JSSDR: Joint-Sparse Sensory Data Recovery in Wireless Sensor Networks

Guangshuo Chen*, Xiao-Yang Liu*, Linghe Kong*†, Jia-Liang Lu*, Wei Shu‡* and Min-You Wu*
*Shanghai Jiao Tong University, China
†Singapore University of Technology and Design, Singapore
‡University of New Mexico, Albuquerque, USA
*{chengs, yanglet, linghe.kong, jlu, mwu}@sjtu.edu.cn, ‡shu@ece.unm.edu

***Abstract*—Data loss is ubiquitous in wireless sensor networks (WSNs) mainly due to the unreliable wireless transmission, which results in incomplete sensory data sets. However, the completeness of a data set directly determines its availability and usefulness. Thus, sensory data recovery is an indispensable operation against the data loss problem. However, existing solutions cannot achieve satisfactory accuracy due to special loss patterns and high loss rates in WSNs. In this work, we propose a novel sensory data recovery algorithm which exploits the spatial and temporal joint-sparse feature. Firstly, by mining two real datasets, namely the Intel Indoor project and the GreenOrbs project, we find that: (1) for one attribute, sensory readings at nearby nodes exhibit inter-node correlation; (2) for two attributes, sensory readings at the same node exhibit inter-attribute correlation; (3) these inter-node and inter-attribute correlations can be modeled as the spatial and temporal joint-sparse features, respectively. Secondly, motivated by these observations, we propose two Joint-Sparse Sensory Data Recovery (JSSDR) algorithms to promote the recovery accuracy. Finally, real data-based simulations show that JSSDR outperforms existing solutions. Typically, when the loss rate is less than 65%, JSSDR can estimate missing values with less than 10% error. And when the loss rate reaches as high as 80%, the missing values can be estimated by JSSDR with less than 20% error.**

***Index Terms*—Wireless sensor networks, data loss, sensory data recovery, joint-sparse, compressive sensing**

I. Introduction

Wireless sensor networks (WSNs) [1] are widely used by researchers for studying the physical word [13]. Through WSNs, scientists gather information and reconstruct environmental data, which is important for them to discover the physical world around. For example, the sensory data of volcanos' temperature and shaking can be used for the prediction of eruption [23][17][21], and the one of the wind speed, air humidity and temperature can help scientists to reveal the plant evolution [12]. Usually, massive data missing is common in WSNs. For instance, the data loss rates are 64% and 35% in the Ocean Sense project [27] and the GreenOrbs project [20]. Hence recovering these lost data with high accuracy is challenging because of this situation.

The high loss rates break the structural features of values. Therefore classical interpolation methods, such as K-Nearest Neighbors (KNN) [7], cannot provide a satisfactory result because values of neighbors are very likely to be missing. Similarly, the performance of other classic interpolation methods is also influenced by high loss rates. A recently proposed compressive sensing approach, the *Environmental Space Time Improved Compressive Sensing* (ESTI-CS) [18][17], can achieve better accuracy. However, the low-rank and sparse features are also effected in the massive data loss scenario where the ESTI-CS experiences the increased estimation error.

In this paper, we propose to further improve the recovery accuracy by the correlation among multiple attributes. Our work is based on the following facts. The first one is that, a WSN's node is able to gather multiple attributes at the same time. For example, in [20], TelosB nodes are used to sense temperature, light illumination and humidity. The second one is that, there are stable correlations among many physical attributes in nature, such as humidity and temperature [19]. These correlations, if formulated mathematically and suitably, can be treated as the supplement of the inner attribute features and be useful for enhancing the accuracy of the estimation. Hence, how to mine and exploit such correlations is the key point hereby.

In this paper, firstly, by observing the real sensory data from the Intel Indoor project [14] and the GreenOrbs project [20], we find that: (1) for one attribute, sensory readings at nearby nodes exhibits inter-node correlation; (2) for multiple attributes, sensory readings at the same node exhibits inter-attribute correlation; (3) these inter-node and inter-attribute correlation can be modeled as the spatial and temporal joint-sparse features, respectively. Secondly, combing with the traditional compressive sensing theory, we propose two novel algorithms, called the Joint-Sparse Sensory Data Recovery (JSSDR), to recover single attribute or multiple attributes jointly by exploiting those correlation. Thirdly, we simulate the proposed approach on real data sets. We compare JSSDR with the classical and state-of-the-art methods such as KNN [7] and ESTI-CS [18].

Our contributions are summarized as following.

1) We mine two large WSN datasets and reveal the inter-node and inter-attribute correlation of them.
2) We design two novel interpolation algorithms, called JSSDR, based on compressive sensing theory.
3) The simulations are made based on real data, the result shows that our approach is very effective for sensory data recovery in WSNs.

The rest parts of this paper are organized as following. In Section II, we present the related work. Section III shows the

978-1-4799-0428-0/13/$31.00 ©2013 IEEE

problem formulation. Section IV mines the internal and external features of attributes in WSNs. Section V proposes our approach, JSSDR. The performance is evaluated in Section VI. In Section VII, the conclusion and future work are presented.

II. Related Work

There are a large number of excellent works with contributions to the field of missing data interpolation. Among them, K-Nearest Neighbors (KNN) [7] is a simple and classic method, which makes the missing data evaluation by utilizing the neighbors' average value. When the situation of data missing is mild, this classic interpolation method can achieve acceptable performance of recovery. With the increasing loss rate, the recovery accuracy of KNN turns bad rapidly, because of the lack of neighbor's information. Another widely used approach is: model the data as a time series and then apply some kind of prediction method, for example, the grey prediction method [11][22].

Recently, there is a popular solution for estimating massive missing data, named Compressive Sensing (CS) [8][5][4]. To deal with the problems in different fields, a series of CS-based or CS-extended solutions are developed, *e.g.*, Beyesian Compressive Sensing [15], as well as Kalman Filtered Compressive Sensing [26] are utilized in the field of signal processing. Further, Multi-Task Compressive Sensing (MTCS) [16] points at both signal and image processing.

