian, Bernoulli, scramble Fourier and scramble block Hadamard ensemble (SBHE) have been shown to be good choices for the measurement matrix $\boldsymbol{\phi}$.

Compressed sampling is a dimension reduction process, which helps reduce the number of collected data from N to M. However, it also makes the recovery of signal $\boldsymbol{x}$ from measurements $\boldsymbol{y}$ an ill-posed problem. The CS theory states that the reconstruction can be formulated as an l_p minimization problem by solving:

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{\theta}||_{l_p} \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{\phi\psi\theta} \tag{3}$$

To solve the above optimization problem, many techniques have been proposed in the literature, e.g., orthogonal matching pursuit (OMP)[28], two-step iterative shrinkage/thresholding (TwIST)[29], gradient projection for sparse reconstruction (GPSR)[30], and sparse reconstru-

Manuscript received September 24, 2013; accepted April 21, 2014
Supported by National Natural Science Foundation of China (61170147), Major Cooperation Project of Production and College in Fujian Province (2012H61010016), and Natural Science Foundation of Fujian Province (2013J01234)
Recommended by Associate Editor HUANG Qing-Ming
1. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China

ction by separable approximation (SpaRSA)[31]. If signal $\boldsymbol{x}$ is 2-D data, the above theory can be directly applied to image compression combining with existing CS acquisition device.

1.2 Compressive video sensing

Relative to the CS imaging, compressive video sensing has more stringent requirements on storage resources and real-time processing. Simultaneous temporal and spatial measurement by the 3D-CS is impractical, and thus one opts for frame-by-frame measurement. The most straight compressive video sensing scheme 2D-CS, adopting frame-by-frame CS for image, takes measurement and solution according to formulas (2) and (3).

For alleviating the huge computation and memory burdens, the Block-CS has been introduced into video coding. In the Block-CS, each frame is divided into N_B non-overlapping blocks $\boldsymbol{x}_j$ (sized $B \times B$, subscript j denotes block indicator), and acquired using a suitable $M_B \times B^2$ measurement matrix $\boldsymbol{\phi}_B$, then the corresponding $\boldsymbol{y}_j$ is

$$\boldsymbol{y}_j = \boldsymbol{\phi}_B \boldsymbol{x}_j, \quad \boldsymbol{x}_j \in \mathbf{R}^{B\times B}, \quad \boldsymbol{y}_j \in \mathbf{R}^{M_B} \tag{4}$$

It is straightforward to see that (4) applying block-by-block to an image is equivalent to a whole-image measurement matrix $\boldsymbol{\phi}$ in (2) with a constrained structure that $\boldsymbol{\phi}$ is block diagonal[14, 22],

$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_B & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\phi}_B & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\phi}_B \end{bmatrix} \tag{5}$$

When sparsity transform $\boldsymbol{\psi}$ is also a block-based operator, the frame can be reconstructed by block at the decoding end. In general, block-independent reconstruction will produce severe blocking artifacts, thus rebuilding by frame is prior to block. In convenience, we focus on the CS for video by frame. Since a structural measurement matrix[32] in the form of (5) is used in this paper, the following scheme is equally applicable to the compressive video sensing by block.

According to the temporal redundancy of video, the correlation model between successive video frames $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ can be expressed as:

$$\begin{cases} \boldsymbol{x}_t = \boldsymbol{x}_c + \boldsymbol{x}_{t\_u} \\ \boldsymbol{x}_{t+1} = \boldsymbol{x}_c + \boldsymbol{x}_{t+1\_u} \end{cases} \tag{6}$$

where $\boldsymbol{x}_c$ is the common portion between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$, while $\boldsymbol{x}_{t\_u}$ and $\boldsymbol{x}_{t+1\_u}$ are the specific portions. The RVCS and DCVS are the two typical schemes for compressive video sensing based on the above correlation model.

