International Journal of Fuzzy Logic and Intelligent Systems, vol. 8, no. 2, June 2008 pp. 126-131

Recursive PCA-based Remote Sensor Data Management System Applicable to Sensor Network

Sung-Ho Kim* and Yui-Su Youk

Dept. of Electronics and Information Engineering, Kunsan National University
Kunsan, Korea
Email : shkim@kunsan.ac.kr
sixofnum@hotmail.com

Abstract

Wireless Sensor Network(WSNs) consists of small sensor nodes with sensing, computation, and wireless communication capabilities. It has new information collection scheme and monitoring solution for a variety of applications. Faults occurring to sensor nodes are common due to the limited resources and the harsh environment where the sensor nodes are deployed. In order to ensure the network quality of service it is necessary for the WSN to be able to detect the faulty sensors and take necessary actions for the reconstruction of the lost sensor data caused by fault as earlier as possible. In this paper, we propose an recursive PCA-based fault detection and lost data reconstruction algorithm for sensor networks. Also, the performance of proposed scheme was verified with simulation studies.

Key Words : Wireless Sensor Network(WSNs), lost data reconstruction, PCA(Principal Component Analysis), Recursive-PCA, remote sensor data management system

1. Introduction

The dramatic advances in wireless communication and electronics have enabled the development of low cost, low power and multifunctional wireless sensor nodes which are equipped with sensor, data processing and communication components. These tiny sensor nodes can be easily deployed into a designated area to form a wireless network and perform specific functions. Wireless sensor network has become a new information collection and monitoring solution for a variety of applications, such as environment and habitat monitoring, disaster management and emergency response[1-4].

Due to the low cost and the deployment of a large number of sensor nodes in a harsh or hostile environment, it is common for the sensor nodes to become faulty. The networks must detect the occurrence of faulty sensors and take any actions for the reconstruction of the lost sensor data caused by the fault as earlier as possible to ensure the network quality of service[5].

Principle component analysis(PCA) is a well-known statistical technique that has been widely applied to solve important signal processing problems like feature extraction, signal estimation and detection[6-10]. Many analytical techniques exit, which can solve PCA once the entire input data is known. However, most of the analytical methods require extensive matrix operations and hence they are not suited for real-time applications.

In this paper, we propose a PCA-based fault detection and lost data reconstruction algorithm for a sensor network that may include frequent occurrence of faulty sensor nodes. For its better performance, we utilize the recursive-PCA algorithm for the automatic adaptation of constantly changing measurement environment.

The paper is organized as follows. We first review the sensor network architecture in section 2. Then, we explain the basic concept of the proposed scheme in section 3. The proposed fault detection and reconstruction algorithm based on recursive PCA is illustrated in section 4. A simulation study and performance analysis of the proposed scheme is reported in section 5. Finally, we conclude the paper.

2. Routing Protocols of Wireless Sensor Network

In general, routing in WSNs can be divided into flat-based routing, hierarchical-based routing, and location-based routing depending on the network structure. In flat-based routing, all nodes are typically assigned equal roles or functionality. In hierarchical-based routing, nodes will play different roles in the network. In location-based routing, sensor nodes' positions are exploited to route data in the network.

Hierarchical or cluster-based routing are well-known techniques with special advantages related to scalability and efficient communication. In a hierarchical architecture, higher energy nodes can be used to perform the sensing in the proximity of the target. Hierarchical routing is an efficient

Manuscript received Feb. 13, 2008; revised Jun. 10, 2008.
*Corresponding author
This research was financially supported by the Ministry of Commerce, Industry and Energy(MOCIE) and Kore Industrial Technology Foundation(KOTEF) through the Human Resource Training Project for Regional Innovation.

way to lower energy consumption within a cluster and by performing data aggregation and fusion in order to decrease the number of transmitted messages to the BaseStation(BS).

