sensors

MDPI

Article

A Subspace Approach to Sparse Sampling Based Data Gathering in Wireless Sensor Networks

Jingfei He *, Xiaoyue Zhang, Yatong Zhou and Miriam Maibvisira

Tianjin Key Laboratory of Electronic Materials and Devices, School of Electronics and Information Engineering, Hebei University of Technology, 5340 Xiping Road, Beichen District, Tianjin 300401, China; zxy_zhangxiaoyue@163.com (X.Z.); ytzhou_hebut@163.com (Y.Z.); miriammaib19@icloud.com (M.M.)

* Correspondence: hejingfei@hebut.edu.cn

Received: 31 December 2019; Accepted: 6 February 2020; Published: 12 February 2020

Abstract: Data gathering is an essential concern in Wireless Sensor Networks (WSNs). This paper proposes an efficient data gathering method in clustered WSNs based on sparse sampling to reduce energy consumption and prolong the network lifetime. For data gathering scheme, we propose a method that can collect sparse sampled data in each time slot with a fixed percent of nodes remaining in sleep mode. For data reconstruction, a subspace approach is proposed to enforce an explicit low-rank constraint for data reconstruction from sparse sampled data. Subspace representing spatial distributions of the WSNs data can be estimated from previous reconstructed data. Incorporating total variation constraint, the proposed reconstruction method reconstructs current time slot data efficiently. The results of experiments indicate that the proposed method can reduce the energy consumption and prolong the network lifetime with satisfying recovery accuracy.

Keywords: wireless sensor networks; data gathering; sparse sampling; subspace; data reconstruction

1. Introduction

Recently, wireless sensor networks (WSNs) have been applied in many fields, such as target tracking, medical care, and environmental monitoring. The component of a WSN includes a number of sensor nodes monitoring and collecting the physical environmental information and a sink aggregating the collected data. The data collection from each node to the sink, known as data gathering, is a significant research issue in WSNs.

In most real-world applications, the sensor nodes in WSNs are always limited by computational capacity and battery power. Lots of data gathering methods have been proposed to reduce energy consumptions in WSNs. These data gathering methods can be mainly classified into two categories: methods utilizing data compression techniques [1–3] and methods based on designing network protocols [4–6]. In the past decades, inspired by the emergence of compressed sensing (CS) [7] and matrix completion [8] theory in signal processing field, data gathering methods based on data compression techniques obtain more attention. It is important to note that the decrease in data transmission between nodes can effectively reduce energy consumption and prolong the network lifetime. Considering the redundant and highly correlated data sensed in neighboring sensors during consecutive times, data gathering methods taking advantage of data compression techniques have contributed to reducing the amount of data transmission and prolonging the network lifetime. Specifically, Compressive Data Gathering (CDG) was proposed [2] based on compressed sensing theory [7]. Instead of collecting each raw sensed data, the sink in CDG receives a weighted sum of all the readings from nodes. Compared with the traditional data gathering methods in WSNs, CDG can balance the energy consumption and prolong the lifetime of the network. Afterwards, many extensions to the CS based data gathering methods in WSNs have been developed [9–11] based on dense sensing

matrix. To further reduce the amount of transmission data, methods utilizing sparse sensing matrix to random sample the raw sensed data in WSNs were also proposed [12–14]. More recently, as the rank of matrix is interpreted as a measure of second-order sparsity, matrix completion method [8] has attracted the attention of researchers, and many data gathering and reconstruction methods in WSNs based on matrix completion were proposed [3,15–19]. By distributing data collected from different nodes in different time slots into an environment matrix, the matrix completion based methods enforce the low-rank constraint on the matrix to take advantage of the spatiotemporal correlation. In particular, Spatio-Temporal Compressive Data Collection (STCDG) [3] was proposed to reduce the amount of traffic and improve the level of recovery accuracy by enforcing the low-rank constraint based on matrix factorization approach. Furthermore, methods utilizing joint low-rank and sparsity constraints were also proposed to further improve the data recovery accuracy [16,19]. Considering the inherent correlation among multi-attribute data, this method [19] extends the low-rank constraint based on matrix to tensor model to further exploit the correlation.

Although these methods enforcing low-rank constraint achieve better data recovery performances than the CS based methods, the requirement to form an environment matrix limits the real-time data gathering and reconstruction in WSNs. Besides, the sparse sampling strategy is adopted in matrix completion based methods, resulting in the existence of several nodes remaining in sleep mode. These inactive nodes will cause the failure of real-time data gathering. To achieve real-time data gathering with matrix completion, this paper proposes a sparse data gathering method based on clustered WSNs. Sparse sampled data can be collected in each time slot even with a fixed percent of nodes remaining in sleep mode. For data reconstruction, a subspace approach is proposed to enforce an explicit low-rank constraint. With subspace representing temporal distributions of the WSNs data estimated from previous reconstructed data, data collected in current time slot can be recovered efficiently. To guarantee data recovery performance even with low sampling ratio, we incorporate total variation constraint to further improve data recovery accuracy.

The rest of this paper is organized as follows. Section 2 reviews related works on matrix completion. Section 3 describes the proposed sparse data gathering method and the subspace approach for data reconstruction. The experimental results and analysis are presented in Section 4. Finally, Section 5 concludes this paper.

