Sparsity-based Online Missing Sensor Data Recovery

Di Guo, Xiaobo Qu, Lianfen Huang, Yan Yao
Dept. Communicaition Engineering
Xiamen University
Xiamen, China
{guodi, quxiaobo, lfhuang}@xmu.edu.cn

Zicheng Liu
Microsoft Research
One Microsoft Way
Redmond, USA
zliu@microsoft.com

Ming-Ting Sun
Dept. Electrical Engineering
University of Washington
Seattle, USA
sun@ee.washington.edu

***Abstract*—In sensor networks, due to power outage at a sensor node, hardware dysfunction, or bad environmental conditions, not all sensor samples can be successfully gathered at the sink. Additionally, in the data stream scenario, some nodes may continually miss samples for a period of time. In this paper, a sparsity-based online data recovery approach is proposed. We construct an overcomplete dictionary composed of past data frames and traditional fixed transform bases. Assuming the current frame can be sparsely represented using only a few elements of the dictionary, missing samples in each frame can be estimated by Basis Pursuit. Our method was tested on data from a real sensor network application: monitoring the temperatures of the disk drive racks at a data center. Simulations show that in terms of estimation accuracy and stability, the proposed approach outperforms existing average-based interpolation methods, and is more robust to burst missing along the time dimension.**

I. INTRODUCTION

Wireless sensor networks are characterized by a dense deployment of sensor nodes that continuously observe a physical phenomenon, such as environmental sensing, habitat monitoring and other emergency cases [1-3]. These distributed sensors collaboratively relay their data to a single sink (base station). Some transmitted sensor data may be lost or corrupted due to power outrage at a sensor node, hardware dysfunction, or bad environmental conditions. Many real-time applications, such as traffic and safety control, and healthcare [4] need to operate on continuous data streams. In this paper, we consider a 2-D (two dimensional) data stream scenario, and the missing data of each 2-D frame need to be estimated at the sink online with low time delay. Fig. 1 shows a sensor network with missing samples in time intervals n-1, n, and n+1, respectively.

Traditional methods interpolating the missing data include inverse distance weighted averaging (IDWA) [5] and Kriging [6]. However, they only consider data within a single frame, and do not take advantage of information in the sequential data frames. Others take temporal factors into consideration. For instance, the work in [7] is restricted to Markov models, where the samples at time interval n+1 are independent of those for any time earlier than n. This assumption is too restrictive.

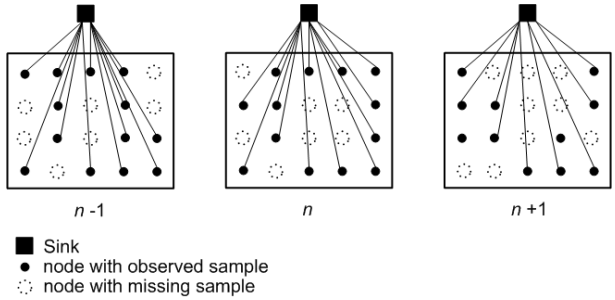


Figure 1. A sensor network with missing samples.

Recently, sparse representation approximates a signal with a linear combination of a small number of elementary signals called atoms [8]. Guo *et al.* [9] adopted a 2-D Discrete Cosine Transform (DCT) basis to sparsely represent the realistic climate sensor data, but there is no guarantee that general transforms such as DCT, or wavelets can sparsely represent the signal of interest [10].

For the online data, temporal correlation between the current and past frames usually exists. This property provides us the opportunity to explore the past frames to represent the current frame. In this paper, we propose to construct an overcomplete dictionary composed of past-data frames and the 2-D DCT basis for the online data recovery. This dictionary is more effective than the fixed DCT basis, because the current frame can be efficiently represented by a few weighted linear combination of previous past frames with the correlation property. No off-line training phase is required. The recovery approach is simple enough to be implemented on the sink, with negligible delay compared to the sampling interval of the sensors.

Our methods were tested on the data from a real sensor network application: monitoring the temperatures of the disk drive racks in a data center. Simulation shows that in terms of estimation accuracy and stability, the proposed approach outperforms existing average-based interpolation methods, and is more robust to burst missing along the time dimension.