Heinzelman, et. al. introduced a hierarchical clustering algorithm for sensor network called Low Adaptive Clustering Hierarchy(LEACH)[9]. LEACH is a cluster-based protocol, which includes distributed cluster formation. LEACH randomly selects a few sensor nodes as cluster-heads(CHs) and rotate this role to evenly distribute the energy load among the sensors in the network. In LEACH, the clusterhead nodes compress data arriving from nodes that belong to the respective cluster, and send an aggregated packet to the base station in order to reduce the amount of information that must be transmitted to the base station. LEACH uses a TDMA/CDMA MAC to reduce inter-cluster and intra-cluster collisions. However, data collection is centralized and is performed periodically. Therefore, it is most appropriate when there is a need for constant monitoring by the sensor network. A user may not need all the data immediately. Hence, periodic data transmissions are unnecessary which may drain the limited energy of the sensor nodes. After a given interval of time, a randomized rotation of the role of clusterhead is conducted so that uniform energy dissipation in the sensor network is obtained.

The operation of LEACH is separated into two phases, the setup phase and the steady state phase. In the setup phase, the clusters are organized and CHs are selected. In the steady state phase, the actual data transfer to the base station takes place. During the steady state phase, the sensor nodes can begin sensing and transmitting data to the cluster-heads. The cluster-head node, after receiving all the data, aggregates it before sending it to the base-station. After a certain time, which is determined a priori, the network goes back to the setup phase again and enters another round of selecting new CH.

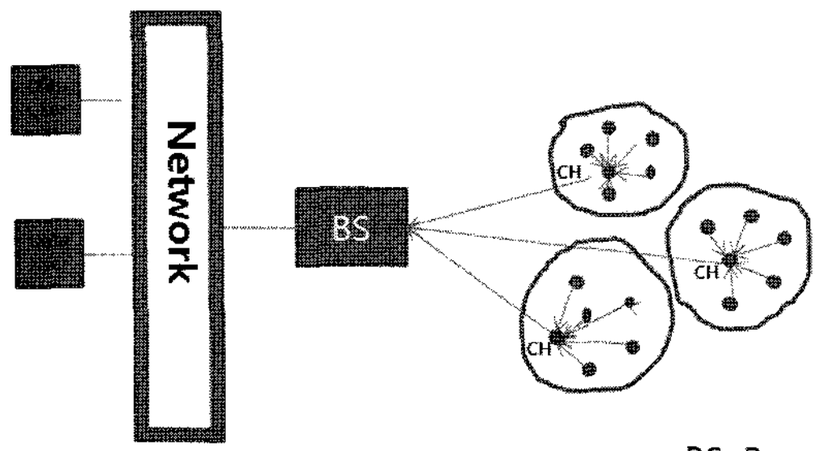


Fig. 1. Operation of LEACH Algorithm

Generally, sensor nodes deployed in the wide area are likely to be faulty. Therefore, the congregated data in BS are not complete owing to the faulty sensor node which can not transmit the sensed data to BS. It may deteriorate the performance of monitoring system using WSN. Therefore, some kind of scheme should be developed to overcome these problems.

3. PCA

The PCA has been widely used in statistical data analysis and pattern recognition[10]. The basic idea of PCA is to derive new variables (principle components arranged in descending order of importance) that are linear combinations of the original variables and are uncorrelated to each other. Principle components are obtained by projecting the multivariate data vectors on the space spanned by the eigenvectors of the covariance matrix of the original data set. One of the advantages of PCA is its ability to describe the data using a small group of underlying variables while preserving as much of the relevant information as possible in the dimensionality reduction process.

3.1 PCA for Data Reconstruction

In PCA we are concerned with finding the latent vector space that explains the greatest amount of variability in a single matrix of data. Let us consider, x(k), a set of m variable.

$$x(k) = [x_1(k) \quad x_2(k) \quad \cdots \quad x_m(k)] \tag{1}$$

Assume that the measurement data are given in a matrix X (n x m) where n represents the number of samples. X is approximated as

$$X \cong \hat{X} = \sum_{i=1}^{k} t_i \cdot p_i^T \tag{2}$$

where p_i is the first k largest eigenvectors $(p_1 \cdot p_k)$ of the covariance matrix of X and the scores t_i are linear combinations of the measurement data and are defined by