2. Matrix Completion in WSNs

2.1. Matrix Completion

Matrix completion problem was proposed to recover a data matrix from its partially sampled entries, and it has been proved that most low-rank matrices can be perfectly recovered from an incomplete set of entries [8]. Specifically, recovering an incomplete matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ can be cast as a rank minimization problem:

$$\begin{array}{ll} \text{minimize} & \text{rank}\,(\mathbf{X}) \\ \text{subject to} & \mathbf{X}_{ij} = \mathbf{M}_{ij} \quad (i,j) \in \Pi, \end{array} \tag{1}$$

where $\mathbf{X}$ is the decision variable, Π the sampled subset of the complete set of entries $[n] \times [m]$ (Here and in the sequel, $[n]$ denotes the list $\{1, ..., n\}$), and rank $(\mathbf{X})$ represents the rank of the matrix $\mathbf{X}$. Essentially, the rank of matrix is treated as the measure of the second order sparsity. Therefore, the low-rank constraint can utilize the spatial and temporal correlation in data matrix. However, the optimization problem (1) is nonconvex and NP-hard. To address this issue, two popular approaches are utilized in numerous applications to enforce the low-rank constraint. One is the nuclear norm based approach, and the other is matrix factorization based approach. For the first one, since the nuclear norm is the best convex surrogate to the rank function over matrices with spectral norm less than or equal to one [20], the nuclear norm minimization is used as an alternative:

$$\begin{aligned} &\text{minimize} && \|\mathbf{X}\|_* \\ &\text{subject to} && \mathbf{X}_{ij} = \mathbf{M}_{ij} \quad (i,j) \in \Pi. \end{aligned} \tag{2}$$

Here, $\|\cdot\|_*$ denotes the nuclear norm, which is a convex function and equal to the sum of the singular values of the matrix. Then, (2) can be rewritten as a regularized unconstrained problem, and singular value thresholding algorithm [21] can be used for the resulting problem. It is worth noting that computing Singular Value Decomposition (SVD) in each iteration is the main computational cost.

For the second one, the incomplete matrix $\mathbf{M}$ with rank r can be expressed as the product of two matrices $\mathbf{LR}$, where $\mathbf{L} \in \mathbb{R}^{n\times r}$, $\mathbf{R} \in \mathbb{R}^{r\times m}$. The matrix factorization is not unique, the factorization where the matrices $\mathbf{L}$ and $\mathbf{R}$ have Frobenius norm as small as possible can be searched. Then (1) can be cast as:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\left(\|\mathbf{L}\|_F^2 + \|\mathbf{R}\|_F^2\right) \\ &\text{subject to} && (\mathbf{LR})_{ij} = \mathbf{M}_{ij} \quad (i,j) \in \Pi, \end{aligned} \tag{3}$$

The matrix factorization based approach has gained great popularity due to fewer requirements for storage capacity and computational overhead. It is worth noting that the problem (3) is a nonconvex quadratic program and can be solved by standard nonlinear optimization algorithms, such as alternating minimization method [22] and gradient descent method.

2.2. Matrix Completion Based Method in WSNs

Inspired by the great success of matrix completion, many data gathering and reconstruction methods in WSNs were proposed based on matrix completion. Considering a WSN consisting of n sensor nodes and one sink with symbol $N_1, N_2, \cdots, N_n$ used to represent sensor nodes. These sensor nodes sense the environment and transmit readings to the sink once every τ time. Therefore, an environment matrix (EM) $\mathbf{M}$ organized by $n \times m$ readings can be obtained during $m\tau$ time (i.e., m time slots):

$$\mathbf{M} = \begin{bmatrix} f(N_1,T_1) & \cdots & f(N_1,T_m) \\ \vdots & \ddots & \vdots \\ f(N_n,T_1) & \cdots & f(N_n,T_m) \end{bmatrix} \in \mathbf{R}^{n\times m}, \tag{4}$$

where $f\left(N_i,T_j\right)$ denotes the reading collected from node N_i in the jth time slot. In most matrix completion based methods, partially sampled readings sensed during m time slots are transmitted to the sink together. Since readings generated by the nodes in a certain area during consecutive times are redundant and highly correlated, the EM $\mathbf{M}$ has approximately low-rank structure. Then, the low-rank constraint can be enforced to exploit the correlation to reconstruct the unsampled readings. Methods based on matrix completion have achieved great success. However, the requirement to form an EM limits the real-time data gathering and reconstruction in WSNs.

3. The Proposed Method

In matrix completion based methods in WSNs, sparse sampling is adopted in data gathering to reduce the amount of transmission data and energy consumption. Typically, the sparse sampled readings sensed during a fixed number of time slots are transmitted to the sink together in the last time slot, which limits real-time data gathering and reconstruction. To implement real-time data gathering in WSNs, the sparse sampled data should be collected and transmitted to the sink in every time slot. However, since only partial sensor nodes wake up and collect data while other nodes remaining in sleep mode, the traditional data gathering methods fail due to the existence of inactive nodes. If the successful aggregation of sparse sampled data is guaranteed at the expense of all nodes waking up to transmit data, the goal of reducing energy consumption through sparse sampling can not be achieved. Moreover, for data reconstruction in real-time case, only data in current time slot need to be

reconstructed. In this paper, a sparse data gathering method based on clustered WSNs is proposed to implement real-time data gathering, and a subspace approach based method is proposed to reconstruct data in current time slot by enforcing an explicit low-rank constraint.

3.1. The Sparse Sampling Data Gathering Scheme

To ensure the success of sparse sampled data gathering without waking up all nodes, a sparse sampling data gathering scheme is proposed in clustered WSNs. In clustered WSNs, all nodes are divided into multiple clusters, and each cluster contains one cluster head (CH) and a number of cluster members (CMs). The cluster head is selected for each round to communicate with the sink, and the cluster members only communicate with CHs in their clusters. The proposed sparse sampling data gathering scheme based on clustered WSNs is shown in Figure 1. For clarity, nodes are divided into two layers: sensor layer containing CMs in each cluster and cluster head layer containing all CHs. Sparse sampling is adopted for CMs in each cluster, that is, only partial CMs wake up and transmit the sensed data to the CH. The CHs transmit the received data and its own sensed data to the sink. It is worth noting that several CHs which locate far from sink node can communicate with sink using multi-hop transmission. For simplicity, only direct communication between CHs and the sink is illustrated in Figure 1.