This work was supported by Tsinghua-Qualcomm Joint Research Program, Fundamental Research Funds for the Central Universities (No. 2011121050), and National Natural Science Foundation of China (No. 61001142).

II. PROPOSED METHOD

Considering a network with M sensor nodes, each node records the physical parameters of an environment at time intervals $1,2,\cdots,n,\cdots$. Samples of all the nodes can be arranged in a vector $\mathbf{f}_n = \left[f_1(n), f_2(n), \cdots, f_M(n)\right]^T$ to form the network data (the n^{th} frame). However, if some of the samples failed to be collected or transmitted to the sink due to hardware failure or environmental limitations, only a subset of $\mathbf{f}_n$ is observed.

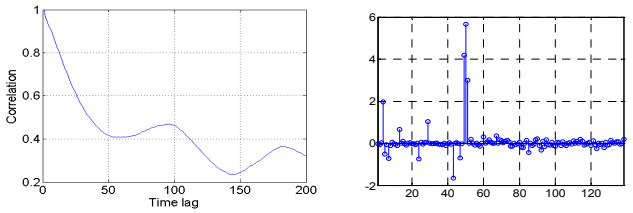
To address this issue, we propose a Sparsity-based online data Recovery method using an Overcomplete Dictionary (SROD). We can group the indices of the entries into two subsets: Λ_n consists of those indices of entries observed in $\mathbf{f}_n$; $\bar{\Lambda}_n$ consists of those indices of entries missed in $\mathbf{f}_n$. Correspondingly, $\mathbf{f}_n^{\Lambda_n}$ and $\mathbf{f}_n^{\bar{\Lambda}_n}$ are denoted as the available data and missing data in $\mathbf{f}_n$, respectively.

Since we have past frames available at the sink, and with the temporal correlation, the current frame could be efficiently represented by a few weighted linear combination of previous past frames, we propose to add some past frames to the dictionary besides the 2-D DCT basis, resulting in an overcomplete dictionary for online data recovery.

We assume the data in the frames are highly temporally correlated which is the case in most sensor network applications, since the sampling rate usually is controlled so that the sensor data do not change drastically within the sampling period. We verify this assumption by plotting the temporal correlation between 200 frames in our temperature dataset [11] as shown in Fig.2. The temporal correlation is defined as:

$$R(\tau) = \frac{1}{\Delta n + 1} \sum_{n=n_0}^{n_0+\Delta n} \mathbf{z}_n^T \mathbf{z}_{n+\tau} \tag{1}$$

where $\mathbf{z}_n$ is a normalized vector containing the sensor data in the n^{th} frame and $\Delta n+1$ is the total number of frame pairs used in calculating $R(\tau)$. This temporal correlation averages the inner products for all the pairs of frames with time lag τ. The high correlation of the frames shown in Fig. 2(a) indicates the feasibility of their sparse representation. We verify the sparse representation of frames in our case by Fig. 2(b), where only a few large coefficients exist when representing the current frame using our proposed dictionary.



(a) (b)

Figure 2. (a) Data correlation between frames. (b) Coefficients when representing a frame using the proposed overcomplete dictionary.

To get started, we assume there is no missing data or the missing data have been recovered in the past L frames. Let $\mathbf{\Psi}$ denote the DCT basis, then the overcomplete dictionary for $\mathbf{f}_n$ is $\mathbf{\Phi}_n = \left[\mathbf{f}_{n-L} \cdots \mathbf{f}_{n-2}\ \mathbf{f}_{n-1}\ \mathbf{\Psi}\right]$, where $\mathbf{\Phi}_n$ is an $M \times (L+M)$ matrix with $\text{rank}(\mathbf{\Phi}_n) = M$, so that any signal can be represented by more than one combination of different atoms. We assume that the current frame can be represented as a sparse linear combination of the atoms in $\mathbf{\Phi}_n$,

$$\mathbf{f}_n = \mathbf{\Phi}_n \boldsymbol{\alpha}_n \tag{2}$$

where $\boldsymbol{\alpha}_n \in \mathbb{R}^{(L+M)}$ is expected to be sparse, i.e. $\|\boldsymbol{\alpha}_n\|_0 \ll L+M$.