$$t_i = X \cdot p_i \tag{3}$$

By applying PCA on matrix X, the information contained in X can be summarized by a lower dimensional score space defined by the principal components and reduce the dimension of the analysis space. Then, identification of the process model by PCA consists in determining eigenvalues and eigenvectors of matrix $X^T \cdot X$. $x(k) \in R^m$ is decomposed into

$$x(k) = x(k)P \cdot P^T + x(k)\tilde{P} \cdot \tilde{P}^T \tag{4}$$

where $P \in R^{mXl}$ are eigenvectors corresponding to the principal eigenvalues $\lambda_1 \geqq, \cdots, \geqq \lambda_l$ of the correlation matrix of $x(k)$ and $\tilde{P}$ are eigenvectors corresponding to the remaining eigenvalues $\lambda_{l+1} \geqq, \cdots, \geqq \lambda_m$. Then, the reconstructed part of $x(k)$ is given by the following expression

$$\hat{x}(k) = x(k)P \cdot P^T = C \cdot x(k) \tag{5}$$

And the residual part of $x(k)$ is given by

$$\tilde{x}(k) = x(k)\tilde{P} \cdot \tilde{P}^T = (I - C)x(k) = \tilde{C} \cdot x(k) \tag{6}$$

consider a set of data in which variable x_i is supposed to be faulty. $x_i(k)$ is arbitrarily set to zero.

$$x_{mi}(k) = [x_1(k)\ x_2(k)\ \cdots\ x_{i-1}(k)\ 0\ x_{i+1}(k)\ \cdots\ x_m(k)] \quad (7)$$

Let $e(k)$ be the reconstruction error defined as

$$e(k) = x(k) - \hat{x}(k) \quad (8)$$

which gives

$$e(k) = x_{mi}\tilde{C} + x_i(k)\tilde{C}_i \quad (9)$$

Then the faulty value can be reconstructed as follows

$$x_i(k) = -\frac{x_{mi}(k) \cdot \tilde{p}\tilde{p}^T \cdot (\tilde{p}_i\tilde{p}_i^T)^T}{\tilde{p}_i\tilde{p}_i^T \cdot (\tilde{p}_i\tilde{p}_i^T)^T} \quad (10)$$

The detection index is the Squared Prediction Error(SPE) defined by

$$d(k) = e(k)^T \cdot e(k) \quad (11)$$

As you can see eq.(11), it is possible to detect the occurrence of sensor fault in the sensor nodes. Furthermore, it is possible to reconstruct the missing data caused by the faulty sensors by using eq.(10). However, it is required to know the eigenvectors of the measurement data X as precisely as possible for the efficient missing data reconstruction.

3.2 Recursive PCA

PCA assumes that data are stationary, which means that there's no change in mean value and covariance. However, this is not the case in general.

Generally, the change in the eigenvalues represents the change in the operational status of the stochastic system. If the stochastic system is assumed to be stationary with zero mean, covariance matrix is constant. Therefore, the changes in the original stochastic system can be easily detected just by monitoring the eigenvalues of the covariance matrix.

Observation of a significant change in the eigenvalues could indicate changes in the following quantities: input excitation, system properties, or noise level, which might be of interest.

Generally, in case of sensor network environment the input measurement data are acquired one at a time, which necessitates sample-by-sample update of the covariance and its eigenvectors which can be required to reconstruct the missing measurement data.

An advanced technique for estimating missing data by using PCA should be developed in the case of non-stationary cases. This problem can be overcome by use of an recursive PCA. The PCA model is continuously updated using an exponential memory. Then, the model has to adapt to slow modification of the process operating condition. At each iteration the covariance matrix is updated recursively as follows (Dayal and MacGregor 1997)

$$X^T \cdot X = \alpha X^T \cdot X(k-1) + (\alpha - 1)x^T(k) \cdot x(k) \quad (12)$$

Where α is a forgetting factor. The value of α is adjusted according to the dynamics of the system.

We can get the recursive formula which can be applied to estimating missing data by using eigenvectors. Eigenvectors are obtained by covariance matrix expressed in eq.(12). So we can always estimate the missing data which undergoes change in system statistics.

