Distributed Multi-View Video Coding Based on Compressive Sensing in WMSN

Luo Hui, Qi Meili, Jiang Nan

1, School of Information Engineering, East China Jiaotong Uni, lh_jxnc@163.com
2, School of Information Engineering, East China Jiaotong Uni, miliqi@sina.cn
3, School of Information Engineering, East China Jiaotong Uni, jiangnan1018@gmail.com

Abstract

In wireless multimedia sensor networks (WMSN), many video sensors distributed in the same scene acquire large quantities of data. However, sensor devices are constrained in terms of power, processing, and bandwidth capacity. This paper presents a new framework of distributed video processing system for WMSN that takes advantage of distributed compressive sensing (DCS) theory. In the encoder analysis the residual joint sparse model. In the decoder, efficient reconstruction is implemented using the GPSR algorithm to recover images. Side information is generated using the temporal correlation between adjacent frames from a single view and the spatial correlation of the two nearest views. Experimental results show that the method presented achieves 1-3dB improvement in PSNR compared to the coding method of single-view videos.

Keywords: Distributed Compressive Sensing; Multi-View; WMSN; Disparity Compensation

1. Introduction

In wireless multimedia sensor networks (WMSN), a number of redundantly distributed nodes acquire data from different perspectives simultaneously, and report them to a central collection point which reconstructs the multi-view images. Communication energy and bandwidth of WMSN are often scarce resources, which imposes important constraints in terms of communication costs or bit rate. Multi-view video data are much larger than single-view video data, so there is an urgent need to identify a simple and high compression efficiency method of multi-view video coding.

The complexity of the encoder is 5 to 10 times that of the decoder in typical video compression standards such as MPEG or H. 264. This cannot meet the WMSN node’s low-cost, low power requirements[1]. The distributed video coding (DVC) theory, based on the Slepian-Wolf and Wyner-Ziv coding, has recently raised many interesting research questions regarding the efficient representation of information in correlated signals captured independently by multi-view sensors. The DVC encoder complexity is low, and can reduce the amount of communication between nodes. However, the DVC encoder is still using the Nyquist sampling theorem and a general transform coding method[2]. The calculation of the transform coding is computation-intensive. In 2006, D. Donoh[3,4] proposed a compressed sensing (CS) theory, in which the sampling number is far less than the Nyquist sampling, and the encoding side can reduce the compression load and increase the coding efficiency. At the same time, the CS perception observation method can reduce the computation intensity of the transform coding, while the CS is typically used for single-channel signal encoding. Distributed compressed sensing (DCS)[5] provides a new method for multi-channel signals coding. In order for the decoder to effectively reconstruct the images of multiple views, the DCS algorithm needs to evaluate the correlation between the video images. Most DCS algorithms rely on sparse representation to take advantage of the correlations between multi-view images. A series of studies on the joint sparse model[6] and the training dictionary of sparse representation[7] were carried out. These methods use the correlation between signals and reduce the number of observations of the encoding side, but the reconstruction of the image is not ideal.

In this paper, we propose a system for coding the multi-view video in WMSN using a distributed compressed sensing algorithm. Based on joint sparse representation, the decoder further exploits the temporal correlation between multi-view images and the spatial correlation between the video sequence in one sensor. In order to preserve good image quality, the system uses disparity estimation and disparity compensation between multi-view images for more precise side information and to optimize...
the reconstruction quality.

2. Related Works

2.1. Multi-view distributed video coding

DVC is based on the Wyner-Ziv theory. The video sequence is generally divided into a Group Of Pictures (GOPs). The first frame of each GOP, also referred to as a keyframe, is encoded using a conventional intra-frame coding technique such as AVC/H.264 in intra-frame mode. The remaining frames in a GOP are encoded using distributed coding principles, and are referred to as WZ frames. Current DVC techniques are divided into pixel domain and transform domain. In a pixel-domain WZ version, the WZ frames first undergo quantization. Alternatively, in a transform-domain version, a transform is applied prior to quantization. Transform domain DVC methods gain higher compression efficiency by taking advantage of the correlation of adjacent pixels.

