Multiple Attributes-based Data Recovery in Wireless Sensor Networks

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

***Abstract*—In wireless sensor networks (WSNs), since many basic scientific works heavily rely on the complete sensory data, data recovery is an indispensable operation against the data loss. Several works have studied the missing value 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 multiple attributes-based recovery algorithm which can provide high accuracy. Firstly, based on two real datasets, the Intel Indoor project and the GreenOrbs project, we reveal that such correlations are strong, *e.g.*, the change of temperature and light illumination usually has strong correlation. Secondly, motivated by this observation, we develop a Multi-Attribute-assistant Compressive-Sensing-based (MACS) algorithm to optimize the recovery accuracy. Finally, real trace-driven simulation is performed. The results show that MACS outperforms the existing solutions. Typically, MACS can recover all data with less than 5% error when the loss rate is less than 60%. Even when losing 85% data, all missing data can be estimated by MACS with less than 10% error.**

I. Introduction

Wireless sensor networks (WSNs) [4][6] are widely used to gather multiple attributes from the physical world and reconstruct environmental data in the cyber world [17]. Such data is significant for scientists to discover the physical world around. For instance, scientists reveal the plant evolution based on wind speed, air humidity and temperature data in the air [10], and predict the eruption by the temperature and shake data of volcano [12][9][18]. However, in WSNs, massive data loss is common, *e.g.*, 64% and 35% of the data are missing in the Ocean Sense project [21] and the GreenOrbs project [20], respectively. Hence, recovering these lost data with high accuracy is challenging.

The high loss rates veil the time and spatial correlations. Therefore classical interpolation methods, such as K-Nearest Neighbors (KNN) [16], cannot provide a satisfactory result due to the lack of one-hop neighbors. A recently proposed compressive sensing approach, the *Environmental Space Time Improved Compressive Sensing* (ESTI-CS) [8], 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.

We are aware of the following two facts. (1) Usually, WSNs gather multiple attributes simultaneously, *e.g.*, TelosB node [20] senses three attributes: temperature, light illumination and humidity. (2) Intuitively, one can expect that those attributes are correlated. For instance, when the sun is arising, the temperature and light illumination outdoor increase simultaneously. And the salinity of sea water also ties with the depth. The empirical study [11] reveals that temperature, dewpoint temperature and relative humidity have linear correlation. The correlations among attributes can be used as the supplement of the internal correlations and benefit the accuracy of the estimation. Hereby, our technical route is how to mine and exploit such correlations for the problem of missing data recovery.

To address this problem, firstly, we study the characteristics of real sensory data from the Intel Indoor project [7] and the GreenOrbs project [20]. The low-rank feature of attributes is revealed. And we propose a joint sparse decomposition method in order to find the cross features among multiple attributes. The energetic common part are found in the two correlated attributes. Secondly, we design an algorithm, named MACS, which can recover multi-attribute datasets jointly, using their correlation. Thirdly, we simulate the proposed approach on real data. We compare MACS with the classical and state-of-the-art methods such as KNN and ESTI-CS.

Our contributions are summarized as following:

- To the best of our knowledge, this is the first work to study the joint data recovery in WSNs.
- We design a novel algorithm, MACS, which is based on compressive sensing theory.
- Real trace driven simulations are performed extensively. The evaluation shows that MACS outperforms other compared solutions.

The rest parts of this paper are organized as following. In Section II, we present the related work. Section III shows the problem formulation. Section IV mines the internal and external features of attributes in WSNs. Section V proposes our approach, MACS. The performance is evaluated in Section VI. Section VII discusses the conclusion and future work.

II. Related Work

Lots of works have contributed in missing data interpolation. The most classic interpolation method is K-Nearest Neighbors (KNN) [16], which utilizes the average value of neighbors to estimate the missing data. This interpolation method performs well in situations where there is a moderate number of missing

978-1-4799-1353-4/13/$31.00 ©2013 IEEE

values. As the loss rate grows, the estimation error increases quickly due to the lack of one-hop neighbors.

Compressive Sensing (CS) [2][3] is currently an advanced and powerful technique for estimating massive missing data. There are a series of CS based solutions being used in different fields, *e.g.*, Distributed Compressive Sensing (DCS) [1][5] and Multi-Task Compressive Sensing (MTCS) [15] 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 [8]. ESTI-CS exploits the low-rank feature and spatial-temporal feature from the sensory data against the special loss patterns of WSNs. However, the low-rank and sparse features are also affected in the massive data loss scenario where the ESTI-CS experiences the increased estimation error.

