COMPRESSIVE VIDEO SENSING BASED ON USER ATTENTION MODEL *

Jie Xu1, 2, Jianwei Ma3, Dongming Zhang1, Yongdong Zhang1, Shouxun Lin1

1 Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
2 Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
3 School of Computational Science, Florida State University, Tallahassee, Florida, 32306, USA

ABSTRACT

We propose a compressive video sensing scheme based on user attention model (UAM) for real video sequences acquisition. In this work, for every group of consecutive video frames, we set the first frame as reference frame and build a UAM with visual rhythm analysis (VRA) to automatically determine region-of-interest (ROI) for non-reference frames. The determined ROI usually has significant movement and attracts more attention. Each frame of the video sequence is divided into non-overlapping blocks of $16 \times 16$ pixel size. Compressive video sampling is conducted in a block-by-block manner on each frame through a single operator and in a whole region manner on the ROIs through a different operator. Our video reconstruction algorithm involves alternating direction $l_1$-norm minimization algorithm (ADM) for the frame difference of non-ROI blocks and minimum total-variance (TV) method for the ROIs. Experimental results showed that our method could significantly enhance the quality of reconstructed video and reduce the errors accumulated during the reconstruction.

Index Terms— compressive sensing, video, user attention model, ROI

1. INTRODUCTION

The recently emerged theory of compressive sensing (CS) based on Candes’ and Donoho’s work suggests a new framework for data acquisition [1-5]. The basic principle of CS is that sparse or compressible signals can be recovered from a surprisingly small number of random measurements generated from those signals with high accuracy. According to the Shannon/Nyquist sampling theorem, to capture a lossless signal, the sampling rate must be at least twice the maximum frequency presented in those signals. However, the CS suggests that one can recover certain signals from far fewer measurements than traditional methods if the signals themselves are compressible. The term “compressible” means that the signals have very limited nonzero elements or nonzero transformed coefficients.

During traditional data acquisition procedure, transform coding plays a crucial role: the signal is full sampled first; then it is transformed into a certain domain; the most important transformed coefficients are located and kept while the rests are discarded. This process is inefficient because we must acquire a mass of raw images or video data and the transform procedure is computationally demanding. CS overcomes these inefficiencies by directly acquiring a compressed signal representation without going through the intermediate stages mentioned above [6]. Many studies have been carried out to explore applications of CS theorem [4, 6].

Wakin developed a one-pixel camera system based on CS [7]. His system can also be used to take streaming measurements of video sequences for video acquisition: frame-by-frame acquisition using 2-D random projections or joint acquisition that acquires 3-D random projections of the entire video sequence. But the computational complexity increases as the number of video frames increases. Zheng proposed a video compressive sensing method which relied on the sparsity properties of video in spatial domain [8]. But his scheme is only suitable for video sequences with relatively small temporal changes and the reconstruction error accumulates badly when a frame is far away from its reference frame.

To overcome these defects, we propose a compressive video sensing method based on UAM for real video acquisition. For every group of consecutive video frames, we set the first frame as reference frame that is fully sampled and build a UAM with VRA to automatically determine ROI for non-reference frames that are sampled by CS. Because the determined ROI has significant temporal changes, we use different CS operators and reconstruction algorithms for ROI and non-ROI.

Our main contributions are as follows: Firstly, due to the block-by-block CS manner on the non-ROI, the algorithm has low complexities. Secondly, via our UAM, the ROI which has significant movement can be efficiently

---

* This work was supported by National Nature Science Foundation of China (60802028).
determined and the proposed model can be implemented in hardware with very low complexity for real time video acquisition. Thirdly, the non-ROI is reconstructed in a block-by-block manner by ADM considering the sparsity of frame difference and ROI is reconstructed in a whole manner considering TV method. Because of the different treatment of ROI and non-ROI, the quality of reconstructed video is better than that conducted by previous methods and the accumulated errors during reconstruction are reduced.

This paper is organized as follows: Section 2 gives an overview of CS from the mathematical point of view. Section 3 describes UAM determination algorithm by VRA. Section 4 proposes our compressive video sensing system from acquisition to reconstruction and experimental results using real video sequence are presented in section 5. Conclusion and future work are given in section 6.