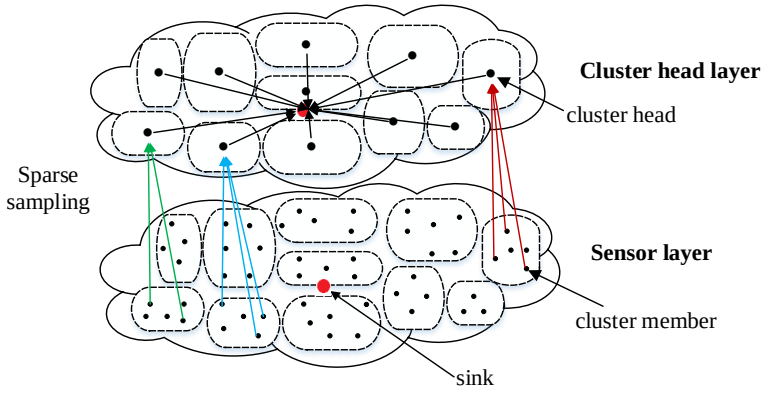


Figure 1. The sparse sampling data gathering scheme in wireless sensor networks (WSNs).

The proposed sparse sampling data gathering scheme can combine with many existing clustering algorithms. In this paper, the well known distributed energy-efficient clustering (DEEC) scheme [4] is selected as the clustering algorithm. In DEEC, cluster heads are selected according to a probability based on the ratio between the residual energy of each node and the average energy of the network. Because of adapting the rotating epoch of each node to its energy, the nodes with high initial and residual energy have a higher probability of becoming CHs than those with low energy. Specifically, the proposed sampling data gathering scheme using DEEC can be divided into rounds, and each round contains two phases as follows.

In the first phase, clusters are formed and the sampling ratio of each node is determined. Specifically, each node decides whether to become a CH based on its own probability threshold, which is related to the residual energy and estimated average energy of networks at current round. After the CHs are selected, the other nodes, cluster members, determine the dependent cluster according to the signal strength of the received information, and notify the corresponding CH to complete the establishment of the clusters. Then the sink broadcasts a fixed sampling ratio ρ to all sensor nodes in the network. Note that the CHs are always awake to ensure the successful aggregation of data. As a result, the sampling ratio for CHs can be set to 1. For the CMs, the sampling ratio is $\rho_{CH} = (\rho n - n_{CH})/(n - n_{CH})$, where n denotes the number of nodes still alive and n_{CH} the number of CHs. $n - n_{CH}$ represents the number of CMs.

In the second phase, sparse sampling and data gathering are held. Specifically, each node in CMs generates a random number between 0 and 1. If the number is less than the sampling ratio ρ_{CH}

then the node senses the environment and transmits readings to the corresponding CH, otherwise the node remains in sleep mode in current time slot. After receiving data from CMs in its cluster, the CH transmits the received data and its own sensed data to the sink. The network will start a new round after a fixed number of time slots.

Instead of using all nodes to collect the information, the proposed sampling data gathering scheme only wakes partial nodes up in each time slot, which greatly reduces the energy consumption and prolongs the lifetime of the network compared to original clustering algorithm.

3.2. The Sliding Window Model

In real-time data gathering with sparse sampling, data from CHs and partial CMs are collected in the sink for current time slot. We use T_c to represent the current time slot, and let T_{c-j} denote the last $(j+1)$th time slot. Figure 2 shows the data sampled from n sensor nodes between time slots T_{c-w} to T_c. To implement the real-time data reconstruction, the sliding window model is utilized by introducing the reconstructed data in previous time slots. As illustrated in Figure 2, two adjacent windows are shown with the width fixed to w time slots. The first window, the previous window of the current window, contains time slots from T_{c-w} to T_{c-1} , and the second window, the current window,contains time slots from T_{c-w+1} to T_c. Data in each window can form an environment matrix, and the window slides forward a time slot each time. That is, the oldest time slot will be deleted, while the current time slot enters the window. Meanwhile, the environment matrix is updated accordingly.

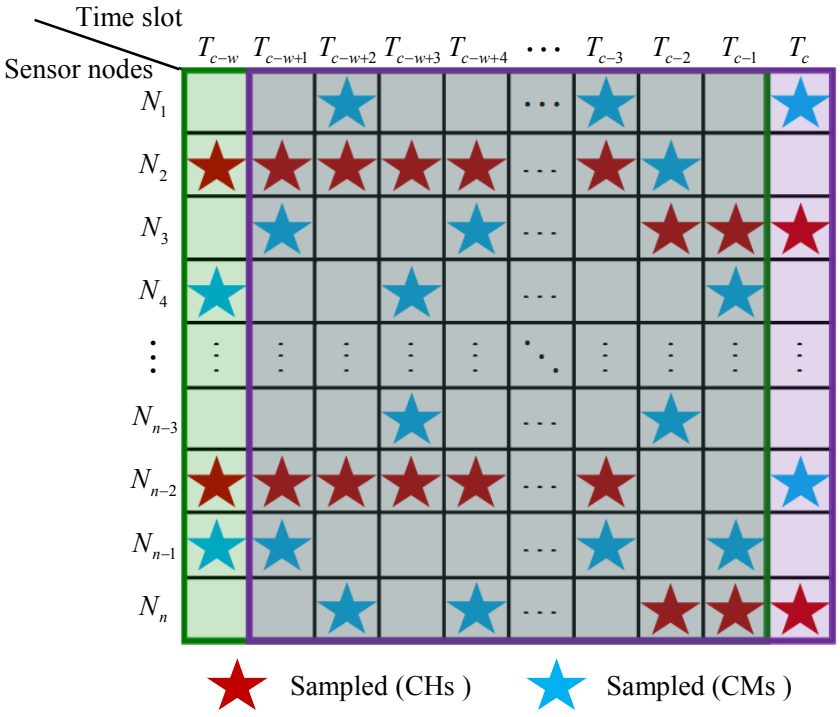


Figure 2. Sampled data using sliding window model.

For example, to reconstruct data $\mathbf{x}_c$ in current time slot T_c, the previous window containing time slots T_{c-w} to T_{c-1} slides forward a time slot. Then the current window including data from T_{c-w+1} to T_c is obtained. Divide the correspondingly current EM $[\mathbf{M}_p\ \mathbf{x}_c]$ into two parts: the previous reconstruction data $\mathbf{M}_p$ (data in time slots T_{c-w+1} to T_{c-1}) and the current data $\mathbf{x}_c$ (collected in current time slot). Since each column in $\mathbf{M}_p$ has been reconstructed in the corresponding time slot, the proposed subspace approach can enforce an explicit low-rank constraint to reconstruct the current data $\mathbf{x}_c$ in real time.