4. Fault Detection and Data Reconstruction Algorithm Based on RPCA

There are plenty of potential applications for intelligent sensor networks: distributed information gathering and processing, monitoring, supervision of hazardous environments, cooperative sensing. The ever-increasing use of sensing units asks for the development of specific data mining architectures. What is expected from these architecture is not only accurate modelling of high dimensional streams of data but also a minimization of the communication and computational effort demanded to each single sensor unit.

The general approach to the analysis of sensor network makes use of a centralized architecture where sensor readings from all the sensors are gathered at Base Station as in fig. 1.

Generally, sensor data in the BS are transmitted to remote server computer for further data processing such as extraction of the higher-level information from the raw data and detection of faulty sensor nodes. If we assume that reasonable-size sensor network is made of thousands of nodes, the limitation of this approach is evident: 1) Transmitting large-size sensor data to the remote server computer via an internet puts a high demand on its communication bandwidth, which is especially true for wireless LAN. 2) Remote data users who log on the server computer might be interested in retrieving sensor data without loss of generality even though the occurrence of some faulty sensor nodes.

In order to facilitate efficient transmission of data to remote sever and perfect sensor data monitoring in the case of the faulty sensor nodes, innovative data management systems are highly desired. For this, in this section, a RPCA-based data management system which is shown in fig. 2 is proposed. The detailed explanation is given briefly.

The BS is equipped with following functions. For the reconstruction of missing data caused by faulty sensors and efficient data compression for the transmission to remote server, Recursive PCA algorithm is adopted. At every step, eigenvalues and eigenvectors for the covariance matrix are calculated. During this process, if the SPE expressed in eq.(11) exceeds the predefined value, it automatically detects the missing data caused by the faulty sensor nodes. And then, the missing data are reconstructed by using eq.(10). Furthermore, if there is an request from the remote server to send the stored data to remote server, data compressed by PCA algorithm are transmitted to remote server. In the remote server computer, the original data can be effectively reconstructed by using transmitted eigenvectors and the compressed data.