Compressive Sensing was first introduced into WSNs [2] to save energy consumption in the data connection process and in turn prolong the network lifetime. Then [3] studied to estimate the received signals taking advantage of joint source-channel communication. Fornasier et al. [10] theoretically analyzed the joint sparsity constraints. Although these works consider the joint features and are advance and powerful in many fields, they neither benefit from inter-attribute correlations nor fall into the sensory data recovery field.

The most related work is ESTI-CS [18] and MACS [6], which are the state-of-the-art CS-based data recovery methods in WSNs. They exploit the low-rank feature, spatial-temporal feature and inter-attribute correlation from the sensory data against the special loss patterns of WSNs. Nevertheless, the low-rank, sparse and inter-attribute correlation features are also affected in the massive data loss scenario where the proposed methods experience the increased estimation error.

To the best of our knowledge, all the methods we find are against missing value problem on a single attribute. In nature, there are stable correlations among many physical attributes such as humidity and temperature [19]. We present our work to go a step further, aiming at improve the recovery accuracy exploiting such correlations.

III. Problem Formulation

A. Sensory Data Recovery

Suppose N nodes are deployed in an area, each of which can measure K different attributes at the same time. The sensing action lasts in a suitable time period which includes T equal time slots. Each node sends a data packet to the sink in every time slot. The format of the data packet is as following:

Node ID	Time Stamp	Attribute 1	Attribute 2	...

Hence $N \times T$ data packets are generated at the sensor nodes after T time slots and each packet contains K different attribute values. The sensory data matrix, the original data matrix and the sample matrix are denoted as $\mathbf{s} \in \mathbb{R}^{K\times N\times T}$, $\mathbf{x} \in \mathbb{R}^{K\times N\times T}$ and $\mathbf{A} \in \mathbb{R}^{N\times T\times T}$, respectively.

Data loss is ubiquitous in wireless sensor networks (WSNs) mainly due to the unreliable wireless transmission, which results in incomplete sensory data sets. We model the data loss problem in the following way. For each node n, the sample pattern is denoted as $\mathbf{A}(n) \in \mathbb{R}^{T\times T}$, which is a diagonal matrix and satisfies $\mathbf{A}(n,i,i) = 1,\ 1 \le i \le T$ if the sensory value is received in the ith time slot, otherwise 0.

Here we assume that all vectors of the same node share the same $\mathbf{A}(n)$ because if one data packet is lost, all attributes in it is missing.

So

$$\mathbf{s}(i,n) = \mathbf{A}(n)\mathbf{x}(i,n). \tag{1}$$

In Eqn.(1), all $\mathbf{x}(i,n)$ and $\mathbf{s}(i,n)$ are $T \times 1$ vectors. Equivalently, for $1 \le i \le K$, $1 \le n \le N$ and $1 \le t \le T$, $\mathbf{s}(i,n,t)$ is represented as following.

$$\mathbf{s}(i,n,t) = \begin{cases} \mathbf{x}(i,n,t), & \text{if } \mathbf{A}(n,t,t) = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Hereby $\mathbf{s}$ is incomplete, so we need to recover the missing values of $\mathbf{s}$.

B. Problem Statement

Our problem is to recover the original data $\mathbf{x}$ from the sensory data $\mathbf{s}$ as precisely as possible. The recovered sensory data, denoted as $\hat{\mathbf{x}} \in \mathbb{R}^{K\times N\times T}$, can be used by scientists to discover the physical world around. The problem is called *sensory data recovery (SDR)* problem.

1) Single Attribute Scenario: Consider the SDR problem under the single attribute scenario, which is defined mathematically as following.

Given $\mathbf{s}$, find an optimal evaluation of $\mathbf{x}$ as $\hat{\mathbf{x}}$, *i.e.*,

$$\begin{aligned} \min \quad & ||\hat{\mathbf{x}}(n) - \mathbf{x}(n)||_1, \\ s.t. \quad & \mathbf{s}(n) = \mathbf{A}(n)\hat{\mathbf{x}}(n), \\ & \forall n,\ 1 \le n \le N, \end{aligned} \tag{3}$$

where $||\cdot||_1$ represents the l_1-norm, which is used in [8][5], *e.g.*, for $\mathbf{x} = [x_1, \cdots, x_n]$, the l_1-norm is defined as $||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$.

2) Multiple Attributes Scenario: Since we focus on exploiting the correlation among multiple attributes, sensory data of different attributes are estimated jointly. In the multiple attributes scenario, the problem is formulated as following.

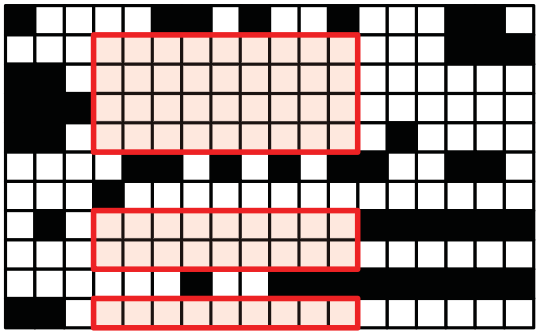


Fig. 1. Filter the the original dataset by selecting the red parts to construct a small but completed dataset as the ground truth [18].

TABLE I
SELECTED DATA SETS AS THE GROUND TRUTH

Data Name	Matrix Size	Time Interval
Intel Indoor	49 nodes × 149 intervals	15 minutes
GreenOrbs	281 nodes × 170 intervals	30 minutes

Given $\mathbf{s}$, find a series of optimal evaluations of $\mathbf{x}$ as $\hat{\mathbf{x}}$, *i.e.*,

$$\begin{aligned} \min \quad & \sum_{i=1}^{K} ||\hat{\mathbf{x}}(i,n) - \mathbf{x}(i,n)||_1, \\ s.t. \quad & \mathbf{s}(i,n) = \mathbf{A}(n)\hat{\mathbf{x}}(i,n), \\ & \forall i,\ 1 \le i \le K, \quad \forall n,\ 1 \le n \le N. \end{aligned} \tag{4}$$

IV. OBSERVATIONS IN SENSORY DATASETS

In this section, we analyze the real datasets of WSNs and discover several features of them, which are the foundations for our data recovery approach.