With the notation of the available data $\mathbf{f}_n^{\Lambda_n}$ and the missing data $\mathbf{f}_n^{\bar{\Lambda}_n}$, the rows of $\mathbf{\Phi}_n$ can also be partitioned into two parts $\mathbf{\Phi}_n^{\Lambda_n}$ and $\mathbf{\Phi}_n^{\bar{\Lambda}_n}$ correspondingly. Thus, the current frame to be recovered in Eq. (1) can be rewritten as,

$$\begin{pmatrix} \mathbf{f}_n^{\Lambda_n} \\ \mathbf{f}_n^{\bar{\Lambda}_n} \end{pmatrix} = \begin{pmatrix} \mathbf{\Phi}_n^{\Lambda_n} \\ \mathbf{\Phi}_n^{\bar{\Lambda}_n} \end{pmatrix} \boldsymbol{\alpha}_n \tag{3}$$

Since $\mathbf{f}_n^{\bar{\Lambda}_n}$ is not known, we are unable to make any use of $\mathbf{f}_n^{\bar{\Lambda}_n} = \mathbf{\Phi}_n^{\bar{\Lambda}_n} \boldsymbol{\alpha}_n$. Our hope for finding $\boldsymbol{\alpha}_n$ relies on the equation corresponding to the available data,

$$\mathbf{f}_n^{\Lambda_n} = \mathbf{\Phi}_n^{\Lambda_n} \boldsymbol{\alpha}_n \tag{4}$$

The number of unknowns is more than the number of equations in (4), thus the system of equations is under-determined. Since we expect a priori that the presentation of the current frame will be sparse, $\boldsymbol{\alpha}_n$ can be estimated by solving the ℓ_1 norm optimization problem:

$$\arg\min_{\boldsymbol{\alpha}_n} \|\boldsymbol{\alpha}_n\|_1 \quad \text{s.t.} \quad \mathbf{f}_n^{\Lambda_n} = \mathbf{\Phi}_n^{\Lambda_n} \boldsymbol{\alpha}_n, \tag{5}$$

as long as $\mathbf{\Phi}_n^{\Lambda_n}$ satisfies the Restricted Isometry Property (RIP) [10]. In other words, among all the solutions that satisfy the constraints, we select the one that has the smallest ℓ_1 norm, i.e., the sparsest solution. One appealing method for solving (5) is Basis Pursuit Denoising (BPDN) [8],

$$\hat{\boldsymbol{\alpha}}_n = \arg\min_{\boldsymbol{\alpha}_n} \frac{1}{2} \left\| \mathbf{f}_n^{\Lambda_n} - \mathbf{\Phi}_n^{\Lambda_n} \boldsymbol{\alpha}_n \right\|_2^2 + \lambda \|\boldsymbol{\alpha}_n\|_1 \tag{6}$$

This solution is robust in the presence of noise, and also gives good performance even when the coefficient vector is not as sparse, which means $\mathbf{f}_n$ can be approximated with some error by truncating the small magnitude coefficients in $\boldsymbol{\alpha}_n$. The final recovered output $\mathbf{A}_n$ is,

$$\mathbf{A}_n = \mathbf{\Phi}_n \hat{\boldsymbol{\alpha}}_n \tag{7}$$

III. RESULTS

A. Dataset and preprocessing

The dataset used in simulations comes from the Microsoft Research Data Center Genome (DC Genome) system [11]. To

get an idea of the scenario, Fig. 3 shows that the temperature across the racks and across different heights of the same rack varies significantly [11].

This dataset was recorded by sensors deployed in an 8×11 grid over a one-day period. Since about 10% of samples in the original dataset are missing, we use 2-D *K*-Nearest Neighbor (KNN) spatial interpolation algorithm [8], one of IDWA interpolation methods, to fill in these missing samples before simulation, so that we have a complete dataset as the ground truth. Then, we generate random and burst missing data patterns for 451st frame to 700th frame, and apply our proposed data recovery scheme to estimate the missing data.