2. COMPRESSIVE SENSING

In this section, we have a brief review of CS theory according to the notation in [4-6]. CS is a novel sampling paradigm which carries data acquisition and compression simultaneously. Mathematically, the compressive sensing problem with incomplete measurements can be expressed by

\[ y = \Phi x, \]

\[ (2.1) \]

where \( \Phi \epsilon R^{M \times N} (M \ll N) \) is a given CS measurement matrix that has much fewer rows than columns. Let \( M \) denote the number of measurements and \( N \) denote the dimension of the signal \( x \). The fundamental of CS is to recover a signal \( x \) from a small set of measurements \( y \). But it seems to be hopeless to solve the ill-conditioned underdetermined linear system because the number of equations is much smaller than the number of unknown variables. However, it has been shown in [1, 5] that if \( x \) expressed as

\[ x = \Psi s \]

\[ (2.2) \]

is \( K \)-sparse in the domain \( \Psi \) for \( K \ll N \) and the measurement matrix satisfies the so-called restricted isometry property (RIP) which states that \( \Phi \) is not correlated with \( \Psi \), \( x \) can be reconstructed with high accuracy from the incomplete measurements \( y \). By substituting \( \Psi \) from (2.2), \( y \) can be written as

\[ y = \Phi x = \Phi \Psi s = \Theta s \]

\[ (2.3) \]

where \( \Theta = \Phi \Psi \) is an \( M \times N \) matrix. Optimization based on the \( l_1 \)-norm

\[ \hat{s} = \arg \min \| s^* \| \text{ such that } \Theta s^* = y \]

\[ (2.4) \]

can exactly recover \( K \)-sparse signals with high probability using only \( M \geq cK \log(N/K) \) iid Gaussian measurements.

Random matrices are often used as measurement matrices because they are incoherent to most transforms and physically realizable. Generally, one can take an orthogonal transform followed by randomly downsampling to build the measurement matrix. In this paper, we build measurement matrices based on noiselet [9] basis which is particularly incoherent to the Harr basis.

ADM is a new development of \( l_1 \)-norm minimization solver and has been successfully applied to variance signal reconstruction applications [10]. We use this method to reconstruct frame difference of non-ROI which is very sparse in spatial domain.

3. UAM THROUGH VISUAL RHYTHM ANALYSIS

The frame difference of ROI is not sparse in spatial domain. To choose the best sparse domain for ROI reconstruction, we did extensive CS reconstruction experiments considering various sparse domains including DCT, block-DCT, DWT and TV. As can be seen from one of the test results in Figure 1, the CS reconstructed real image via TV minimization has the best quality measured by PSNR.

Figure 1. (a) 256×256 Size image (b) PSNR of reconstructed image using different numbers of measurements and considering different sparse domain

The frame difference of most videos is sparse, but it is no longer sparse for some videos which contain moving object or significantly temporal changing area. CS reconstruction algorithms based on frame difference sparsity are not effective for these videos. So it is desirable to efficiently identify ROI which usually has significant movement. Our UAM for ROI is based on Ming-Chieh’s work in [11].

The visual rhythm is an abstraction of a video. It uses four sampling line, including diagonal, anti-diagonal, vertical and horizontal lines, to capture the temporal information of a video sequence which can be used to extract moving object. The definition of visual rhythm [11] derived from the diagonal sampling line is

\[ D_j = \begin{bmatrix}
    P_1(0, [0]) \\
    P_1(1, [1 \times r_y]) \\
    \vdots \\
    P_1((n-2), [(n-2) \times r_y]) \\
    P_1((n-1), [(n-1) \times r_y])
\end{bmatrix}, \]

\[ (3.1) \]

where \( [x] \) is a rounding operation. \( r_y \) and \( r_x \) are defined as
\[ r_d = -r_p = \frac{m-1}{n-1} \quad (3.2) \]

where \( m \) and \( n \) are width and height of a video frame. Promising results are shown in Figure 2.