A. Real World Sensor Network Projects

1) The Intel Indoor Project: Intel Indoor project [14] was carried by the Intel Berkeley Research Lab between February 28th and April 5th, 2004. In the project, 54 nodes were deployed in the room, which can collect sensory data such as humidity, temperature, light and voltage once every 31 seconds. Data was collected using the TinyDB in-network query processing system, built on the TinyOS platform.

2) The GreenOrbs Project: The GreenOrbs [20] project aims at all-year round ecological surveillance in the forest, collecting sensory data such as temperature, humidity and light illumination, and content of carbon dioxide.

GreenOrbs employs the TelosB nodes with a MSP430 processor and CC2420 transceiver. The software on the GreenOrbs nodes is developed on the basis of TinyOS 2.1.

The project started at the year of 2008. Ever since then, GreenOrbs has experienced a number of deployments at different places, with different scales, and for different duration. Currently, over 1000 nodes are deployed.

B. Data Sets

To reveal inter and intra features among attributes, it is in need to make observations on complete data sets. And integrated data sets are also required for evaluating the performance of our approach. So the integrality of data sets is important in this paper.

The original data sets are gathered from two projects, GreenOrbs [20] and Intel Indoor [14]. After investigating the raw data, the loss rates of these two data sets are 35% and 23%, respectively. In order to obtain the data sets as the ground truth, two small but completed data sets are selected as shown in Table.I. The selection method is shown in Fig.1, which considers the maximization of the integrality in both time and space. Each data set contains subsets of two attributes: temperature and light illumination, both of which share the same selecting entries.

C. Spatial Joint-Sparse Feature

The spatial correlation is already revealed in our recent work [18]. Here, we study the spatial correlation of sensory data in WSNs further. It is known that environments are often smooth in a small area, which leads to the face that the readings of nearby sensors are close. Hence, we mine the inherent structure or redundancy of sensory vectors gathered by nodes nearby.

According to [18], sensory vectors are approximately sparse under the wavelet field and the Discrete Cosine Transform (DCT) field. In this paper, DCT is used as the basis for exploiting sparsity.

To environment data of an attribute, denoted as $\mathbf{x} \in \mathbb{R}^{N\times T}$, consider several nodes, which are neighbors of a center node in a r-radius circle. The attribute vectors gathered by these nodes are represented as $\mathbf{x}(a_1), ..., \mathbf{x}(a_n)$, where $1 \le a_1, \cdots, a_n \le N$ are indices of these neighbors. Since the spatial correlation is revealed, it is resonable to decompose them under the DCT basis as following,

$$\mathbf{x}(a_j) = \mathbf{\Psi}\theta_c + \mathbf{\Psi}\theta(a_j). \tag{5}$$

where $\mathbf{\Psi}$ is the DCT basis, $\theta_c \in \mathbb{R}^T$ and $\theta(a_j) \in \mathbb{R}^T,\ \forall a_j$.

The decomposition can be made by using the compressive sensing method. Firstly, integrate Eqn.(5) of all a_j, as,

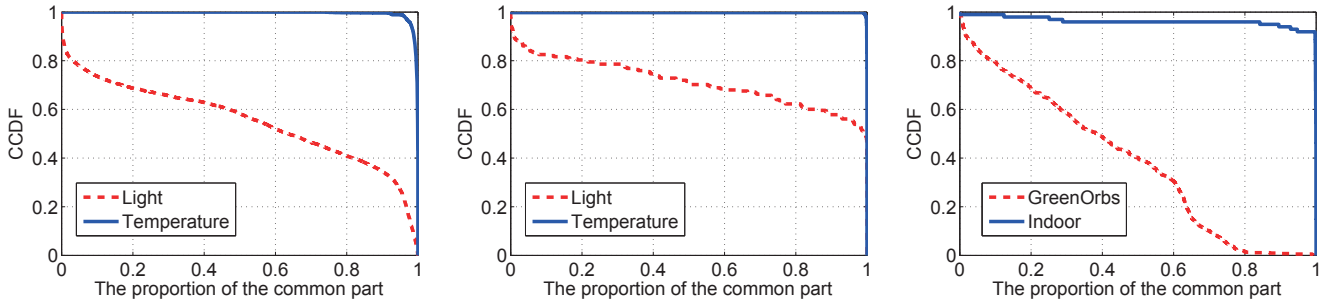
$$\begin{bmatrix} \mathbf{x}(a_1) \\ \vdots \\ \mathbf{x}(a_n) \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi} & \mathbf{\Psi} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Psi} & 0 & \cdots & \mathbf{\Psi} \end{bmatrix} \begin{bmatrix} \theta_c \\ \theta(a_1) \\ \vdots \\ \theta(a_n) \end{bmatrix}, \tag{6}$$

which is represented simply as $\mathbf{y} = M\theta$. Secondly, solve the l_1-norm normalization problem, *i.e.*, $\min ||\theta||_1,\ s.t.\ \mathbf{y} = M\theta$. Thirdly, obtain θ_c and all $\theta(a_j)$ from θ.

After the decomposition, $||\theta_c||_1/(||\theta_c||_1 + ||\theta(a_j)||_1)$ are calculated. As shown in Fig.2(a) and Fig.2(b), θ_c contains the main part of $\mathbf{x}(a_j)$ under the basis $\mathbf{\Psi}$ to data sets of indoor/outdoor temperature and light illumination. It is also observed that θ_c and all $\theta(a_j)$ are sparse. Hence the spatial joint-sparse feature of sensory data is revealed.