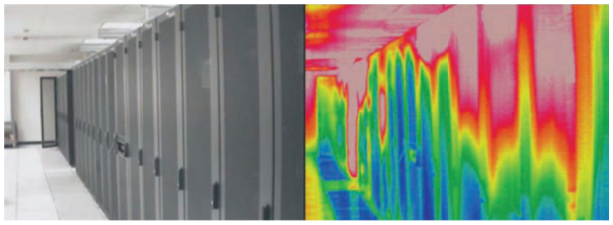


Figure 3. The thermal image of an aisle in a data center. The infrared thermal image shows significant variations on intake air temperature across racks and at different height [11].

B. Simulation setup

In the simulations, to extend 2-D *KNN* to 3-D *KNN* for our online data recovery scenario, the neighbors can only be chosen from the current frames and the past frames. Let (x, y) be the spatial location of a node, and n be the frame index. Adapting to anisotropic spatial and temporal correlation, we use a parameter η as the weighting of the temporal correlation relative to the spatial correlation, thus the distance is computed as,

$$d = \sqrt{(x-x_0)^2 + (y-y_0)^2 + \eta(n-n_0)^2} \tag{8}$$

where (x_0, y_0, n_0) and (x, y, n) are the coordinates of a node with missing sample and a neighboring node, respectively.

In order to evaluate the accuracy and stability of 3-D *KNN* and the proposed methods, we use Root Mean Squared Error (RMSE) and Maximal Absolute Error (MAE) calculated over all missing entries. Suppose $\mathbf{f}$ and $\hat{\mathbf{f}}$ are the original and recovered data vectors, respectively, and I is the total number of missing samples in $\mathbf{f}$. The RMSE is defined as,

$$\mathrm{RMSE}(\mathbf{f}, \hat{\mathbf{f}}) = \sqrt{\frac{\sum_{i=1}^{I} (f_i - \hat{f}_i)^2}{I}}, \quad i = 1, 2, \cdots, I, \tag{9}$$

where f_i and $\hat{f}_i$ stand for ith missing entry of $\mathbf{f}$ and $\hat{\mathbf{f}}$, respectively. Besides, the MAE of all missing entries is used to assess the estimation stability. The MAE is defined as,

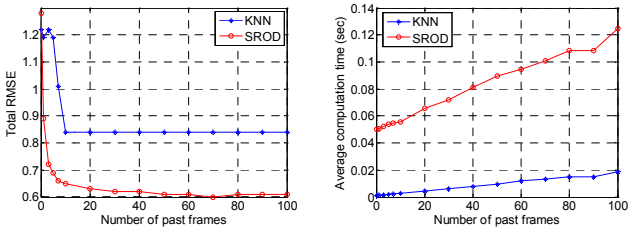
$$\mathrm{MAE}(\mathbf{f}, \hat{\mathbf{f}}) = \max\left[\left\{\left|f_i - \hat{f}_i\right|\right\}\right] \text{ for } i = 1, 2, \cdots, I, \tag{10}$$

Specifically, we choose four evaluation metrics, including: (a) MAE of all nodes in each frame, (b) MAE of all frames in each node, (c) RMSE of all nodes in each frame, and (d) RMSE of all frames in each node. They evaluate the accuracy and stability of the methods node-by-node and frame-by-frame.

To solve the ℓ_1 norm minimization in (8), the "Sparselab 2.1" toolbox [12] is used. The parameters are $\lambda = 1 \times 10^{-3}$, $\eta = 0.1$, $K = 9$. The size of each frame is 8×11. We generate burst missing patterns, i.e., the same node continuously missing samples along the temporal dimension, and the duration of time is defined as the *burst missing length*. In the simulations, we choose different missing rates (u) and burst missing lengths (v), and compare the above four evaluation metrics of 3-D *KNN* and the proposed approach.