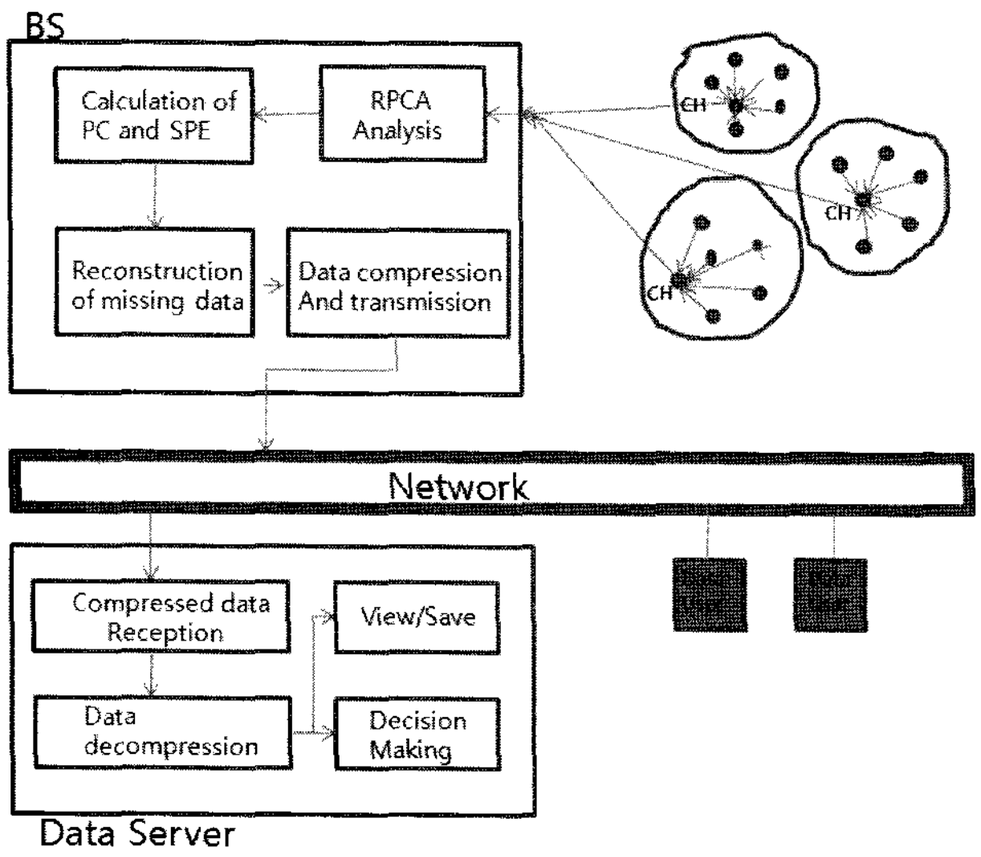


Fig. 2. Proposed data management system

5. Simulation Study

In this section, the simulation studies are carried out to verify the feasibility of the proposed method. The sensor data used for the simulation are obtained from sixteen sensor nodes which are installed on an inclined plane. Each sensor nodes are equipped with temperature sensor. A total of 1,788 sampling points were taken from each sensor nodes. All changes in the temperature is shown in fig. 3. The verification of proposed method was performed using measured sensor data and Matlab program and a discussion of the result is given in the following subsections.

5.1 Calculation of PC and SPE

As discussed earlier, occurrence of missing data caused by faulty sensor nodes can be detected through monitoring SPE calculated by eq.(11). The detected changes in SPE would decide when the reconstruction of missing data starts. To verify the feasibility of the detection of missing data, we assume that fifth sensor node is faulty during the 1,108th and 1,443th step, and eleventh sensor is faulty during 1,001th and 1,210th step. Sensor data comprising missing data is plotted in fig. 4.

Fig. 5 shows missing data in fifth and eleventh sensor node and its changes in SPE which is calculated by using principle components obtained from PCA algorithm.

5.2 Reconstruction of Missing Data

The reconstruction of missing data caused by the faulty sensor noses is conducted after the faulty sensor has been detected by monitoring SPE. Fig. 6 shows the reconstruction of missing data. The blue line in fig. 6 shows the original data and the red line is the reconstructed data by using eq.(10). A we can see in fig. 6, the reconstructed data follows quite fairly the fault free sensor values.