All above methods aim at missing value estimation based on a single attribute. However, many physical attributes in nature have strong correlations such as humidity and temperature [11]. This work is to further improve the recovery accuracy exploiting such correlations. To the best of our knowledge, this is the first missing data recovery work using multiple attributes in WSNs.

III. Problem Formulation

A. Environment Data Recovery Problem

Suppose n nodes are deployed in an area, each of which equips k sensors to measure attributes. The monitoring period includes t time slots. The format of the data packet is as following:

Sensor ID	Time Stamp	Attribute 1	Attribute 2	...

Hereafter, let k attributes be denoted by M_i, $i = 1, 2, \cdots, k$. Each M_i is a $n \times t$ matrix. M_i is usually an incomplete matrix due to the data loss in WSNs. The available information about M_i is a sampled set of entries $(M_i)_{pq}$, $(p, q) \in \Omega_i$, where Ω_i is a subset of the complete set of entries in M_i. This process is represented by using a sampling operator $\mathcal{P}_\Omega(\cdot)$, which is defined as:

$$[\mathcal{P}_\Omega(X)]_{ij} = \begin{cases} X_{ij}, & (i,j) \in \Omega; \\ 0, & otherwise. \end{cases} \tag{1}$$

Therefore, the matrices we obtain are $\mathcal{P}_{\Omega_i}(M_i)$, $i = 1, \cdots, k$.

Our problem is to recover a series of matrices $M_1, \cdots, M_k$ (complete environmental data) from their sampled matrices $\mathcal{P}_{\Omega_i}(M_1), \cdots, \mathcal{P}_{\Omega_i}(M_k)$ (incomplete data gathered by WSN) as precisely as possible, so-called *Environment Data Recovery (EDR)* problem .

B. Problem Statement

Since we focus on exploiting the correlation among multiple attributes, multiple matrices are estimated jointly. For simplicity, in the most parts of this paper, we discuss the EDR problem under the situation of two attributes as an example. Our analysis and approach can be easily extended to the case of more attributes.

Formally, when $k = 2$, the problem is defined as follows:

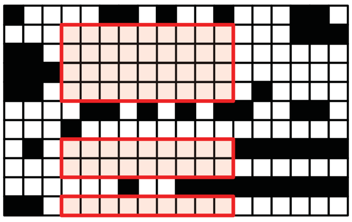


Fig. 1. Filter the original dataset by selecting the red parts to construct a small but completed dataset as the ground truth [8].

TABLE I
Selected Datasets As The Ground Truth

Data Name	Matrix Size	Time Interval
Intel Indoor	49 nodes × 149 intervals	1.5 minutes
GreenOrbs Temperature	281 nodes × 170 intervals	10 minutes

Give subsets of M_1, M_2 as $\mathcal{P}_{\Omega_1}(M_1)$, $\mathcal{P}_{\Omega_2}(M_2)$, find an optimal solution as $\hat{M}_1$ and $\hat{M}_2$, *i.e.*,

$$\begin{aligned} \text{minimize} \quad & ||\hat{M}_1 - M_1||_F + \mu||\hat{M}_2 - M_2||_F, \\ \text{subject to} \quad & \mathcal{P}_{\Omega_1}(\hat{M}_1) = \mathcal{P}_{\Omega_1}(M_1), \\ & \mathcal{P}_{\Omega_2}(\hat{M}_2) = \mathcal{P}_{\Omega_2}(M_2), \end{aligned} \tag{2}$$

where $||\cdot||_F$ represents the Frobenius norm, which is used in [19][8]. For instance, to a matrix $X = (x(i,j))_{p \times q}$, $||X||_F = \sqrt{\sum_{i,j}(x(i,j))^2}$. Because the magnitudes of attributes are not equal, which may cause one matrix overshadowing another, μ is used a tradeoff coefficient.

IV. Datasets in Sensor Networks

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. Ground Truth

The original datasets are gathered from two projects, GreenOrbs [20] and Intel Indoor [7]. After investigating the raw data, the loss rates of these two datasets are 35% and 23%, respectively. Hence, in order to obtain the ground truth, two small but completed datasets 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 dataset contains subsets of two attributes: temperature and light illumination, which share the same selecting entries.

B. Low-rank Structure

Consider the fact that the readings of nearby sensors are correlated and the readings in short time periods are close, we mine the inherent structure or redundancy of environment datasets.