**Figure 2. ROIs of some video sequences**

Because the ROI changes frame by frame, we have to design ROI measurement matrix for each frame. This will lead to producing huge overhead. In this paper, we first determine the center of ROI denoted as \((x_c, y_c)\) through VRA. Then we expand the same number of blocks centered at \((x_c, y_c)\) for each non-reference frame. So we just need to design one noiselet measurement matrix for all of ROIs in a video sequence.

### 4. COMPRESSIVE VIDEO SENSING BASED ON UAM

![Diagram of the system](image)

**Figure 3. The framework of our system**

The framework of our compressive video sensing system is shown in Figure 3.

Firstly, we divide a video sequence into several groups. Each group contains 15 consecutive frames. The first frame is always set as reference frame. The reference frame is full sampled. Subsequent frames are considered to be non-reference frames. Each frame is divided into non-overlapping 16x16 blocks.

Secondly, we use the method presented in section 4 to extract ROI in each non-reference frame. After that, each block can be identified as ROI or non-ROI.

Thirdly, we use the same measurement operator to conduct CS for non-ROI blocks. The non-ROI CS operation is:

\[ \mathbf{R}_{i,j} = \mathbf{F}_{\text{non-ROI}} (f_{i,j} - f_{i-1,j}) = \mathbf{F}_{\text{non-ROI}} f_{i,j} - \mathbf{F}_{\text{non-ROI}} f_{i-1,j} \quad (4.1) \]

where \( i \) is the frame number, \( j \) is the non-ROI block index. As can be seen from (4.1), subtracting measurement results is equivalent to making the frame difference measurement, so we just need to save or transmit the measurement difference of non-ROI blocks.

The ROI CS operation is:

\[ \mathbf{R}_{i,ROI} = \mathbf{F}_{ROI} f_{i,ROI} \quad (4.2) \]

where \( i \) is the frame number.

In the reconstruction process, we use TV method to reconstruct ROI directly. To reconstruct non-ROI blocks, we use ADM to reconstruct sparse frame differences of non-ROI blocks first and then add them to the reference frame sequentially.

### 5. EXPERIMENTAL RESULTS

![Graphs of PSNR and SSIM](image)

**Figure 4. Objective evaluation of reconstructed video frames**

We apply our system described in section 4 to the Y-components of some standard Quarter Common Intermediate Format (QCIF) sequences. We use Silent and News [12] to illustrate our experimental results. Since both of them have relatively large background and significant local movement.

ROI corresponding to each video sequence is shown in Figure 2. We use 30% compressive sampling rate for both ROI and non-ROI. The structural similarity index (SSIM) and the peak signal-to-noise ratio (PSNR) are used for reconstructed video quality assessment. As can be seen from Figure 4, our compressive video sensing system with UAM has better performance than the system just using frame difference sparsity without UAM in both SSIM and PSNR. Figure 5 demonstrates that our method can not only enhance the visual quality of reconstructed video frames but also correctly reconstruct the object which has significant movement such as the dancer in News.
6. CONCLUSION AND FUTURE WORK

In this paper, we present a compressive video sensing scheme based on UAM which can overcome poor quality of reconstructed video frames due to temporal changes. And we apply VRA to identify ROI that has significant temporal changes. In the video acquisition process, we use noiselet to construct measurement matrix. In the video reconstruction process, we use a new development of $l_1$–norm minimization solver called ADM for non-ROI video blocks and TV for ROI. The experiments demonstrate that our method presented in this paper can enhance reconstructed video quality and reduce errors accumulated in reconstruction process to some extent.

When a video contains a larger moving object, we have to increase the sampling rate and the size of ROI to ensure better reconstruction visual quality. But this will increase the quality of data needed to be saved or transmitted. If the size of ROI is large enough, the corresponding frame can be set to reference frame. Therefore adaptive shot-cut detection can help to improve reconstructed video quality. In addition, we will implement compressive video sensing in other two color channels and study how to incorporate efficient motion estimation into our system in the future work.

Acknowledgments. J.Ma would like to thank Junfeng Yang for useful discussions and for providing the ADM code.

7. REFERENCES