D. Temporal Joint-Sparse Feature

The temporal stability feature is also revealed in [18]. And the inter-node temporal correlation leads to the sparsity of sensory vectors under the DCT basis. However, the paper [18]



(a) Spatial joint-sparse in GreenOrbs. (b) Spatial joint-sparse in Intel Indoor. (c) Temporal joint-sparse in GreenObrs and Intel Indoor.

Fig. 2. The decomposition in the DCT basis for real sensory data sets.

only talks about the inter-node temporal correlation feature of sensory vectors.

In this paper, we mine the inter-attribute temporal joint-spare feature. The relationship usually exists among natural attributes. For instance, the empirical study [19] reveals that temperature, dewpoint temperature and relative humidity have linear correlation under some special cases. Intuitively, light illumination and temperature have the same trend of change in outdoor environment. Such a trend may be not clear in the time field but can be reflected in other fields. Hence, we try to mine the feature of attributes under the DCT field by using decomposition.

Consider $\mathbf{x}(i) \in \mathbb{R}^T,\ 1 \leq i \leq K$, which are sensory vectors of different attributes gathered by a given node. After the process of normalization, decompose them under the DCT field as following,

$$\mathbf{x}(i) = \mathbf{\Psi}\vartheta_c + \mathbf{\Psi}\vartheta(i),\ 1 \leq i \leq K. \tag{7}$$

where $\vartheta_c, \vartheta(i) \in \mathbb{R}^T$.

During the observation of $\vartheta(1), \cdots, \vartheta(K)$ and ϑ_c, we find that (1) all of them are sparse and (2) $||\vartheta_c||_2$ is far larger than any $||\vartheta(i)||_2,\ 1 \leq i \leq K$, which means that sensory vectors of different attributes share a common part under the DCT basis as shown in Fig.2(c). Since all vectors are obtained from the time field, we can say that the temporal joint-spare feature exists in multiple attribute scenario.

V. Our Approach

To address the SDR problem, we propose a novel relative data estimation approach named *Joint-Sparse Sensory Data Recovery (JSSDR)*, which is designed to jointly recover the attributes in a WSN.

We propose the approach in the single attribute scenario first, and then in the multiple attributes scenario.

A. Single Attribute Scenario

1) Compressive Sensing: Assume that $\mathbf{x}(n)$ is k-sparse under the basis $\mathbf{\Psi}$ and the following condition holds, *i.e.*,

$$T \geq c \cdot \mu^2(\mathbf{A}(n), \mathbf{\Psi}) \cdot k \cdot logT, \tag{8}$$

where c is a positive constant value and $\mu(\Phi, \Psi)$ is a metric measuring the largest correlation between any two elements of Φ and Ψ, which is defined as following,

$$\mu(\Phi, \Psi) = \sqrt{T} \cdot \max_{1 \leq i,j \leq T} |\langle \phi_i, \psi_j \rangle|, \tag{9}$$

where $\langle \cdot, \cdot \rangle$ is the inner product operator. According to the compressive sensing theory, $\mathbf{x}(n)$ can be recovered by solving an l_1-norm normalization problem, *i.e.*,

$$\begin{aligned} \min \quad & ||\theta(n)||_1 \\ s.t. \quad & \mathbf{s}(n) = \mathbf{A}(n)\mathbf{\Psi}\theta(n), \\ & \hat{\mathbf{x}}(n) = \mathbf{\Psi}\theta(n), \\ & \forall n,\ 1 \leq n \leq N, \end{aligned} \tag{10}$$

where $\hat{\mathbf{x}}(n)$ is the evaluation of $\mathbf{x}(n)$, and $||\cdot||_1$ is the l_1 norm.

Further, to avoid the overfitting problem, Eqn.(10) is relaxed as following.

$$\begin{aligned} \min \quad & ||\theta(n)||_1 \\ s.t. \quad & ||\mathbf{s}(n) - \mathbf{A}(n)\mathbf{\Psi}\theta(n)||_2 < \varepsilon, \\ & \hat{\mathbf{x}}(n) = \mathbf{\Psi}\theta(n), \\ & \forall n,\ 1 \leq n \leq N, \end{aligned} \tag{11}$$

where ε is a predefined threshold and $||\cdot||_2$ is the l_2 norm, *e.g.*, for $\mathbf{x} = (x_1, \cdots, x_n)$, the l_2 norm is defined as $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

2) Spatial Joint Sparse Recovery: Eqn.(11) can be solved by compressive sensing. As the spatial joint-sparse feature is revealed, our approach is designed to benefit from this feature.

Suppose $\mathbf{x}(a_1), \cdots, \mathbf{x}(a_n),\ 1 \leq a_1, \cdots, a_n \leq N$ are vectors in a r-radius circle, whose values are close. Hence, because of the spatial joint-sparse feature, it is reasonable to assume that $\hat{\mathbf{x}}(a_1), \cdots, \hat{\mathbf{x}}(a_n)$ satisfy the same feature and they can be decomposed under the basis $\mathbf{\Psi}$ into a common part and individual parts, *i.e.*,

$$\hat{\mathbf{x}}(a_j) = \mathbf{\Psi}\theta_c + \mathbf{\Psi}\theta(a_j),\ 1 \leq a_1, \cdots, a_n \leq N, \tag{12}$$

where $\theta(a_1), ..., \theta(a_n)$ and θ_c are sparse vectors under the basis $\mathbf{\Psi}$.