The singular value decomposition (SVD) is adopted. The SVD of a $n \times t$ matrix X is:

$$X = U\Sigma V^T = \sum_{i=1}^{min(n,t)} \sigma_i u_i v_i^T, \tag{3}$$

where $\sigma_i \geq \sigma_{i+1}, i = 1, \cdots, min(n, t)$, $(\cdot)^T$ is the transpose operator, U is a $n \times n$ orthogonal matrix, V is a $t \times t$

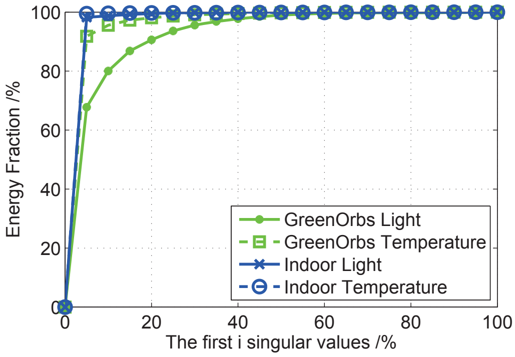


Fig. 2. The first 5% singular values contributes to over 90% total energy in GreenOrbs temperature and Indoor light/temperature. The number is 20% in GreenOrbs light.

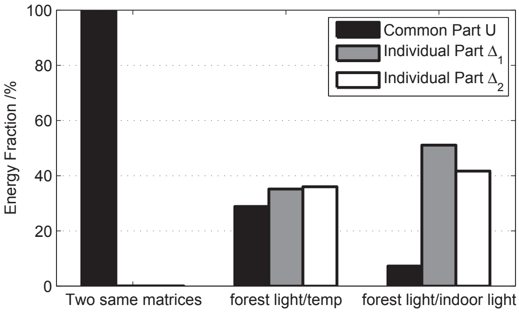


Fig. 3. Correlation analysis by joint sparse decomposition.

orthogonal matrix and Σ is a $n \times t$ diagonal matrix containing all singular values σ_i of X. Suppose $r = rank(X)$, so σ_i in $\Sigma = diag(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$.

The sum of all singular values represent the total energy of X. According to [19], if a matrix X is low-rank, the sum of its first r singular values occupy the total or near total energy, *i.e.*, $\sum_{i=1}^{r} \sigma_i \approx \sum_{i=1}^{min(n,t)} \sigma_i$. Fig.2 is a CDF to show the distribution of singular values' energy. The top 5% singular values contain all energy in Indoor temperature and Indoor light. The top 5% and 20% σ_i include 90% of energy in GreenOrbs temperature and GreenOrbs light, respectively. The above results show that rank minimization is suitable for our data recovery problem.

C. Inter-Correlation between Attributes

The relationship usually exists among natural attributes. For instance, the empirical study [11] reveals that temperature, dewpoint temperature and relative humidity have linear correlation under some special cases. However, in most cases, the correlations cannot be directly measured as a simple function.

In order to exploit the relationship of attributes in a WSN, we propose the *joint sparse decomposition (JSD)* to divide two matrices into a common part and two individual parts. Suppose $M_1 = (m_1^{(1)}, \cdots, m_1^{(t)})$ and $M_2 = (m_2^{(1)}, \cdots, m_2^{(t)})$. For both column vector $m_1^{(k)}$ and $m_2^{(k)}$, the goal is to split them, *i.e.*,

TABLE II
LOW-RANK FEATURES AFTER JOINT SPARSE DECOMPOSITION

Data Name	Matrix Name	XX% σ contain 90% Energy
Intel Indoor light/temperature	U	14%
	Δ_1	8%
	Δ_2	2%
GreenOrbs light/temperature	U	27%
	Δ_1	28%
	Δ_2	21%

$$\begin{aligned} m_1^{(k)} &= u^{(k)} + \delta_1^{(k)} \\ m_2^{(k)} &= u^{(k)} + \delta_2^{(k)} \\ u^{(k)} &= \Psi v^{(k)} \end{aligned} \tag{4}$$

where $u^{(k)}$ is the common part of $m_1^{(k)}$ and $m_2^{(k)}$, which is the multiplication of a certain basis Ψ (*e.g.*, a wavelet basis) and a sparse vector $v^{(k)}$. The individual parts are represented by $\delta_1^{(k)}, \delta_2^{(k)}$, respectively. Furthermore, Eqn.(4) is rewritten in a matrix formulation, *i.e.*,

$$\begin{bmatrix} m_1^{(k)} \\ m_2^{(k)} \end{bmatrix} = \begin{bmatrix} \Psi & I & 0 \\ \Psi & 0 & I \end{bmatrix} \begin{bmatrix} v^{(k)} \\ \delta_1^{(k)} \\ \delta_2^{(k)} \end{bmatrix} \tag{5}$$