With its ability to exploit inter-camera correlation at the decoder side, without communication between cameras, DVC is also well suited for multi-view video coding where it could offer a noteworthy architectural advantage. Multi-view video DVC coding can theoretically achieve the same bit rate and joint decoding. The architecture is shown in Figure 1. The main advantage of distributed video coding of multi-view video is that it can meet the requirements for low-power video sensor network nodes and low complexity. However, the emergence of compressed sensing provides a new solution for multi-view video coding, as it is more suitable for the communications and limited computing power of WMSN.

![Figure 1. The process of distributed video coding](Image)

2.2. Video coding based on compressed sensing

The recently emerged theory of CS provides an alternative management scheme to code video data. Consider a length-L, real-valued signal x of any dimension indexed as \( x(n) \), \( n \in \{1, 2, \ldots, L\} \). Suppose that the basis \( \Psi = \{\psi_1, \psi_2, \ldots, \psi_L\} \) provides a K-sparse representation of x; that is, x can be represented as \( x = \Psi \theta \) and \( \theta \) with length L can be approximated well using only \( K << L \) non-zero entries\(^3\). CS states that x can be accurately reconstructed by taking only:

\[
M = O( K \log(L/K) )
\]

where \( K < M << L \), linear and non-adaptive measurements from:

\[
b = \Phi X = \Phi \Psi \theta
\]

where b is an \( M \times 1 \) vector, \( \Phi \) is an \( M \times L \) measurement matrix that is incoherent with \( \Psi \), and \( D = \Phi \Psi \). More specifically, the M measurements in b are random linear combinations of the entries of \( \theta \), which can be viewed as the compressed and encrypted version of X. To reconstruct \( \theta \) from b, CS is based on either solving the convex optimization problem, or using some iterative computation-intensive algorithms. Finally, x can be reconstructed via \( \hat{x} = \Psi \hat{\theta} \), where \( \hat{\theta} \) is the reconstructed \( \theta \).

Distributed compressed sensing (DCS) theory is proposed based on CS and DVC. The DCS
framework provides an interesting new paradigm for multi-view video coding; the architecture is shown in Figure 2. DCS is to simplify the encoding of correlated signals, directly collecting the random measurements using CS hardware as shown in function (2). After the randomized encoding, the measurement vectors b1, b2, ..., bL-1, bL are then transmitted to a central node for joint decoding.

Figure 2. The process of video coding based on distributed compressed sensing

2.3. Joint Sparse model

A collection of joint sparsity (JS) models\[^9\] is incorporated in the DCS video coding framework. In a JS we represent each signal \( x_j \in \mathbb{R}^L \) in terms of a decomposition \( x_j = Z_c + Z_j \), in which \( Z_c \in \mathbb{R}^L \) is a common component that is assumed to be present in all \( \{x_j\} \), and \( Z_j \in \mathbb{R}^L \) is a novel component that differs for each signal. Depending on the application, different sparsity assumptions may be imposed on \( Z_c \) and \( Z_j \). The data between multi-view videos have a strong correlation in WMSN, which can be used in a JS model for sparse representation. For signal ensembles that are well-modeled by a JS, DCS reconstruction can offer a significant savings in the sampling rates. While each sensor must take enough samples to account for its novel component \( Z_j \), all sensors can share the burden of measuring the common component \( Z_c \).