It is worth noting that the sink needs the indexes of the sparse sampled data to reconstruct the current data. To avoid overhead incurred by index information, the preset random seed can be used. At the beginning of each round, each CM generates a random seed as a pseudo-random number

generator and sends it to the CHs and sink. The generated random number determines whether the CM wakes up and collects data. With all the information of random seeds, the sink can obtain the indexes of the sparse sampled data without extra overhead.

3.3. The Subspace Approach for Reconstruction

With the proposed sparse sampling data gathering scheme, the sink can obtain the current sparse sampled data $\mathbf{d} = \Omega(\mathbf{x}_c)$, where $\Omega : \mathbb{R}^{n\times 1} \to \mathbb{R}^{s\times 1}$ is the sparse sampling operator to obtain the partially known data from the current data. According to the sliding window model, the current EM can be expressed as $[\mathbf{M}_p\ \mathbf{x}_c]$. Readings in EM are collected from nodes in a certain area during a consecutive time, then EM has approximately low-rank structure. Using the matrix factorization based approach, we have $[\mathbf{M}_p\ \mathbf{x}_c] = \mathbf{LR}$. Here, $\mathbf{L} \in \mathbb{R}^{n\times r}$ and $\mathbf{R} \in \mathbb{R}^{r\times w}$ can be regarded as the subspace representing the spatial distributions and the temporal distributions of the WSNs data, respectively. With $\mathbf{R}$ rewritten as $[\mathbf{R}_p\ \mathbf{r}_c]$, where $\mathbf{R}_p \in \mathbb{R}^{r\times(w-1)}$ and $\mathbf{r}_c \in \mathbb{R}^{r\times 1}$, we have $[\mathbf{M}_p\ \mathbf{x}_c] = \mathbf{L}[\mathbf{R}_p\ \mathbf{r}_c] = [\mathbf{LR}_p\ \mathbf{Lr}_c]$. That is, $\mathbf{M}_p = \mathbf{LR}_p$ and $\mathbf{x}_c = \mathbf{Lr}_c$. Since the subspace $\mathbf{L}$ representing the spatial distributions can be estimated from previous reconstructed data $\mathbf{M}_p$, the current data can be reconstructed by

$$\begin{aligned} \hat{\mathbf{r}}_c &= \arg\min_{\mathbf{r}_c} \quad \|\mathbf{d} - \Omega(\mathbf{Lr}_c)\|_2^2 \\ \hat{\mathbf{x}}_c &= \mathbf{Lr}_c. \end{aligned} \tag{5}$$

To further improve the recovery accuracy, the total variation constraint $\|\nabla_x([\mathbf{M}_p\ \mathbf{x}_c])\|_1$ can be jointly utilized to enforce the temporal stability, where ∇_x represents the horizontal finite difference operator. It is worth noting that the vertical finite difference operator is not used. Since adjacent nodes in the matrix may not adjacent in space, constraining the vertical finite difference can not utilize the spatial stability. Besides, $\mathbf{M}_p$ is known by previous reconstruction, then the constraint $\|\nabla_x([\mathbf{M}_p\ \mathbf{x}_c])\|_1$ can be simplified as $\|\nabla_x([\mathbf{m}_p\ \mathbf{x}_c])\|_1$, where $\mathbf{m}_p$ is the last column of $\mathbf{M}_p$. By jointly enforcing the constraint and introducing a quadratic penalty term, (5) can be converted into a corresponding unconstrained formulation as:

$$\begin{aligned} \hat{\mathbf{r}}_c &= \arg\min_{\mathbf{r}_c} \|\mathbf{d} - \Omega(\mathbf{Lr}_c)\|_2^2 + \lambda\|\nabla_x([\mathbf{m}_p\ \mathbf{Lr}_c])\|_1 \\ \hat{\mathbf{x}}_c &= \mathbf{Lr}_c. \end{aligned} \tag{6}$$

where λ denotes the regularization parameter. The proposed subspace approach incorporates both the explicit low-rank constraint and the temporal stability constraint in a single formulation (i.e., (6)). To solve the optimization problem in (6), an efficient algorithm based on alternating direction method of multipliers (ADMM) [23,24] is developed. First, using variable splitting, we can convert (6) into the following equivalent constrained optimization problem:

$$\begin{aligned} \{\hat{\mathbf{r}}_c, \mathbf{y}\} &= \arg\min_{\hat{\mathbf{r}}_c, \mathbf{y}} \|\mathbf{d} - \Omega(\mathbf{Lr}_c)\|_2^2 + \lambda\|\mathbf{y}\|_1 \\ \text{s.t.} \quad & \mathbf{y} = \nabla_x([\mathbf{m}_p\ \mathbf{Lr}_c]). \end{aligned} \tag{7}$$

Second, the augmented Lagrangian function for (7) can be obtained:

$$\mathcal{L}(\mathbf{r}_c, \mathbf{y}, \mathbf{a}) = \|\mathbf{d} - \Omega(\mathbf{Lr}_c)\|_2^2 + \lambda\|\mathbf{y}\|_1 + \langle \mathbf{a}, \mathbf{y} - \nabla_x([\mathbf{m}_p\ \mathbf{Lr}_c])\rangle + \frac{\alpha}{2}\left\|\mathbf{y} - \nabla_x([\mathbf{m}_p\ \mathbf{Lr}_c])\right\|_2^2, \tag{8}$$

where $\mathbf{a}$ is the Lagrangian multiplier, and $\alpha > 0$ is the penalty parameters. Third, (8) can be minimized alternatively as following:

$$\mathbf{r}_c^{k+1} = \arg\min_{\mathbf{r}_c} \mathcal{L}(\mathbf{r}_c, \mathbf{y}^k, \mathbf{a}^k) = \arg\min_{\mathbf{r}_c} \|\mathbf{d} - \Omega(\mathbf{L}\mathbf{r}_c)\|_2^2 + \frac{\alpha}{2} \left\| \mathbf{y}^k - \nabla_x([\mathbf{m}_p\ \mathbf{L}\mathbf{r}_c]) + \frac{\mathbf{a}^k}{\alpha} \right\|_2^2 \tag{9}$$