According to compressive sensing theory [2][3], it is able to obtain $(v^{(k)T}, \delta_1^{(k)T}, \delta_2^{(k)T})^T$ by solving an l_1-norm minimization problem as following:

$$\hat{\vartheta} = \arg\min_{\vartheta} ||\vartheta||_1 \quad s.t.\ m = A\vartheta, \tag{6}$$

where $|| \cdot ||_1$ is the l_1-norm, $\vartheta = (v^{(k)T}, \delta_1^{(k)T}, \delta_2^{(k)T})^T$, $m = (m_1^{(k)T}, m_2^{(k)T})^T$ and $A = (\Psi, I, \mathbf{0}; \Psi, \mathbf{0}, I)$. And then $u^{(k)T}$, $\delta_1^{(k)T}$, and $\delta_2^{(k)T}$ are calculated from ϑ.

Applying JSD onto every column vector, M_1 and M_2 are decomposed, *i.e.*,

$$\begin{aligned} M_1 &= U + \Delta_1 \\ M_2 &= U + \Delta_2 \end{aligned} \tag{7}$$

Fig.3 shows that the common part occupies bigger ratio if two matrices have stronger correlation. For instance, when JSD is operated on two same matrices, which have definitely highest correlation, the common part contains 100% energy and individual parts $\Delta_1 = \Delta_2 = 0$. Intuitively, in an outdoor WSN, the sensory light and temperature have correlation. Fig.3 shows the common part of light/temperature in forest contains 29% of total energy, while the individual parts contain 35% and 36%. We also verify JSD on two irrelevant matrices. The common part of forest light and indoor light, where no relationship exists, contains only 7%.

In addition, the low rank feature of matrices after JSD is also revealed. As shown in TABLE.II, over 90% energy of attributes are contained in the first 30% singluar values. This means that the derived matrices of JSD still exhibit the low-rank feature.

V. OUR APPROACH

To address the EDR problem, we propose a novel relative data estimation approach named *Multi-Attribute-assistant Compressive Sensing (MACS)*, which is designed to jointly recover the attributes in a WSN.

A. Approach Design

Normalization: In Eqn.(2), the choice of μ has a significant effect on the accuracy of estimation. Since the relationship between M_1 and M_2 is unknown, it is difficult to find the best μ. To overcome the difficulty, a simple method is to normalize each matrix, and then set $\mu = 1$. The real maximum value is possible to loss, hence we adopt the maximum value in gathered datasets instead, *i.e.*, for each sensory matrix $\mathcal{P}_{\Omega_i}(M_i)$, $\max(\mathcal{P}_{\Omega_i}(M_i))$ is used on the normalization. This operation 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 in terms of the magnitude, *i.e.*, for $i = 1, 2$

$$\max(M_i) - \max(\mathcal{P}_{\Omega_i}(M_i)) \ll \max(M_i) \tag{8}$$

Low-Rank Matrix Approximation: Eqn.(2) contains the parameters M_1 and M_2, so the problem cannot be directly solved. However, since the low-rank features are revealed in Sec.IV, the problem is calculated by converting to a rank minimization problem. Through the inverse process of SVD, using k largest singular value of X, an optimal k-rank approximation [8] of X under the Frobenius norm $||\cdot||$ of errors can be obtained as $\hat{X} = \sum_1^k \sigma_i u_i v_i^T$. In our problem, this method is infeasible since we do not know the complete M_i its proper rank. However, it is reasonable to assume that estimated $\hat{M_i}$ is low-rank due to the low-rank feature of the original M_i. Thus the optimal M_i is evaluated by the problem: $min(rank(\hat{M_i})),\ s.t.\ \mathcal{P}_{\Omega_i}(M_i) = \mathcal{P}_{\Omega_i}(\hat{M_i})$.

Still two problems are up against us: (1) the rank calculating operator $rank(\cdot)$ is not convex. (2) there is no connection between M_1 and M_2.