2.4. Gradient projection for sparse reconstruction (GPSR)

In a variant of the CS reconstruction algorithm, the gradient projection method GPSR achieves higher computing speed and good reconstruction results, expressed as:

\[
\min_\theta \frac{1}{2} \left\| b - D \theta \right\|_2^2 + \tau \left\| \theta \right\|_1,
\]

where \( b \) is the received observation vector, \( b = \Phi x; D = \Phi \Psi; x = \Psi \theta; \left\| . \right\|_2 \) is L2 norm; \( \left\| . \right\|_1 \) is L1 norm; and \( \tau \) a non-negative balance parameter. The first iterative reconstruction of GPSR restores value \( \tilde{\theta} \) for \( \theta \), and then reconstructs the original signal \( x \) by \( \tilde{x} = \Psi \tilde{\theta} \). In general, in the GPSR algorithm the initial condition of \( \theta \) is zero vector. The iterative algorithm is stopped when the relative change in the number of non-zero components in \( \theta \) is smaller than the iteration condition of \( T_\alpha \) (default \( T_\alpha = 0.01 \)). The traditional GPSR algorithm has some shortcomings, such as the signal being overly sparse can lead to excessive iterations, and the reconstruction quality is not ideal over a long reconstruction time.
3. Multi-view video coding based on DCS

3.1. The main idea of multi-view video coding based on DCS

The basic idea of the coding method is to use a reference frame, \( W_{re} \), to lower the observation rate. \( W_{re} \) can be used in the encoder and decoder in order to ensure the low complexity encoding characteristics of the system. \( W_{re} \) makes available a simple estimate of the current frame, such as simple interpolation or extrapolation to generate decoded key frames. A reference frame can reduce the observed rate because the encoder uses \( W_{re} \) to generate residual frame \( x \), where \( x = W - W_{re} \). In the same sparse base, due to the stronger sparsity of \( x \) than the original frame \( W \), fewer observations will be required to accurately reconstruct the image \([10,11]\).

This section presents an effective multi-view video coding model by using residual\([11]\) and DCS theory to reduce the amount of data according to the framework shown in Figure 3. For keyframe \( K \) using a conventional intra-frame encode and decode technique, get the reconstruction frame \( K' \). The method of coding non-keyframe \( W \) will be introduced later.

![Figure 3. Architecture of the proposed coding method](image)

3.2. DCS encoder

The process of encoding non-keyframe:
(a) Generate a reference frame \( W_{re} \) using average interpolation method after recovering the key frame \( K' \);
(b) Generate a residual frame, \( X = W - W_{re} \);
(c) Derive the sparse representation coefficients from the residual frame \( X \) using the sparse representation based JS model;
(d) Derive \( b \) (the measurement values) using the measurement matrix.

This assumes the existence of the video sequence of the L perspective, \( x_L \) on behalf of L-view video residual frame, as defined by joint sparse model (JS)\([9]\):

\[
X_i = X_i + Z_i \\
\vdots \\
X_L = X_L + Z_L
\]

For observation \( b_i \) of each video sensor, the random projection equation is defined as

\[
f : b_i = \Phi_i X_i
\]
Paper [9] defines a random matrix \( A \in \mathbb{R}^{d \times D} (d < D) \) as a super-complete dictionary
\[
\begin{bmatrix}
    b_1 \\
    \vdots \\
    b_L
\end{bmatrix} =
\begin{bmatrix}
    A_1 & A_1 & 0 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \cdots & \vdots \\
    A_L & 0 & \cdots & 0 & A_L
\end{bmatrix}
\begin{bmatrix}
    X_c \\
    Z_1 \\
    \vdots \\
    Z_L
\end{bmatrix}
\]
\[\iff b' = A'X' \in \mathbb{R}^{d'} \quad \quad (6)\]
where the block Hadamard ensemble is selected as \( \Phi \). After completion of step (d), the observed value \( b \) is transmitted to the decoder.

3.3. DCS decoder and reconstruction algorithm analysis

The multi-view image reconstruction algorithm is analysed, taking advantage of the strong correlation between the multi-view images. Disparity estimation[12,13] and disparity compensation[12, 13] are incorporated as side information into the GPSR recovery algorithm. We exploit the strong correlations between multi-view images by reconstructing the residual frames between images and their predictions of disparity compensation as a means for refining the accuracy of GPSR reconstruction.