$$\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}} \mathcal{L}(\mathbf{r}_c^{k+1}, \mathbf{y}, \mathbf{a}^k) = \arg\min_{\mathbf{y}} \|\mathbf{y}\|_1 + \frac{\alpha}{2\lambda} \left\| \mathbf{y} - \left(\nabla_x([\mathbf{m}_p\ \mathbf{L}\mathbf{r}_c^{k+1}]) - \frac{\mathbf{a}^k}{\alpha} \right) \right\|_2^2 \tag{10}$$

$$\mathbf{a}^{k+1} = \mathbf{a}^k + \alpha \left(\mathbf{y}^{k+1} - \nabla_x([\mathbf{m}_p\ \mathbf{L}\mathbf{r}_c^{k+1}]) \right). \tag{11}$$

To solve the subproblem in (9), which is a quadratic optimization problem, the preconditioned conjugate gradient (PCG) algorithm is applied in this paper. To solve the subproblem in (10), the well-known soft-thresholding formula [25] can be utilized. Then we have

$$\mathbf{y}^{k+1} = \mathcal{S}\left(\nabla_x([\mathbf{m}_p\ \mathbf{L}\mathbf{r}_c^{k+1}]) - \frac{\mathbf{a}^k}{\alpha}, \frac{\lambda}{\alpha} \right), \tag{12}$$

where $\mathcal{S}(\mathbf{q}, \tau)_i := \text{sign}(\mathbf{q}_i) \max(|\mathbf{q}_i| - \tau, 0)$ is a soft-thresholding operator for each element $\mathbf{q}_i$ in $\mathbf{q}$. The procedures of the reconstruction algorithm based on ADMM for solving (6) can be summarized in Table 1. In practical implementation, we initialize $\mathbf{r}_c^0$, $\mathbf{y}^0$, and $\mathbf{a}^0$ with zeros vectors. Since (6) is a convex optimization problem, the ADMM based method is guaranteed to have global convergence from any initializations [24]. The stopping criteria for the algorithm are $\|\mathbf{r}_c^k - \mathbf{r}_c^{k-1}\|_2 / \|\mathbf{r}_c^{k-1}\|_2 \leq \epsilon$ and $k > K_{max}$, where ϵ and K_{max} are the predefined tolerance parameter and the maximum number of iterations, respectively. The algorithm is terminated until each criterion is satisfied.

Table 1. The reconstruction algorithm to solve the optimization problem in (6).

Input:
Initialized $\mathbf{r}_c^0$,$\mathbf{y}^0$,and $\mathbf{a}^0$ with zeros vectors;
The subspace $\mathbf{L}$ representing the spatial distributions;
The regularization parameter λ and the penalty parameter α;
The measurements $\mathbf{d}$, the sampling operation Ω, and the iteration number $k = 0$;
do
1)Iteration number $k = k + 1$;
2)Update $\mathbf{r}_c^k$ by solving (9);
3)Update $\mathbf{y}^k$ by solving (12);
4)Update $\mathbf{a}^k$ by solving (11);
while $\frac{\|\mathbf{r}_c^k - \mathbf{r}_c^{k-1}\|_2}{\|\mathbf{r}_c^{k-1}\|_2} > \varepsilon$ **and** $k \leq K_{max}$;
Output: $\hat{\mathbf{x}}_c = \mathbf{L}\mathbf{r}_c^k$

4. Experiments and Analysis

In this section, the experimental environments are established to verify the effectiveness of the proposed method. For the reconstruction experiments, the performance of the subspace based reconstruction method was compared with local interpolation method K-Nearest Neighbors (KNN) [26], the CS based method [13], Seq-Prog-CS method [10], and joint CS and matrix completion method (CSMC) [16], and the real dataset collected from GreenOrbs was adopted. For the energy consumption experiments, the proposed sparse sampling data gathering scheme using DEEC was compared with the original DEEC to verify the benefit of the proposed method in reducing the energy consumption. All experiments were conducted using MATLAB R2017a on a computer with 1.8GHz Intel core i7-8550U CPU and 8GB RAM.

4.1. Data Reconstruction Performance

To verify the effectiveness of the subspace approach for reconstruction, a small real dataset collected from GreenOrbs [27] was selected as the ground truth. The readings in the small real dataset were collected from 94 nodes in 124 time slots for three attributes: temperature, humidity, and light. Then we verified the effectiveness of the proposed method from two aspects. One is the recovery of data in current time slot. Another one is the long-term recovery of data in consecutive time slots.

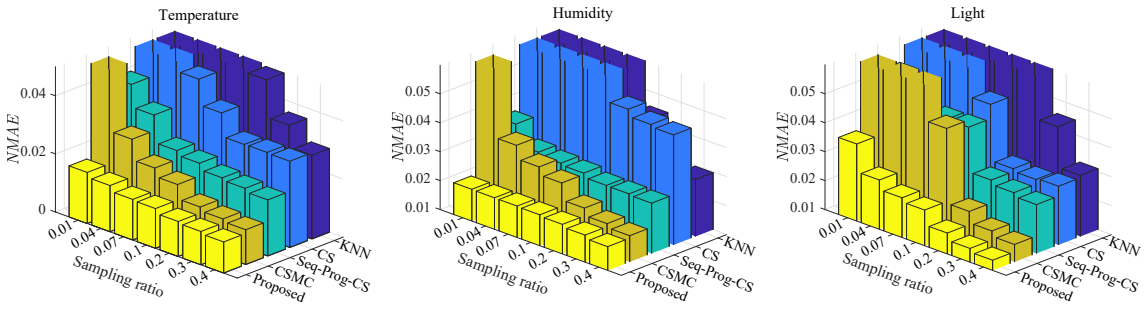
For the first aspect, the 50th time slot in the dataset was selected as the current time slot, and the width of sliding window was set to $w = 41$. $\mathbf{x}_c \in \mathbb{R}^{n\times 1}$ denote the data in current time slot, and $\mathbf{M}_p \in \mathbb{R}^{n\times 40}$ denotes the previous reconstructed data. Here, $n = 94$ is the number of nodes. With predefined sampling ratio ρ, only partial readings $\mathbf{d} \in \mathbb{R}^{\lfloor n\rho \rfloor \times 1}$ can be sampled using the proposed sparse sampling data gathering scheme. Mathematically, the sampling procedure can be expressed as $\mathbf{d} = \Omega(\mathbf{x}_c)$. The proposed method and compared methods were used to reconstruct $\hat{\mathbf{x}}_c$ from $\mathbf{d}$. The Normalized Mean Absolute Error (NMAE) was adopted to measure the recovery performance and defined as:

$$NMAE = \frac{\sum_{i\in\Pi} |\mathbf{x}_{c_i} - \hat{\mathbf{x}}_{c_i}|}{\sum_{i\in\Pi} |\mathbf{x}_{c_i}|}. \tag{13}$$

Here, $\hat{\mathbf{x}}_c$ is the recovered data, and Π represents the unsampled subset of the complete set of entries $[n] \times 1$. That is, we only calculated the recovery accuracy of the unsampled data.

In the simulation, the spatial distributions $\mathbf{L}$ need to be estimated for the proposed subspace approach, which need $\mathbf{M}_p$ and the rank r. For simplicity, the historical data collected from 10th to 49th time slots were used as $\mathbf{M}_p$, and the rank r was set to 15 according to singular values of $\mathbf{M}_p$. Besides, to make the comparison more convincing, the same historical data were also used in CSMC and Seq-Prog-CS method. The sampling ratio ρ was set as $[0.4, 0.3, 0.2, 0.1, 0.07, 0.04, 0.01]$, and the sampling operator Ω was generated with a uniform random sampling pattern. Each experiment was repeated 100 times to calculate the mean NMAE value for each method. For each experiment, the same sampling operator and sampled data were used for all the methods.

Figure 3 shows the recovery performance of methods for the temperature, humidity, and light data in GreenOrbs, respectively.



(a) Temperature (b) Humidity (c) Light

Figure 3. The recovery performance of KNN, CS, Seq-Prog-CS, CSMC, and the proposed method for three attributes in GreenOrbs : (**a**) temperature, (**b**) humidity, and (**c**) light.

As shown in Figure 3, the proposed method achieves the lowest NMAE with each sampling ratio for all attributes. With the decrease of sampling ratio, the performance of other three methods degrades dramatically while the proposed method keeps satisfying performance. Even with the sampling ratio as low as 0.01, the proposed method obtains NMAE $= 0.018, 0.021, 0.037$ for temperature, humidity, and light data, respectively.

Figure 4 shows the running time performance of methods for temperature data in GreenOrbs. The average running time of KNN, CS, CSMC, and the proposed method is graphically represented. As shown in Figure 4, the proposed method achieves the lowest running time. The reason is that

the proposed method avoids the singular value decomposition (SVD) in (6). For traditional matrix completion methods, the main computational complexity is from the SVD of EM for each iteration, and exact SVD of a matrix with size $n \times m$ has time complexity $O(\min\{nm^2, n^2m\})$. In the proposed subspace approach, the SVD is used to obtain the subspace $\mathbf{L}$ from previous reconstructed data $\mathbf{M}_p$. The calculation only needs to be run once and can be done between two time slots. Then the computational complexity in each iteration to solve (6) is negligible.

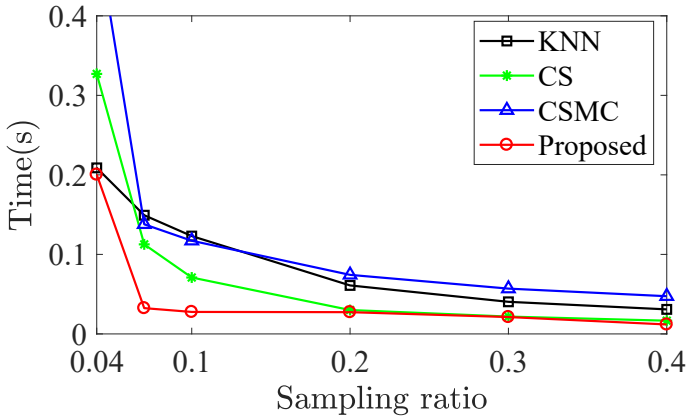


Figure 4. Running time over sampling ratio for temperature data in GreenOrbs.

The historical data were used as $\mathbf{M}_p$ in the above simulation, while $\mathbf{M}_p$ should be the previous reconstructed data in the practical implementation. It is necessary to verify the effectiveness of the proposed method in long-term reconstruction, which updates $\mathbf{M}_p$ with previous reconstruction and estimates spatial distribution $\mathbf{L}$ for next time slot recovery. Then for the second aspect, the long-term reconstruction, data collected from the 50th time slot to the 124th time slot in the dataset were selected. Set the number of cluster head n_{CH} as $[1, 5, 9]$ and the sampling ratio $\rho = 0.8$. Other settings were set as the previous simulation. Figure 5 shows the long-term recovery performance of the proposed method for the temperature data in GreenOrbs.

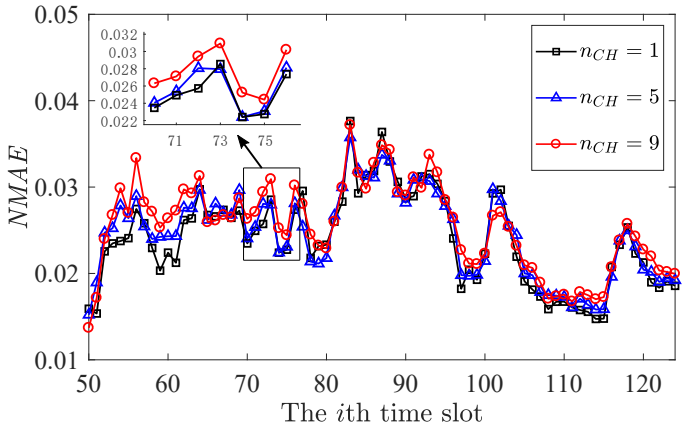


Figure 5. Long-term recovery of the proposed method for the temperature data in GreenOrbs.