To bypass the difficulty (1), we utilize SVD-like factorization [8] as $\hat{X} = LR^T$ where L is a $n \times k$ matrix and R is a $t \times k$ matrix, k is an approximation of the proper rank. According to the progress of the matrix compressive-sensing literature [14][19], rank minimization is exactly equivalent to the nuclear norm minimization when a certain technical condition holds on $\mathcal{P}_\Omega(\cdot)$ (the restricted isometry property [14]). Further, if the rank of X is less than the rank of LR^T, $min(rank(\hat{X}))$ is equivalent to $min(||L||_F^2 + ||R^T||_F^2)$.

Compressive Sensing-based Joint Matrix Decomposition: To overcome the difficulty (2), we need to find the correlation between M_1 and M_2, and then exploit it into finding an optimal solution.

Through the joint sparse decomposition proposed in Sec.IV, the correlation between two matrices can be revealed. Hence, we separate the approximation $\hat{M_1}$ and $\hat{M_2}$ by JSD as:

$$\begin{aligned} \hat{M_1} &= \hat{U} + \hat{\Delta}_1 \\ \hat{M_2} &= \hat{U} + \hat{\Delta}_2 \end{aligned} \tag{9}$$

Assume that $\hat{U}$, $\hat{\Delta}_1$ and $\hat{\Delta}_2$ are low-rank based on the low-rank structure analysis in Sec.IV. The problem is reformulated as:

$$\begin{aligned} \text{minimize} \quad & ||\hat{U}||_* + ||\hat{\Delta}_1||_* + ||\hat{\Delta}_2||_* \\ \text{subject to} \quad & \mathcal{P}_{\Omega_1}(\hat{U} + \hat{\Delta}_1) = \mathcal{P}_{\Omega_1}(M_1) \\ & \mathcal{P}_{\Omega_2}(\hat{U} + \hat{\Delta}_2) = \mathcal{P}_{\Omega_2}(M_2) \end{aligned} \tag{10}$$

where $||\cdot||_*$ is the nuclear norm which is defined as the sum of singular values, *e.g.*, $||X||_* = \sum_{i=1}^r \sigma_i(X)$.

Further more, by using SVD-like factorization into $\hat{U}$, $\hat{\Delta}_1$ and $\hat{\Delta}_2$, Eqn.(10) is rewritten as:

$$||L_U||_F^2 + ||R_U^T||_F^2 + ||L_1||_F^2 + ||R_1^T||_F^2 + ||L_2||_F^2 + ||R_2^T||_F^2 \tag{11}$$

where L_U, L_1, L_2 are $n \times r$ matrices and R_U, R_1, R_2 are $t \times r$ matrices. Moreover, $\hat{U} = L_U R_U^T$, $\hat{\Delta}_1 = L_1 R_1^T$ and $\hat{\Delta}_2 = L_2 R_2^T$.

To avoid overfitting, we convert the problem to a non-stationary optimization problem by using the Lagrange multiplier method, *i.e.*,

$$\begin{aligned} \text{minimize} \quad & ||\mathcal{P}_{\Omega_1}(L_U R_U^T + L_1 R_1^T) - S_1||_F^2 \\ + \quad & ||\mathcal{P}_{\Omega_2}(L_U R_U^T + L_2 R_2^T) - S_2||_F^2 \\ + \quad & \lambda(\sum_L ||L_j||_F^2 + \sum_R ||R_j||_F^2) \end{aligned} \tag{12}$$

where $S_1 = \mathcal{P}_{\Omega_1}(M_1)$ and $S_2 = \mathcal{P}_{\Omega_2}(M_2)$. The Lagrange multiplier λ allows a tunable tradeoff between the rank minimization and the accuracy fitness.

Eqn.(12) is solvable because (1) Ω_1, Ω_2, S_1 and S_2 are known, (2) each $||\cdot||_F^2$ is non-negative, (3) the optimal value can be reached by minimizing all non-negative parts to zero. Hence, $\hat{M_1}$ and $\hat{M_2}$ can be estimated by combining Eqn.(12) with Eqn.(9).

Extension: Our approach is also suitable for the case of more attributes. For instance, if three attributes is measured in one WSN, represented as M_1, M_2 and M_3. Similarly, rewrite Eqn.(10) as following:

$$||\hat{U}||_* + ||\hat{\Delta}_1||_* + ||\hat{\Delta}_2||_* + ||\hat{\Delta}_3||_* \tag{13}$$

where $i = 1,\ 2,\ 3$, $\hat{M_i} = \hat{U} + \hat{\Delta}_i$. This Equation can be solved by a similar method like Alg.1. In cases of more attributes, it is able to extend Eqn.(10) by this way.