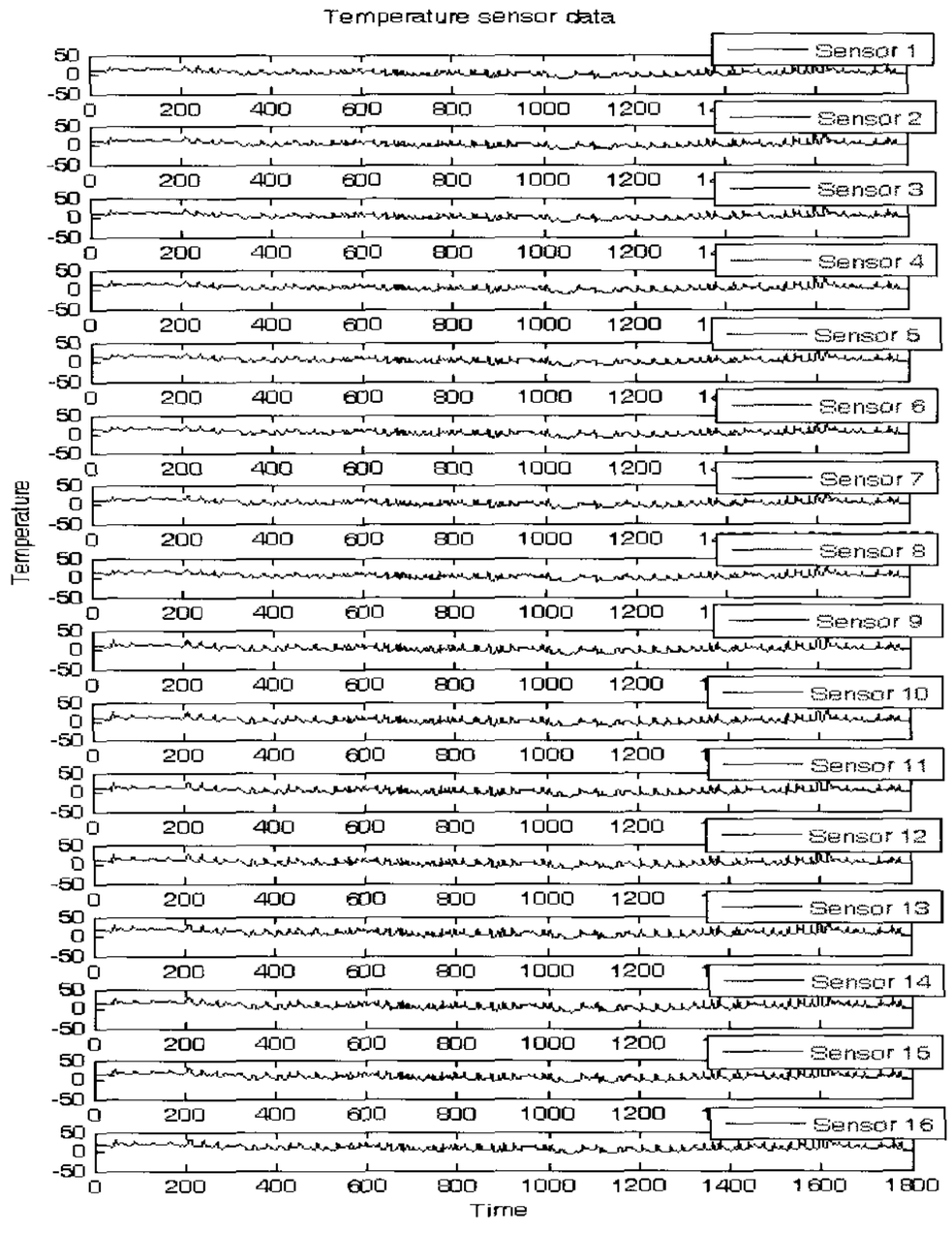


Fig. 3. Temperature sensor data obtained from sixteen sensor nodes

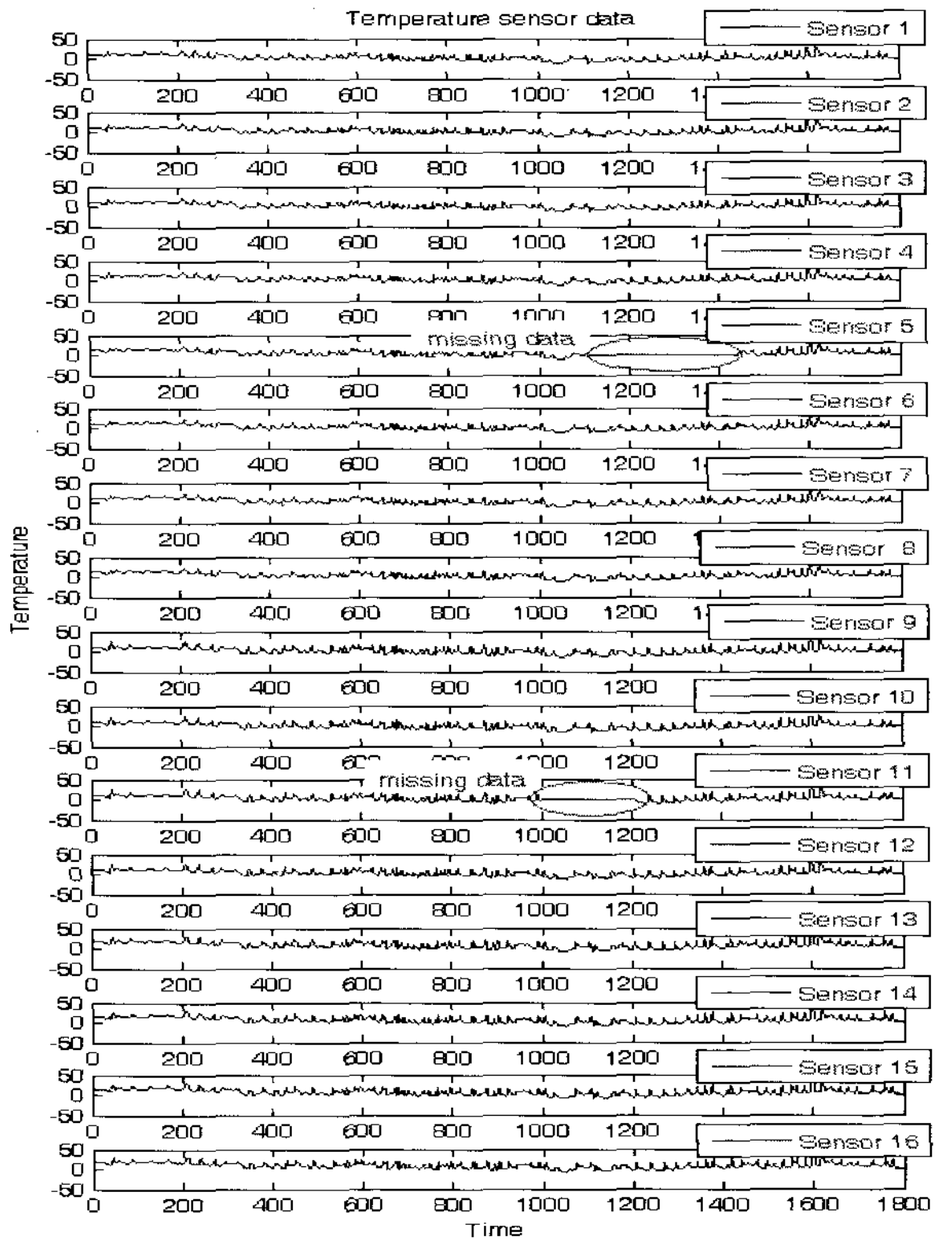