C. Performance comparison

First, we discuss how the performance and complexity of *SROD* change with respect to the number of previous frames in the overcomplete dictionary. Compared with pure DCT basis, adding past frames into the dictionary can dramatically reduce the recovery error. This recovery error can be further reduced by increasing the number of past frames. However, further reduction of recovery error is not obvious when the number of past frame exceeds a threshold, as shown in Fig. 4(a). On the other hand, increasing the number of past frames will increase the computation time [see Fig. 4(b)]. In our observations, using 50 past frames is a reasonable choice to tradeoff the recovery error and the computation time. This is the default number of past frames in the following simulations. Fig. 5 shows an overcomplete dictionary, which is composed of 50 past 8x11 2-D frames and an 8x11 2-D DCT basis.



(a) (b)

Figure 4. Recovery error and computation time of *KNN* and *SROD* when u = 20% and v = 10. (a) Recovery error with respect to the numbers of past frames in the dictionary. (b) Computation time with respect to the numbers of past frames in the dictionary.

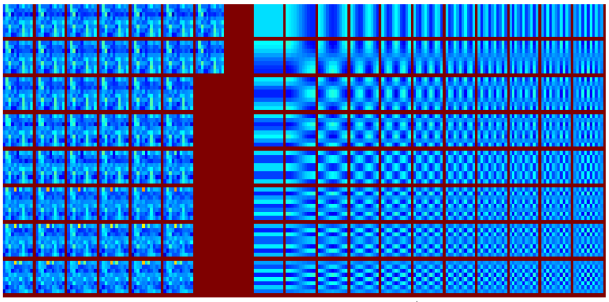


Figure 5. The overcomplete dictionary for the 451st frame. Left half contains 50 past data frames which change over time, and each frame is with mean zero and normalized to 1. Right half is the fixed 8x11 2-D DCT basis.

With this dictionary, Table I shows the performance of the *SROD* and *KNN*, by summarizing the mean values of the above performance metrics for all the 451st ~700th recovered frames. The missing rate u = 10% and 20%, the burst missing length v = 5 and 10. The proposed approach outperforms *KNN* in terms of these evaluation metrics with more than 20% improvement. The mean error values of *KNN* and *SROD* both increase as the missing rate and missing burst length go up, but the error of the proposed method is still much lower than that of *KNN*. The maximal error values demonstrate that the proposed approach is more robust to different missing rates and missing burst lengths.

TABLE I. PERFORMANCE COMPARISON OF *KNN* AND *SROD*

Methods / Mean	*KNN*				*SROD*			
	10%		20%		10%		20%	
	5	10	5	10	5	10	5	10
MAE_frame	1.31	1.40	1.54	1.88	0.88 (32.8%)	1.11 (20.7%)	1.19 (22.7%)	1.49 (20.7%)
MAE_node	1.48	1.48	1.75	1.80	1.06 (28.4%)	1.21 (18.2%)	1.39 (20.6%)	1.50 (16.7%)
RMSE_frame	0.66	0.69	0.66	0.78	0.43 (34.8%)	0.53 (19.7%)	0.47 (28.8%)	0.58 (25.6%)
RMSE_node	0.68	0.67	0.69	0.77	0.43 (36.8%)	0.52 (23.5%)	0.47 (31.9%)	0.56 (27.3%)
Total RMSE	**0.76**	**0.75**	**0.73**	**0.84**	**0.47 (38.2%)**	**0.57 (25.0%)**	**0.51 (30.1%)**	**0.63 (25.0%)**

From 451st to 700th frame. u = 10% and 20%, v = 5 and 10. The numbers between brackets are percentage improvement of the evaluation criteria of the *SROD* relative to that of *KNN*, at the same u and v.

D. Impact of noise

In reality, the data are usually corrupted with noise. We add white Gaussian noise to the available data, and the noisy measurement vector is written as

$$\mathbf{y} = \mathbf{f}_n^{\Lambda_n} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\boldsymbol{\varepsilon}$ is the Gaussian noise, whose power is controlled by signal-to-noise ratio (SNR) defined as

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\frac{\left\|\mathbf{f}_n^{\Lambda_n}\right\|_2^2}{\left\|\boldsymbol{\varepsilon}\right\|_2^2}, \tag{12}$$

and then use BPDN algorithm

$$\hat{\boldsymbol{\alpha}}_n = \arg\min_{\boldsymbol{\alpha}_n}\frac{1}{2}\left\|\mathbf{y}-\boldsymbol{\Phi}_n^{\Lambda_n}\boldsymbol{\alpha}_n\right\|_2^2 + \lambda\left\|\boldsymbol{\alpha}_n\right\|_1 . \tag{13}$$

Fig. 6 shows the estimation error of *SROD* under different noise levels. As the noise increases, i.e., SNR drops, the curves of total RMSE basically remain the same, until SNR is as low as 10dB. This demonstrates that *SROD* is robust to low and medium noise.