The basic idea of RVCS[17] comes from the traditional inter-frame coding. By using the same measurement matrix in a group of pictures (GOP), the difference of measurements between adjacent frames is equivalent to the projection of inter-frame residuals, i.e.,

$$\boldsymbol{y}_{t+1} - \boldsymbol{y}_t = \boldsymbol{\phi}\boldsymbol{x}_{t+1} - \boldsymbol{\phi}\boldsymbol{x}_t = \boldsymbol{\phi}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) \tag{7}$$

Therefore, video residuals can be acquired by the single-pixel camera and a subtraction operation. For the scene of slow motion or video surveillance, the neighboring frames are much similar and inter-frame residuals have an intensive sparsity, so it is more conducive to be measured and rebuilt via CS. But for a general video sequence, the reconstructing performance changes with inter-frame residuals.

The DISCOS[19] and DVC-CS[20] introducing the traditional video coding method to key-frame coding demand the traditional camera to sample those data. In view of low cost, we only discuss the DCVS[21] in which all frames can be acquired by CS camera. The DCVS combines the idea of distributed compressed sensing (DCS) and distributed video coding (DVC), and it regards video sequences as the relative sources in the joint sparse model (JSM). Each frame is taken CS measurement individually at the encoding end, and the non-key frames are jointly reconstructed based on the side information at the decoding end. As the coding scheme of DCVS is concise and the decoding algorithm is very flexible, it is especially suitable for many fields such as low-cost digital camera, power-saving and mobile video collecting equipment, distributed sensor network, and so on. However, its reconstruction performance still needs further improving.

2 RDCVS-DSI

2.1 The basic idea and framework for RDCVS-DSI

In general, the coding algorithms of RVCS and DCVS improve the 2D-CS algorithm only in view of inter-frame correlation between video frames, but not increase the coding efficiency through their own characteristics. According to the correlation between successive frames described in (6), finding an appropriate side information ($\boldsymbol{x}_c$), the $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ can be encoded through compressing the $\boldsymbol{x}_{t\_u}$ and $\boldsymbol{x}_{t+1\_u}$. As the residual sparsity is very strong, it is more advantageous to carry on CS coding.

This study intends to regard the low quality image of the original frame $\boldsymbol{x}_l$ (similar to $\boldsymbol{x}_c$) as the side information of the $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ for reference. The successive inter-frame model in (6) can be expanded to a related model between the key frame $\boldsymbol{x}_k$ and the multiple non-key frames $\boldsymbol{x}_{nk}$ in a GOP, then

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{x}_{k\_l} + \Delta\boldsymbol{x}_k \\ \boldsymbol{x}_{nk} = \boldsymbol{x}_{nk\_l} + \Delta\boldsymbol{x}_{nk} \\ \boldsymbol{x}_{nk\_l} = f(\boldsymbol{x}_{k\_l}) \end{cases} \tag{8}$$

where $\boldsymbol{x}_{k\_l}$ represents the low quality version of the key frame, $\boldsymbol{x}_{nk\_l}$ represents the low quality version of the non-key frame, $\triangle\boldsymbol{x}_k$ and $\triangle\boldsymbol{x}_{nk}$ represent the residuals between the key/non-key and its low quality version, and $f(\cdot)$ indicates the relationship between the key and non-key frames in low quality version.

In order to guarantee quickly obtaining the reference frame in low quality version both in the encoding and decoding ends, the first side information (SI1) is considered to be generated with a large amount of information and a few data at the encoding end, and it is sent to the decoding end together with the measurement value, so as to be quickly converted to the reference information for decoding. As wavelet transform has a characteristic of time-frequency scalability, and its main energy concentrates in the low frequency, the wavelet coefficients are taken in the lowest layer as SI1.

Because there is a strong sparsity in detailed information of the difference between the key/non-key frame and its low quality version in the same GOP, the residuals for a key or non-key frame can be measured respectively. As the SBHE has the advantages of good performance, simple operation, less memory, etc., it is suitable for video measuring