Algorithm 1 JSSDR for the single attribute scenario

Input:
$\mathbf{x}$: sensory data
$\mathbf{A}$: sample matrix
$\mathbf{P}$: positions of nodes
r: radius of neighbor circle
Output:
$\hat{\mathbf{x}}$: estimated environment data
Notation:
$N_r(n)$: a set of neighbors of the nth node
$\sharp N_r(n)$: the number of neighbors of the nth node
$SolveLasso$: the least angle regressio[9] solver
Main Procedure:

```
for n ∈ 1 to N do
    b_n = ♯N_r(n);
    y ← Eqn.(13)
    M ← Eqn.(13)
    θ = SolveLasso(y, M)
    x̂_n(n) ← Eqn.(14)
    for i ∈ N_r(n) do
        x̂_n(i) ← Eqn.(14)
    end for
end for
for n ∈ 1 to N do
    x̂(n) ← Eqn.(15)
end for
return x̂
```

Then, to Eqn.(11) in $a_1, \cdots, a_n$, integrate them as,

$$\begin{bmatrix} \mathbf{s}(a_1) \\ \vdots \\ \mathbf{s}(a_n) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(a_1)\mathbf{\Psi} & \mathbf{A}(a_1)\mathbf{\Psi} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}(a_n)\mathbf{\Psi} & 0 & \cdots & \mathbf{A}(a_n)\mathbf{\Psi} \end{bmatrix} \begin{bmatrix} \theta_c \\ \theta(a_1) \\ \vdots \\ \theta(a_n) \end{bmatrix}, \tag{13}$$

which is represented as $\mathbf{y} = \mathbf{M}\theta$, for simplicity.

$\mathbf{x}(a_1), \cdots, \mathbf{x}(a_n)$ is able to be recovered by solving the following minimization problem,

$$\begin{aligned} \min \quad & ||\theta||_1 \\ s.t. \quad & ||\mathbf{y} - \mathbf{M}\theta||_2 < \varepsilon, \\ & \hat{\mathbf{x}}(a_i) = \mathbf{\Psi}(\theta_c + \theta(a_j)), \\ & \theta = [\theta_c^T, \theta(a_1)^T, \cdots, \theta(a_n)^T]^T \\ & 1 \le a_1, \cdots, a_n \le N \end{aligned} \tag{14}$$

3) Weighted Average of multiple calculations: If a node i is involved in m r-radius circles, $\hat{\mathbf{x}}(i)$ will be calculated by m times. Here, a weighted average method is proposed to obtain $\hat{\mathbf{x}}(i)$.

Suppose each r-radius circle contains b_j nodes, then

$$\hat{\mathbf{x}}(i) = \frac{\sum_{j=1}^{m} b_j \cdot \hat{\mathbf{x}}_j(i)}{\sum_{j=1}^{m} b_j} \tag{15}$$

where $\hat{\mathbf{x}}_j(i)$ is the evaluation of $\mathbf{x}(i)$ in the jth circle.

The detail of the approach in the single attribute scenario is shown in Alg.1.

B. Multiple Attributes Scenario

1) Normalization: : Because attributes are in different dimensions, the process of normalization is required to ensure that all vectors are in the same dimension. The process is in need of the maximum value of each vector, *i.e.*, $\max(\mathbf{x}(i,n))$. But the real maximum value is possible to loss, hence we adopt the maximum value in gathered vectors instead, *i.e.*, $\max(\mathbf{s}(i,n))$ is used on the normalization for $1 \le i \le K$.

This process is based on the observation that the natural attributes changes gradually. In other words, the gap between maximum values of the observed matrix and the original matrix is small compared with the magnitude, *i.e.*, for $1 \le i \le K$

$$\max(\mathbf{x}(i,n)) - \max(\mathbf{s}(i,n)) \ll \max(\mathbf{x}(i,n)) \tag{16}$$

.

2) Temporal and Spatial Joint Sparse Recovery: After the process of normalization, all vectors are in the same dimension. Since we find the temporal joint-sparse feature of vectors among attributes, the estimation approach can benefit from this feature.

Suppose sensory vectors of K different attributes in a r-radius circle, represented as $\mathbf{x}(i,a_1), ..., \mathbf{x}(i,a_n)$, where $1 \le i \le K$. Because of both the temporal and spatial joint-sparse feature, $\mathbf{x}(i,n)$ can be decomposed according to Eqn.(5) and Eqn.(7), *i.e.*,

$$\mathbf{x}(i,n) = \vartheta + \theta_i + \delta(i,n),\ 1 \le i \le K \tag{17}$$

where ϑ is the inter-attributes common component, θ_i is the inter-node common component and $\delta(i,n)$ is the individual component.

Similarly, it is reasonable to assume that $\hat{\mathbf{x}}(i,n)$ satisfies Eqn.(17). Hence, the decomposition is used in our estimation approach in multiple attributes scenario.

Consider the integrated matrix as following,

$$\mathbf{Y} = \mathbf{M}\eta \tag{18}$$

where $\mathbf{M}$ is defined as following,

$$\begin{bmatrix} \mathbf{A}(a_1)\mathbf{\Psi} & \mathbf{A}(a_1)\mathbf{\Psi} & \cdots & 0 & \mathbf{A}(a_1)\mathbf{\Psi} & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots & \ddots & \vdots \\ \mathbf{A}(a_n)\mathbf{\Psi} & 0 & \cdots & \mathbf{A}(a_n)\mathbf{\Psi} & 0 & \cdots & \mathbf{A}(a_n)\mathbf{\Psi} \end{bmatrix}, \tag{19}$$

and

$$\mathbf{y} = [\mathbf{s}(1,a_1)^T, ..., \mathbf{s}(1,a_1)^T, \cdots, \mathbf{s}(K,a_1)^T, \cdots, \mathbf{s}(K,a_n)^T]^T \tag{20}$$

as well as

$$\eta = [\vartheta^T, \theta_1^T, ..., \theta_K^T, \delta(1,a_1)^T, ..., \delta(K,a_n)^T]^T \tag{21}$$

Eqn.(18) can be solved like Eqn.(14). For calculations in different r-radius circles, the weighted average process is still used to obtain $\mathbf{x}(i,n)$.

The pseeudo-code of the approach in the multiple attributes scenario is presented in Alg.2.

C. Complexity Analysis

In JSSDR, the main operation is the joint-sparse recovery. The normalization and weighted average calculating operations are both in the complexity of $O(N)$, which are negligible compared with the one of the joint-sparse recovery operation. Hence the complexity of our approach depends on two issues, *i.e.*,

1) the complexity of the method for solving l_1-norm minimization problem.
2) the network topology of the environment where the sensors deployed.