Assume the current image to be recovered is \( iX \), measurement matrix \( i\Phi \) is known, observations are \( b = \Phi_iX_i \), and sparse base is \( \Psi \). To adapt the GPSR algorithm to the multi-view scenario, assume that we know images adjacent to \( xi \); specifically, we know the closest images to the “left” and “right”, defined as \( xi-1, xi+1 \) respectively. The reconstruction method process is shown in Figure 4. The process of decoding is as follows:

(a) Generate the reference frame \( W_r \) and generate side information frame \( D \) by using the average interpolation method based on reconstructed key frame \( K' \), and generate a residual side information\[14\] frame \( R' = D - W_r \).

(b) Generate the current perspective of the residual image side information by bidirectionally interpolating the adjacent perspective views as \( X_i' = \text{ImageInterpolation}(\hat{x}_i, \hat{x}_i) \).

(c) Apply the linear combination of fusion method, which is proposed in paper [2], to fuse \( R' \) and \( X' \) to produce \( x' \).

(d) Acquire the residual \( r = b_i - b \), where \( b_i \) is the observation resulting from the projection of \( X_i' \) using the same measurement matrix \( \Phi \) of acquiring observation \( b \).

(e) Restore the original residual frame, using the GPSR reconstruction algorithm.

(f) Obtain the left and right disparity vectors \( V_i-1 \) and \( V_i+1 \) from disparity estimation (DE) applied to the current reconstruction \( \hat{x}_i \), and the left and right adjacent images of the current image.

(g) Produce the current prediction via disparity compensation (DC) driven by DE; the reconstructed residual \( \hat{r} \) is further refined with reverse DC.

(h) Repeat (e)(f)(g) many times.
Several iteration stopping criteria for the GPSR algorithm are proposed in paper [5], based on the correlation between the current frame and side information. Therefore, set $X_i'$ as the side information of frame $X_i$. With a few iterations, we can acquire accurate reconstruction of the images.

4. Simulation results

To validate the performance of the above method, this paper selects three-view video image sequences with GOP size=3. The method of generating side information in the reconstruction process is introduced in paper [1]. The DE algorithm [12,13] and the DC algorithm are used to improve the accuracy of the side information. The traditional single perspective DCVS algorithm was selected to compare with the multi-view video coding theory described in this paper. Non-key frames in the middle perspective of the reconstructed image are presented as the verification results, shown in Figure 5. Image (a) is the reconstruction image of the traditional single view DCVS algorithm, and image (b) is recovered by the multi-view DCS algorithm proposed in this paper. Subjectively speaking, the quality of image (a) is poor, and there is some distortion; the quality of image (b) is significantly better than (a).

We further compared the reconstruction accuracy of these two algorithms by analyzing the peak signal to noise ratio (PSNR). The graphed data are shown in Figure 6. The horizontal axis represents the measurement value per frame, the vertical axis represents the value of the peak signal to noise ratio; the solid line represents the traditional single view DVCS algorithm, and the dotted line represents the proposed coding method. In the case of small measurement values, the proposed method, improves the PSNR performance by 1-3dB over the single-view video coding algorithm, and the quality of the reconstructed image is good, as can be seen from Figure 5 (b).
Figure 5. Two subjective evaluations of the reconstruction algorithm

5. Conclusions

In this paper, we examined a method for the multi-view video coding of WMSN data using the distributed compressed sensing theory. The method uses the residual theory to eliminate the inter-frame redundancy of the video, greatly reducing the quantity of data transferred. The accuracy of the prediction of the current frame is improved via fusion of the temporal and spatial side information; the side information is made more accurate by using disparity estimation and disparity compensation between the multi-views to optimize the quality of the reconstruction. The simulation results demonstrate the effectiveness of the method in reconstructing high quality images, increasing the peak signal to noise ratio by 1-3dB.
6. Acknowledgement

This work is supported by the National Natural Science Foundation “Research on Video Image Processing Method Based on Compressed Sensing for Railway Foreign Invasion Monitoring in the Mountainous” (No.61261040).

7. References