Fig. 4. Temperature sensor data comprising missing data

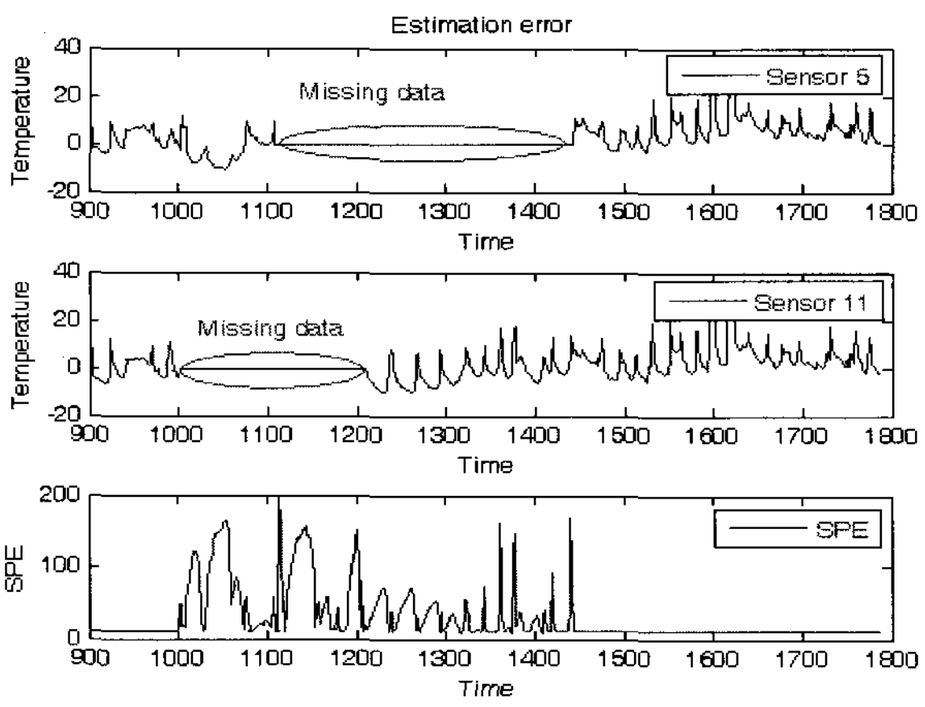
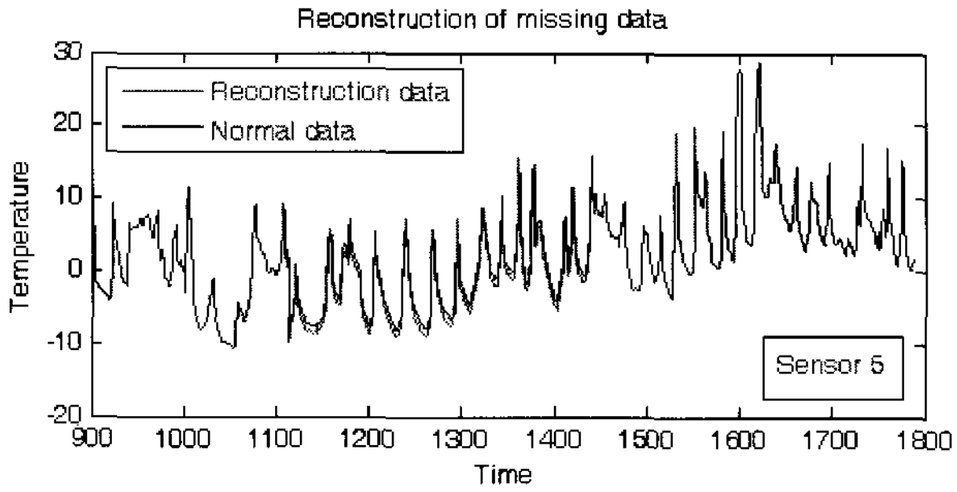