For example, suppose the complexity of solving an l_1-norm minimization problem is $f(p,q)$ where p is the number of measurements and q is the size of the spare vector, each node is involved by m r-radius circles and each circle contains z nodes in average. Then in the single attribute scenario, the number of the l_1-norm minimization problems is $N \cdot m$, so that the average complexity of JSSDR is $N \cdot m \cdot f(z \cdot T, (z+1) \cdot T)$. Similarly, the complexity is $K \cdot N \cdot m \cdot f(z \cdot K \cdot T, (z \cdot K+1) \cdot T)$ in the multiple attributes scenario.

If the least angle regression [9], whose complexity is $O(p^3 + qp^2)$ in average, where p is the number of measurements and q is the size of the spare vector, is used in solving l_1-norm minimization problems, the complexity will be $O(T^3)$ in both scenarios because $N, K, m, z \ll T$. In this paper, the least angle regression [9] method is adopted to solve all the l_1-norm minimization problems.

Algorithm 2 JSSDR for the multiple attributes scenario

Input:
$\mathbf{x}$: sensory data
$\mathbf{A}$: sample matrix
$\mathbf{P}$: positions of nodes
r: radius of neighbor circle
Output:
$\hat{\mathbf{x}}$: estimated environment data
Notation:
$N_r(n)$: a set of neighbors of the nth node
$\sharp N_r(n)$: the number of neighbors of the nth node
$SolveLasso$: the least angle regressio[9] solver
Main Procedure:

for $n \in 1\ to\ N$ **do**
  for $k \in 1\ to\ K$ **do**
    $\mathbf{s}(k,n) = \mathbf{s}(k,n)/max(\mathbf{s}(k,n))$
  end for
  $b_n = \sharp N_r(n)$
  $\mathbf{y} \leftarrow$ Eqn.(20)
  $\mathbf{M} \leftarrow$ Eqn.(19)
  $\eta = SolveLasso(\mathbf{y}, \mathbf{M})$
  for $k \in 1\ to\ K$ **do**
    $\hat{\mathbf{x}}_n(k,n) \leftarrow$ Eqn.(21)
    $\hat{\mathbf{x}}_n(k,n) = \hat{\mathbf{x}}_n(k,n) * max(\mathbf{s}(k,n))$
  end for
  for $i \in N_r(n)$ **do**
    for $k \in 1\ to\ K$ **do**
      $\hat{\mathbf{x}}_n(k,i) \leftarrow$ Eqn.(21)
      $\hat{\mathbf{x}}_n(k,i) = \hat{\mathbf{x}}_n(k,i) * max(\mathbf{s}(k,i))$
    end for
  end for
end for
for $n \in 1\ to\ N$ **do**
  for $k \in 1\ to\ K$ **do**
    $\hat{\mathbf{x}}(k,n) \leftarrow$ Eqn.(15)
  end for
end for
return $\hat{\mathbf{x}}$

VI. Performance Evaluation

A. Compared Methods

Lots of works have contributed in missing data interpolation.

1) K-Nearest Neighbors (KNN) method: The most classic interpolation method is K-Nearest Neighbors (KNN) [7]. Simple nearest neighbors uses the nearest neighbor for missing value interpolation. KNN extends this by using a weighted average of the k nearest-neighbors' values. The KNN perform well in common situations where a moderate number of values are missing. As loss rate grows, the estimation error increases quickly due to the lack of one-hop neighbors.

2) ESTI-CS: Compressive Sensing (CS) [8][5] is currently an advanced and powerful technique for estimating massive missing data.

Originally, the goal of CS is to recover a signal $\mathbf{x} \in \mathbb{R}^m$ from random measurements $\mathbf{b} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $n \ll m$. This problem is ill posed in general and has many solutions. The basic idea of CS is to seek the sparsest solution. A condition that $\mathbf{x}$ is sparse under a given basis $\mathbf{P}$ is required. Then in CS, the problem becomes

$$\hat{\mathbf{s}} = \arg\min ||\mathbf{s}||_1 \quad s.t.\ \mathbf{b} = \mathbf{A}\mathbf{P}\mathbf{s} \tag{22}$$

where $||\cdot||_1$ is the l_1-norm. The recovered signal is $\hat{x} = P\hat{s}$.

There are a series of CS based solutions being used in different fields, *e.g.*, Beyesian Compressive Sensing [15], Kalman Filtered Compressive Sensing [26] and Multi-Task Compressive Sensing (MTCS) [16] are utilized in the fields of signal processing and image processing.

The state-of-the-art CS based interpolation method utilized in the field of WSNs is ESTI-CS [18]. ESTI-CS exploits the low-rank feature and spatial-temporal feature from the sensory data against the special loss patterns of WSNs.