B. Algorithm

To solve the estimation in the optimization problem defined by Eqn.(12), we propose an efficient algorithm. The detail of this algorithm is shown in Alg.1.

The algorithm solves the optimization by a iterative manner. First, all L and R matrices are initialized randomly except R_U. Fixing L_U, R_U can be calculated from other L and R matrices by solving the equation:

$$\begin{bmatrix} \mathcal{P}_{\Omega_1}(L_U R_U^T) \\ \mathcal{P}_{\Omega_2}(L_U R_U^T) \\ \sqrt{\lambda} R_U^T \end{bmatrix} = \begin{bmatrix} S_1 - L_1 R_1^T \\ S_2 - L_2 R_2^T \\ 0 \end{bmatrix} \tag{14}$$

this equation is solvable by calculating each line. Rewrite it as:

$$\begin{bmatrix} \mathcal{P}_{\Omega_1}(L_U)_{(i)} R_{U(i)}^T \\ \mathcal{P}_{\Omega_2}(L_U)_{(i)} R_{U(i)}^T \\ \sqrt{\lambda} R_{c(i)}^T \end{bmatrix} = \begin{bmatrix} (S_1 - L_1 R_1^T)_{(i)} \\ (S_2 - L_2 R_2^T)_{(i)} \\ 0 \end{bmatrix} \tag{15}$$

Algorithm 1 MACS Algorithm

```
Input:
    $\Omega_1$ and $\Omega_2$: sensory entry set
    $S_1$ and $S_2$: incomplete sensory data
    $r$: rank estimation of $M_1$ and $M_2$
    $\lambda$: tradeoff coefficient
    $k$: iteration times
Output:
    $\hat{M_1}$ and $\hat{M_2}$: estimated environment matrices
Main Procedure:
Normalization
$\alpha_1 \leftarrow max(S_1)$; $\alpha_2 \leftarrow max(S_2)$;
    $S_1 \leftarrow S_1./\alpha_1$; $S_2 \leftarrow S_2./\alpha_2$;
Approximation
$L_U \leftarrow rand(n,r)$; $L_1 \leftarrow rand(n,r)$; $L_2 \leftarrow rand(n,r)$;
$R_1 \leftarrow rand(r,t)$; $R_2 \leftarrow rand(r,t)$;
for 1 to k do
    $A_1 = S_1 - \mathcal{P}_{\Omega_1}(L_1 R_1^T); A_2 = S_2 - \mathcal{P}_{\Omega_2}(L_2 R_2^T);$
    $R_U \leftarrow crossInverse(\Omega_1, \Omega_2, L_U, \lambda, r, A_1, A_2)$
    $L_U \leftarrow crossInverse(\Omega_1^T, \Omega_2^T, R_U, \lambda, r, A_1^T, A_2^T)$
    $B_1 = S_1 - \mathcal{P}_{\Omega_1}(L_U R_U^T); B_2 = S_2 - \mathcal{P}_{\Omega_2}(L_U R_U^T);$
    $R_1 \leftarrow singleInverse(\Omega_1, L_1, \lambda, r, B_1)$
    $L_1 \leftarrow singleInverse(\Omega_1^T, R_1, \lambda, r, B_1^T)$
    $R_2 \leftarrow singleInverse(\Omega_2, L_2, \lambda, r, B_2)$
    $L_2 \leftarrow singleInverse(\Omega_2^T, R_2, \lambda, r, B_2^T)$
    $v \leftarrow Eqn.(12)$
    if $v < \hat{v}$ then
        $\hat{L_U} \leftarrow L_U; \hat{R_U} \leftarrow R_U; \hat{L_1} \leftarrow L_1; \hat{R_1} \leftarrow R_1; \hat{L_2} \leftarrow$
        $L_2; \hat{R_2} \leftarrow R_2; \hat{v} \leftarrow v;$
    end if
end for
$\hat{M_1} \leftarrow \alpha_1(\hat{L_U}\hat{R_U}^T + \hat{L_1}\hat{R_1}^T)$
$\hat{M_2} \leftarrow \alpha_2(\hat{L_U}\hat{R_U}^T + \hat{L_2}\hat{R_2}^T)$
Procedure $Y = singleInverse(\Omega, L, \lambda, r, S)$:
for i=1 to t do
    $P_i \leftarrow [\mathcal{P}_\Omega(L)(:,i); \sqrt{\lambda} * I_r]$
    $Q_i \leftarrow [S(:,i); \mathbf{0}_r]$
    $Y(:,i) = (P_i^T * P_i) \backslash (P_i^T * Q_i)$
end for
Procedure $Y = crossInverse(\Omega_1, \Omega_2, L, \lambda, r, S_1, S_2)$:
for i=1 to t do
    $P_i \leftarrow [\mathcal{P}_{\Omega_1}(L)(:,i); \mathcal{P}_{\Omega_2}(L)(:,i); \sqrt{\lambda} * I_r]$
    $Q_i \leftarrow [S_1(:,i); S_2(:,i); \mathbf{0}_r]$
    $Y(:,i) = (P_i^T * P_i) \backslash (P_i^T * Q_i)$
end for
```