Despite the reduction in the recovery accuracy of the proposed method for long-term reconstruction, the overall reconstruction accuracy is still satisfying. The average NMAE is $[0.024, 0.0245, 0.0254]$ for the proposed method with $[1, 5, 9]$ cluster heads, respectively. It is obvious that less number of cluster head leads to lower NMAE. With a fixed number of CH, only $\lfloor n\rho \rfloor - n_{CH}$ sampling nodes are randomly selected while n_{CH} nodes are fixed. As a result, if $n_{CH} = 0$, all sampling nodes are randomly selected, which can achieve better recovery.

4.2. The Energy Consumption Performance

To further verify the benefit of the proposed method in reducing the energy consumption, we established the experimental environments for the proposed sparse sampling data gathering scheme using DEEC as the clustering algorithm.

In the simulation, a two-level heterogeneous network containing advanced nodes and normal nodes was utilized. The initial energy of normal node was E_0, and the initial energy of advance node was aE_0. There were 100 sensors (80 advanced nodes and 20 normal nodes) randomly distributed in a $100m \times 100m$ area and one sink located in the center of the area.

The parameters used in simulations are shown in Table 2. The transmission process adopted free space model and the multipath fading model for distance $d \leq 87m$ and $d > 87m$, respectively. There were 100 times data gathering in one round, and the cluster heads were reselected at the beginning of each round. The number of cluster head was set to 10 for both methods, and the sampling ratio was set as $[0.2, 0.4, 0.6]$ for the proposed method.

Table 2. Network energy consumption model.

Description	Value
a	4
Initial energy of normal node E_0	0.5J
Energy for transmit per bit	50nJ/bit
Energy for receiving per bit	50nJ/bit
Amplifier energy for free space model	10pJ/bit/m^2
Amplifier energy for multipath fading model	0.0013pJ/bit/m^4

Figures 6 and 7 show the energy consumption of original DEEC and the proposed sparse sampling data gathering scheme using DEEC. As shown in Figure 6, the nodes start to die in 798th round for original DEEC, while the proposed method keeps all nodes alive until the 1356th, 1941th, 3504th round with $\rho = 0.6$, $\rho = 0.4$, $\rho = 0.2$, respectively.

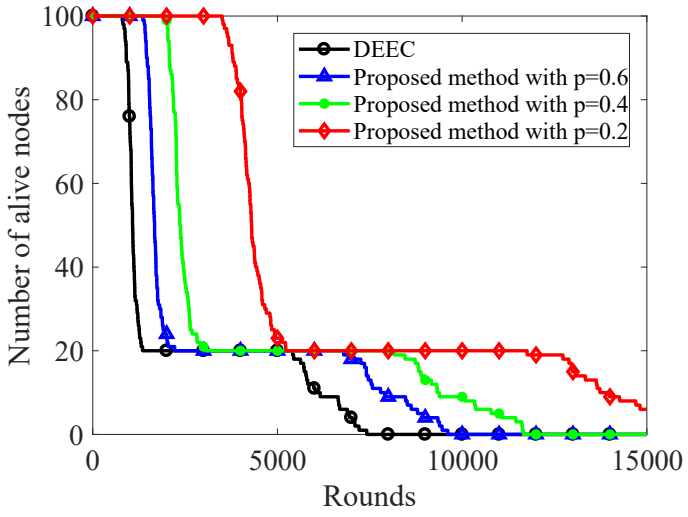


Figure 6. Number of alive nodes over rounds.

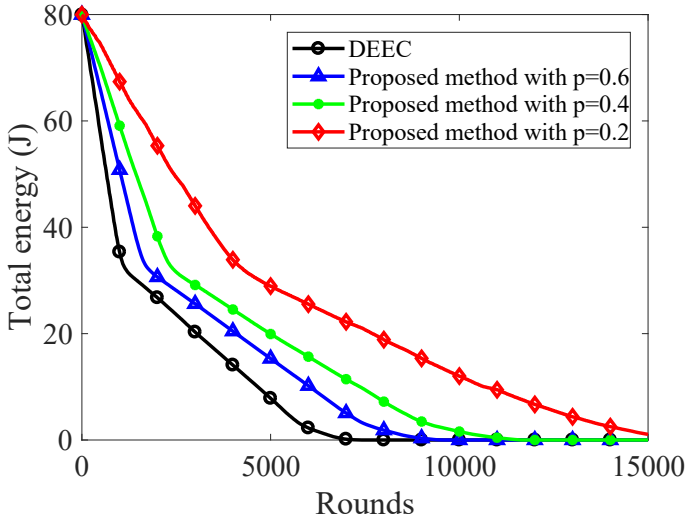


Figure 7. Total energy over rounds.

To keep 50% nodes alive, the original DEEC can only carry out 1081 rounds. The proposed method with $\rho = 0.2$ can carry out 4300 rounds, which is about 4 times than that of original DEEC. As the sampling ratio decreases, more network nodes can keep alive after a fixed number of rounds. As shown in Figure 7, the total energy of network decreases slowly using the proposed sparse sampling data gathering scheme. When the sampling ratios are 0.2, 0.4, and 0.6, the total energy of the proposed method is 33.9J, 24.55J, and 20.46J, respectively, while the original DEEC only had 14.1J at the 4000th round. To keep 20% total energy, the original DEEC can only carry out 3692 rounds, while the proposed method with $\rho = 0.2$ can carry out 8829 rounds. It is obvious that the proposed method can reduce the energy consumption significantly compared with the original DEEC.