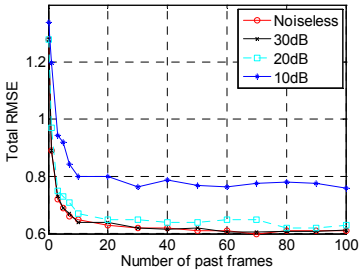


Figure 6. Total RMSE of *SROD* for noisy data. u = 20%, v = 10.

IV. CONCLUSIONS

We presented a new approach for online data recovery in sensor networks. A sparse linear relationship between a current frame and its past frames are modeled to estimate the missing data in the current frame. We design an overcomplete dictionary composed of the past data and the DCT basis to sparsely represent the current frame. Data recovery is achieved through ℓ_1 norm minimization. Simulation results on a real sensor data set demonstrate that the proposed approaches outperform the average-based interpolation methods in terms of both accuracy and robustness. In the future, we plan to leverage available data in the next frame to correct the current frame, and when delay is acceptable, more future frames can be incorporated to further improve the recovery accuracy.

ACKNOWLEDGMENT

The authors would like to thank Dr. Jie Liu in Microsoft providing the real sensor dataset. D. Guo and X. Qu would like to thank China Scholarship Council for financial support. D. Guo thanks Qirong Ma for discussion.

REFERENCES

[1] T. Mitchell. (1999, March 27). "50" km resolution daily precipitation for the Pacific Northwest, 1949–94 [Online]. Available: http://www.jisao.washington.edu/data/widmann/

[2] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin, "Habitat monitoring with sensor networks," *Communications of the ACM*, vol. 47, no. 6, pp. 34–40, 2004.

[3] K. Lorincz, D. J. Malan, T. R. F. Fulford-Jones, A. Nawoj, A. Clavel, V. Shnayder, G. Mainland, M. Welsh, and S. Moulton, "Sensor networks for emergency response: Challenges and opportunities," *Pervasive Comput.*, vol. 3, no. 4, pp. 16–23, 2004.

[4] A. Gaddam, S.C. Mukhopadhyay, G.S. Gupta, "Elder Care Based on Cognitive Sensor Network," *IEEE Sensors J.*, vol. 11, no. 3, pp. 574–581, Mar. 2011.

[5] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers & Geosciences*, vol. 34, no.9, pp. 1044–1055, 2008.

[6] M. Umer, L. Kulik, and E. Tanin, "Kriging for Localized Spatial Interpolation in Sensor Networks," in *Scientific and Statistical Database Management*. vol. 5069, B. Ludäscher and N. Mamoulis, Eds. New York: Springer Berlin / Heidelberg, 2008, pp. 525-532.

[7] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *Proc. 30th int. conf. Very Large Data Bases* 2004, pp. 588-599.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129-159, 2001.

[9] D. Guo, X. Qu, L. Huang, and Y. Yao, "Sparsity-Based Spatial Interpolation in Wireless Sensor Networks," *Sensors*, vol. 11, no.3, pp. 2385–2407, 2011.

[10] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York: Springer, 2010, pp. 227–228.

[11] J. Liu, F. Zhao, J. O'Reilly, A. Souarez, M. Manos, C. J. M. Liang, and A. Tersiz. (2008, December). Project genome: Wireless sensor network for data center cooling. *The Architecture Journal* [Online], vol. 18, pp. 28–34. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=78813

[12] D. Donoho, Stodden, V., Tsaig, Y. (2007, May 26). Sparselab 2.1 [Online]. Available: http://sparselab.stanford.edu/.