where i ranges from 1 to t. Eqn.(15) can be treated as a linear least square problem. R_U can be obtained by the inverse procedure given in Alg.1 as $crossInverse$. And L_U can be computed using the same procedure by fixing R_U.

L_i and R_i are obtained by a similar procedure, which is defined in $singleInverse$ as the pseudo code.

Moreover, in Alg.1, the rank approximation r and the lagrange tradeoff coefficient λ are significant influential in the accuracy of estimation. Hence, λ is tuned by the method in [13]. And our evaluation uses $r = 20\% min(n,t)$, since 20% singular values contributes to over 90% energy in the datasets.

The complexity of the algorithm is $O(rntk)$. Because the key operation in Alg.1 is the inverse computation, whose complexity is $O(nrt)$ [8], and the algorithm iterates k times.

VI. Performance Evaluation

A. Methodology

Performance evaluation is based on real-trace driven simulation.

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

Compared Methods: To verify the effectiveness of our approach, two methods for missing data recovery in WSNs are chosen for comparison. They are the classical interpolation method, K-Nearest Neighbor (KNN) [16], and the state-of-the-art method, Environmental Space Time Improved Compressive Sensing (ESTI-CS) [8].

Metric: To compare results evaluated from different matrices, the error rate of approximation under the Frobenius norm, $err(\hat{M}, M, \Omega)$, is applied [13], which is defined as:

$$err(\hat{M}, M, \Omega) = \frac{||\mathcal{P}_{\overline{\Omega}}(M) - \mathcal{P}_{\overline{\Omega}}(\hat{M})||_F^2}{||\mathcal{P}_{\overline{\Omega}}(M)||_F^2} \tag{16}$$

where $\overline{\Omega}$ is the complementary set of Ω.

Procedure: The procedure of simulation is as following:

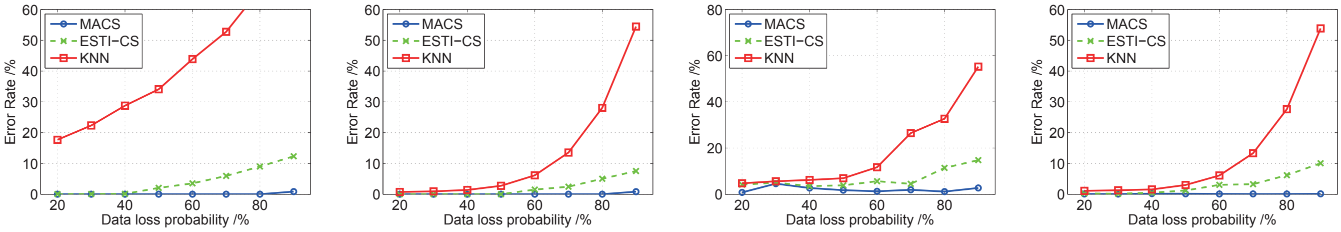
1) Randomly lose the data from the ground truth to simulate the gathered data in WSNs. Generate the subset Ω from the random loss pattern and then set $\Omega_1 = \Omega_2 = \Omega$. The quantity of data loss is from 20% to 90%.
2) Using datasets of two physical attributes, compute $\mathcal{P}_\Omega(M_1)$ and $\mathcal{P}_\Omega(M_2)$.
3) $\mathcal{P}_\Omega(M_1)$ and $\mathcal{P}_\Omega(M_2)$ serve as the inputs of the estimation algorithms including KNN, ESTI-CS, and MACS, separately or together. Then we obtain the approximations of M_1 and M_2 as $\hat{M_1}$ and $\hat{M_2}$.
4) Compare the performance of the algorithms on the error rate defined by Eqn.(16).