5. Conclusions

This paper proposes an efficient sparse sampling data gathering method in clustered WSNs including a sparse sampling data gathering scheme and a subspace based reconstruction algorithm. The proposed sparse sampling data gathering scheme not only uses the sparse sampling strategy to reduce the amount of transmission data and the energy consumption, but also ensures the success of real-time sparse sampled data gathering in clustered WSNs. Then the previous reconstructed data are introduced with the sliding window model and used to estimate the subspace representing spatial distributions of the WSNs data. In this way, the proposed subspace based reconstruction approach can enforce an explicit low-rank constraint for data reconstruction from sparse sampled data. Besides, the total variation constraint is joint utilized to further improve the recovery accuracy. The experimental results show that the proposed method outperforms the state-of-the-art methods in data recovery and reduces the energy consumption of network efficiently.

Author Contributions: Conceptualization, J.H. and Y.Z.; methodology, J.H. and X.Z.; writing–original draft preparation, J.H. and X.Z.; writing–review and editing, J.H. and M.M.; supervision, J.H.; project administration, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (No. 61801164), Natural Science Foundation of Tianjin City (No. 18JCQNJC01700), Natural Science Foundation of Hebei Province (No. F2019202387), and Foundation of Hebei Educational committee (No. QN2018092).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yoon, S.; Shahabi, C. The Clustered AGgregation (CAG) technique leveraging spatial and temporal correlations in wireless sensor networks. *ACM Trans. Sens. Networks (TOSN)* **2007**, *3*, 3–es. [CrossRef]
2. Luo, C.; Wu, F.; Sun, J.; Chen, C.W. Compressive data gathering for large-scale wireless sensor networks. In Proceedings of the 15th Annual International Conference on Mobile Computing and Networking, Beijing, China, 20–25 September 2009; pp. 145–156.
3. Cheng, J.; Ye, Q.; Jiang, H.; Wang, D.; Wang, C. STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 850–861. [CrossRef]
4. Qing, L.; Zhu, Q.; Wang, M. Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. *Comput. Commun.* **2006**, *29*, 2230–2237. [CrossRef]
5. Lindsey, S.; Raghavendra, C.; Sivalingam, K.M. Data gathering algorithms in sensor networks using energy metrics. *IEEE Trans. Parallel Distrib. Syst.* **2002**, *13*, 924–935. [CrossRef]
6. Zhang, H.; Shen, H. Balancing energy consumption to maximize network lifetime in data-gathering sensor networks. *IEEE Trans. Parallel Distrib. Syst.* **2008**, *20*, 1526–1539. [CrossRef]
7. Donoho, D.L. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306. [CrossRef]
8. Candès, E.J.; Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.* **2009**, *9*, 717–772. [CrossRef]
9. Xiang, L.; Luo, J.; Vasilakos, A. Compressed data aggregation for energy efficient wireless sensor networks. In Proceedings of the 2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, Salt Lake City, UT, USA, 27–30 June 2011.
10. Markus, L.; Marian, C.; Markku, J. Sequential compressed sensing with progressive signal reconstruction in wireless sensor networks. *IEEE Trans. Wirel. Commun.* **2015**, *14*, 1622–1635.
11. Zhao, C.; Zhang, W.; Yang, Y.; Yao, S. Treelet-based clustered compressive data aggregation for wireless sensor networks. *IEEE Trans. Veh. Technol.* **2015**, *64*, 4257–4267. [CrossRef]
12. Ebrahimi, D.; Assi, C. Compressive data gathering using random projection for energy efficient wireless sensor networks. *Ad Hoc Networks* **2014**, *16*, 105–119. [CrossRef]
13. Wu, X.; Xiong, Y.; Yang, P.; Wan, S.; Huang, W. Sparsest random scheduling for compressive data gathering in wireless sensor networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 5867–5877. [CrossRef]
14. Ebrahimi, D.; Assi, C. On the interaction between scheduling and compressive data gathering in wireless sensor networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 2845–2858. [CrossRef]
15. Kong, L.; Xia, M.; Liu, X.Y.; Chen, G.; Gu, Y.; Wu, M.Y.; Liu, X. Data loss and reconstruction in wireless sensor networks. *IEEE Trans. Parallel Distrib. Syst.* **2014**, *25*, 2818–2828. [CrossRef]
16. He, J.; Sun, G.; Zhang, Y.; Wang, Z. Data Recovery in Wireless Sensor Networks With Joint Matrix Completion and Sparsity Constraints. *IEEE Commun. Lett.* **2015**, *19*, 2230–2233. [CrossRef]
17. Xie, K.; Wang, L.; Wang, X.; Xie, G.; Wen, J. Low cost and high accuracy data gathering in WSNs with matrix completion. *IEEE Trans. Mob. Comput.* **2018**, *17*, 1595–1608. [CrossRef]
18. Xu, Y.; Sun, G.; Geng, T.; He, J. Low-Energy Data Collection in Wireless Sensor Networks Based on Matrix Completion. *Sensors* **2019**, *19*, 945. [CrossRef]
19. He, J.; Zhou, Y.; Sun, G.; Geng, T. Multi-Attribute Data Recovery in Wireless Sensor Networks With Joint Sparsity and Low-Rank Constraints Based on Tensor Completion. *IEEE Access* **2019**, *7*, 135220–135230. [CrossRef]
20. Recht, B.; Fazel, M.; Parrilo, P.A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **2010**, *52*, 471–501. [CrossRef]
21. Cai, J.F.; Candès, E.J.; Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **2010**, *20*, 1956–1982. [CrossRef]
22. Hardt, M. Understanding alternating minimization for matrix completion. In Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, Philadelphia, PA, USA, 18–21 October 2014; pp. 651–660.
23. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2011**, *3*, 1–122. [CrossRef]
24. Parikh, N.; Boyd, S.P. Proximal Algorithms. *Found Trends Optim.* **2014**, *1*, 127–239. [CrossRef]
25. Donoho, D.L. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **1995**, *41*, 613–627. [CrossRef]

26. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
27. He, Y.; Mo, L.; Liu, Y. Why are long-term large-scale wireless sensor networks difficult: early experience with GreenOrbs. *ACM SIGMOBILE Mob. Comp. Commun. Rev.* **2010**, *14*, 10–12. [CrossRef]