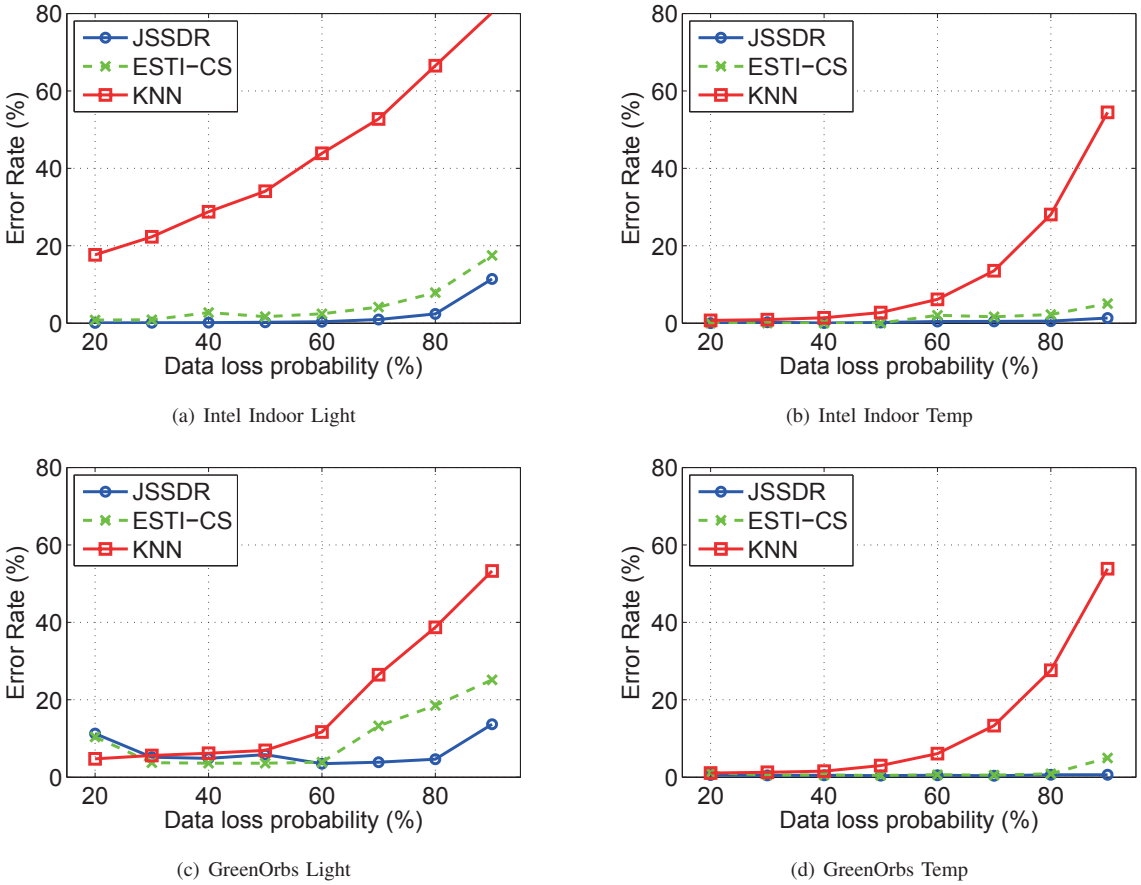
B. Methodology

Performance evaluation is based on real data driven simulation.

1) Ground Truth: The real data including the temperature and light attributes from GreenOrbs and Intel Indoor projects. In Sec.IV-B, we have presented the method to obtain the ground truth from raw data in detail.

2) Metric: To compare results evaluated from different data sets, the error rate of approximation under the l_2 norm, $err(\mathbf{x}, \hat{\mathbf{x}})$, is applied [18][25], which is defined as:

$$err(\mathbf{x}, \hat{\mathbf{x}}) = \frac{||\hat{\mathbf{x}} - \mathbf{x}||_2^2}{||\mathbf{x}||_2^2}. \tag{23}$$



(a) Intel Indoor Light

(b) Intel Indoor Temp

(c) GreenOrbs Light

(d) GreenOrbs Temp

Fig. 3. The accuracy of missing value estimation methods.

3) Procedure: The procedure of simulation is as following. Firstly, actively lose the data from the ground truth to simulate the gathered data in WSNs. Generate $\mathbf{A}$ randomly. The quantity of data loss is from 20% to 90%. And using data sets of two physical attributes, compute $\mathbf{s}$. Secondly, $\mathbf{s}$ and $\mathbf{A}$ serve as the inputs of the estimation algorithms, *i.e.*, KNN, ESTI-CS and JSSDR. Finally, Compare the performance of the algorithms on the error rate defined by Eqn.(23).

C. Simulation Results

In Fig.3, we plot the comparison result of three algorithms in the case of two attributes. According to the simulation, JSSDR can obtain less than 5% error rate under the loss rate less than 60%, where ESTI-CS can provide 10% and KNN is far weaker. Even in high loss rate (80%), the error rate of JSSDR is still less than 10%. The main reason is that JSSDR uses the correlation between two attributes. Hence, the accuracy of estimating missing values increases if the correlation exists. And even there are no relation between two data sets, the performance of JSSDR is as equal as ESTI-CS.

The recovery accuracy of the temperature is higher than the one of the light illumination. The main reason is that the temperature in outdoor WSNs changes slowly and has small amplitude, which leads to its strong time and space stabilities benefiting estimation methods. While the accuracy of light illumination in GreenOrbs is a little weak, the reason is that light illumination varies considerably in nature.

As shown in Fig.3, the estimation performance of KNN is barely satisfactory and reduces quickly as the increasing of data loss rate. The possible reason is that the massive data loss in WSNs veils the time and spatial correlations between attributes. Hence the interpolation methods can not benefit well from these features.

Totally, JSSDR outperforms ESTI-CS and KNN in random loss pattern, whatever the correlation between attributes exists or not.

VII. Conclusion

We investigated the *sensory data recovery (SDR)* problem of WSNs in this paper. We observed environmental data sets from real projects, *i.e.*, Intel indoor and GreenOrbs. The inter-node spatial joint-sparse feature and the intra-attribute temporal joint-sparse feature were revealed. After making such observations, we are enlightened and designed the JSSDR algorithm to estimate the missing values. The algorithm is extended from the basic compressive sensing method and can benefit from both the inner and intra correlations of attributes. It turns out that JSSDR outperforms existing interpolation

methods under the test of data-driven simulations.

The future works are as following. First, considering the integrity of the environment sensory data and developing better estimating approach. Second, improving the time and space complexity of our approach. Third, generalizing the multiple attributes data reconstruction to more fields.

ACKNOWLEDGMENT

This research was supported by NSF of China under grant No. 61073158, No. 61100210, STCSM Project No. 12dz1507400, No. 13511507800, Doctoral Program Foundation of Institutions of Higher Education under grant No. 20110073120021, Singapore-MIT International Design Center IDG31000101, iTrust Cyber Physical System Protection project, and the Singapore NRF under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office, Media Development Authority.