B. Simulation Results

In Fig.4, we plot the comparison result of three algorithms in the case of two attributes. According to the simulation, MACS can obtain less than 5% error rate under the loss rate less than 60%, where ESTI-CS can provide 10% and KNN performs worse. Even in high loss rate (80%), the error rate of MACS is still less than 10%. The main reason is that MACS uses the correlation between two attributes. Hence, the accuracy of estimating missing values increases if the correlation exists. And even when there are no relation between two attributes, the performance of MACS 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.4, the estimation performance of KNN is barely satisfactory and reduces quickly as the increasing of



(a) Indoor Light (b) Indoor Temp (c) GreenOrbs Light (d) GreenOrbs Temp

Fig. 4. The accuracy of missing value estimation methods.

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

VII. Conclusion

In this paper, we studied the Environment Data Recovery Problem in WSNs. We proposed the joint sparse decomposition to reveal the correlation among multiple attributes. The low-rank feature was exhibited by both the original and the JSD derived data. Driven by these observations, we designed the MACS algorithm to approximate the missing data. The algorithm combines the benefits of compressive sensing and the correlation of attributes. Data-driven simulations illustrated that MACS outperforms existing interpolation methods.

The future works are as following. First, considering to use the Bayesian Model into the prediction and the data reconstruction. Second, studying the relationship between the computation time and the accuracy. 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] D. Baron, M.B. Wakin, M.F. Duarte, S. Sarvotham, and R.G. Baraniuk. "Distributed compressed sensing", Preprint, 2005.

[2] D.L. Donoho. "Compressed sensing", *IEEE Transactions on Information Theory*, Vol. 4, No. 4, pp. 1289-1306, 2006.

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

[4] F.L. Lewis. "Wireless sensor networks", *Smart Environments: Technologies, Protocols, and Applications*, pp. 11-46, 2004.

[5] G. Chen, X.-Y. Liu, L. Kong, J.-L. Lu, W. Shu, and M.-Y. Wu. "JSSDR: Joint-Sparse Sensory Data Recovery in Wireless Sensor Networks", *IEEE WiMob*, 2013.

[6] IF. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. "Wireless sensor networks: a survey", *Elsevier Computer Networks*, Vol. 38, No. 4, pp. 393-422, 2002.

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

[8] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu. "Data loss and reconstruction in sensor networks", *IEEE INFOCOM*, pp. 1701-1710, 2013.

[9] L. Kong, M. Zhao, X.-Y. Liu, J.-L. Lu, Y. Liu, M.-Y. Wu, and W. Shu. "Surface coverage in sensor networks", *IEEE Transactions on Parallel and Distributed Systems*, 2013.

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

[11] M.G. Lawrence. "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, No. 2, pp. 225-233, 2005.

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

[13] S. Rallapalli, L. Qiu, Y. Zhang and YC. Chen. "Exploiting temporal stability and low-rank structure for localization in mobile networks", *ACM MobiCom*, pp. 161-172, 2010.

[14] B. Recht, M. Fazel, and PA. Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization", *SIAM*, Vol. 52, No. 3, pp. 471-501, 2010.

[15] S. Ji, D. Dunson, and L. Carin. "Multitask compressive sensing", *IEEE Transactions on Signal Processing*, Vol. 57, No. 1, pp. 92-106, 2009.

[16] T. Cover and P. Hart. "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27, 1967.

[17] T. He, S. Krishnamurthy, J.A. Stankovic, T. Abdelzaher, L. Luo, R. Stoleru, T. Yan, L. Gu, J. Hui, and B. Krogh. "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.

[18] X.-Y. Liu, K. Wu, Y. Zhu, L. Kong, and M.-Y. Wu. "Mobility increases the surface coverage of distributed sensor networks", *Elsevier Computer Networks*, 2013.

[19] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu. "Spatio-temporal compressive sensing and internet traffic matrices", *ACM SIGCOMM*, Vol. 39, No. 4, pp. 267-278, 2009.

[20] Y. Liu, Y. He, M. Li, J. Wang, K. Liu, L. Mo, W. Dong, Z. Yang, M. Xi, and J. Zhao. "Does wireless sensor network scale? A measurement study on greenorbs", *IEEE INFOCOM*, pp. 873-881, 2011.

[21] Z. Yang, M. Li, and Y. Liu. "Sea depth measurement with restricted floating sensors", *IEEE RTSS*, pp. 469-478, 2007.