REFERENCES

[1] Akyildiz, Ian F., Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. Wireless sensor networks: a survey. *Elsevier Journal of Computer Networks*, Vol. 38, No. 4, pp. 393-422, 2002.

[2] Bajwa, Waheed and Haupt, Jarvis and Sayeed, Akbar and Nowak, Robert. Compressive wireless sensing. *ACM Proceedings of the 5th international conference on Information processing in sensor networks (IPSN)*, pp. 134-142, 2006.

[3] Bajwa, Waheed and Haupt, J and Sayeed, A and Nowak, R. Joint source–channel communication for distributed estimation in sensor networks. *IEEE Transactions on Information Theory*, Vol. 53, No. 10, pp. 3629-3653, 2007.

[4] Candès, Emmanuel J and Wakin, Michael B. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, Vol. 25, No. 2, pp. 21-30, 2008.

[5] Candès, Emmanuel J and Romberg, Justin and Tao, Terence. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, Vol. 52, No. 2, pp. 489-509, 2006.

[6] Chen, Guangshuo and Liu, Xiao-Yang and Kong, Linghe and Lu, Jia-Liang and Shu, Wei and Wu, Min-You. Multiple Attributes-based Data Recovery in Wireless Sensor Networks. *IEEE GlobeCom*, 2013.

[7] Cover, Thomas and Hart, Peter. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27, 1967.

[8] Donoho, David Leigh. Compressed sensing. *IEEE Transactions on Information Theory*, Vol. 52, No. 4, pp. 1289-1306, 2006.

[9] Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, Vol. 32, No. 2, pp. 407-499, 2004.

[10] Fornasier, Massimo and Rauhut, Holger. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, Vol. 46, No. 2, pp. 577-613, 2008.

[11] Fu, Cai and Gao, Xiang and Liu, Ming and Liu, Xiao-Yang and Han Lansheng and Chen Jing. GRAP: Grey risk assessment based on projection in ad hoc networks. *Elsevier Journal of Parallel and Distributed Computing*, Vol. 71, No. 9, pp. 1249-1260, 2011.

[12] Heil, Martin and Karban, Richard. Explaining evolution of plant communication by airborne signals. *Trends in Ecology & Evolution*, Vol. 25, No. 3,pp. 137-144, 2010.

[13] He, Tian and Krishnamurthy, Sudha and Stankovic, John A and Abdelzaher, Tarek and Luo, Liqian and Stoleru, Radu and Yan, Ting and Gu, Lin and Hui, Jonathan and Krogh, Bruce. Energy-efficient surveillance system using wireless sensor networks. *ACM Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pp. 270-283, 2004.

[14] Intel lab data. http://www.select.cs.cmu.edu/data/labapp3/index.html.

[15] Ji, Shihao and Xue, Ya and Carin, Lawrence Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, Vol. 56, No. 6, pp. 2346-2356, 2008.

[16] Ji, Shihao and Dunson, David and Carin, Lawrence. Multitask compressive sensing. *IEEE Transactions on Signal Processing*, Vol. 57, No. 1, pp. 92-106, 2009.

[17] Kong, Linghe and Zhao, Mingchen and Liu, Xiao-Yang and Lu, Jialiang and Liu, Yunhuai and Wu, Min-You and Shu, Wei. Surface coverage in sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 2013.

[18] Kong, Linghe and Xia, Mingyuan and Liu, Xiao-Yang and Wu, Min-You and Liu, Xue. Data loss and reconstruction in sensor networks. *IEEE INFOCOM*, pp. 1702-1710, April 2013.

[19] Lawrence, Mark G. The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. *Bulletin of the American Meteorological Society*, Vol. 86, pp. 225-233, 2005.

[20] Liu, Yunhao and He, Yuan and Li, Mo and Wang, Jiliang and Liu, Kebin and Mo, Lufeng and Dong, Wei and Yang, Zheng and Xi, Min and Zhao, Jizhong and et al. Does wireless sensor network scale? A measurement study on greenorbs. *IEEE INFOCOM*, pp. 873-881, 2007.

[21] Liu, Xiao-Yang and Wu, Kailiang and Zhu, Yanmin and Kong, Linghe and Wu, Min-You. Mobility increases the surface coverage of distributed sensor networks. *Elsevier Computer Networks*, Vol. 57, No. 11, pp. 2348C2363, 2013.

[22] Liu Xiao-Yang and Fu, Cai and Yang, Jiandong and Han Lansheng. Grey Model-Enhanced Risk Assessment and Prediction for P2P Nodes. *IEEE The Fourth International Conference on Frontier of Computer Science and Technology*, pp. 681-685, 2009.

[23] Rudolph, Maxwell L and Karlstrom, Leif and Manga, Michael. A prediction of the longevity of the lusi mud eruption, indonesia. *Elsevier Earth and Planetary Science Letters*, Vol. 308, No. 1, pp. 124-130, 2011.

[24] Recht, Benjamin and Fazel, Maryam and Parrilo, Pablo A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization *SIAM Review*, Vol. 52, No. 3, pp. 471-501, 2010.

[25] S. Rallapalli, L. Qiu, Z. Yin and YC. Chen. Exploiting temporal stability and low-rank structure for localization in mobile networks. *ACM Proceedings of the sixteenth annual international conference on Mobile computing and networking (MobiCom)*, pp. 161-172, 2010.

[26] Vaswani, Namrata. Kalman filtered compressed sensing. *IEEE International Conference on Image Processing (ICIP)*, pp. 893-896, 2008.

[27] Yang, Zheng and Li, Mo and Liu, Yunhao. Sea depth measurement with restricted floating sensors. *IEEE International Conference on Real-Time Systems Symposium (RTSS)*, pp. 469-478, 2007.

[28] Zhang, Yin and Roughan, Matthew and Willinger, Walter and Qiu, Lili. Spatio-temporal compressive sensing and internet traffic matrices. *ACM SIGCOMM*, Vol.39, No. 4, pp. 267-278, 2009.