Fig. 5. Missing data in fifth and eleventh sensor node and its SPE



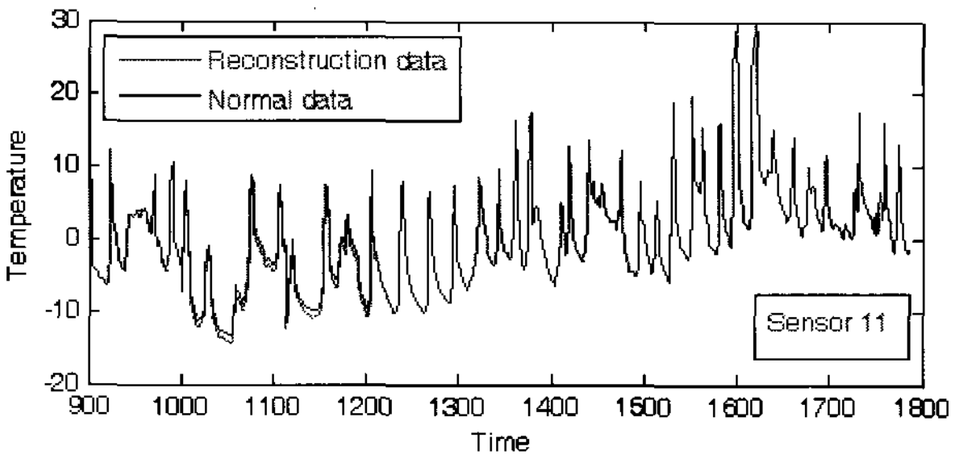


Fig. 6. Reconstruction of missing data

5.3 Compression and Decompression of Senor Data

As mentioned in section 4, gathering of sensor data and reconstruction of missing data are performed in BS. Futhermore, all the data collected in BS should be transmitted to remote sever computer for better information distribution. For efficient data transmission, transmitted data should be compressed by using eigenvectors which are obtained from RPCA algorithm. In this case stuy, we take five eigenvectors as principle components to compress the transmitted data. Fig. 7 shows the loading vectors which should be transmitted to remote server computer along with its corresponding five eigenvectors.

Decompression process should be taken in the remote server computer to get the original data from the transmitted loading vectors and the eigenvectors. Fig. 8 shows the reconstructed data in the remote server computer.

Fig. 9 shows the reconstruction error. As we can see the fig. 9, the reconstructed error remains within a reasonable boundary.

From the results of Fig. 6 and Fig. 9, we could confirm to reconstruct error of missing data and reconstruct error in the remote server computer. And we could confirm that the original data and the reconstruction data is almost similar.

Fig. 10. compares reconstruction error using PCA with reconstruction error using R-PCA. From the result of Fig. 10, we could confirm the excellent performance of R-PCA.

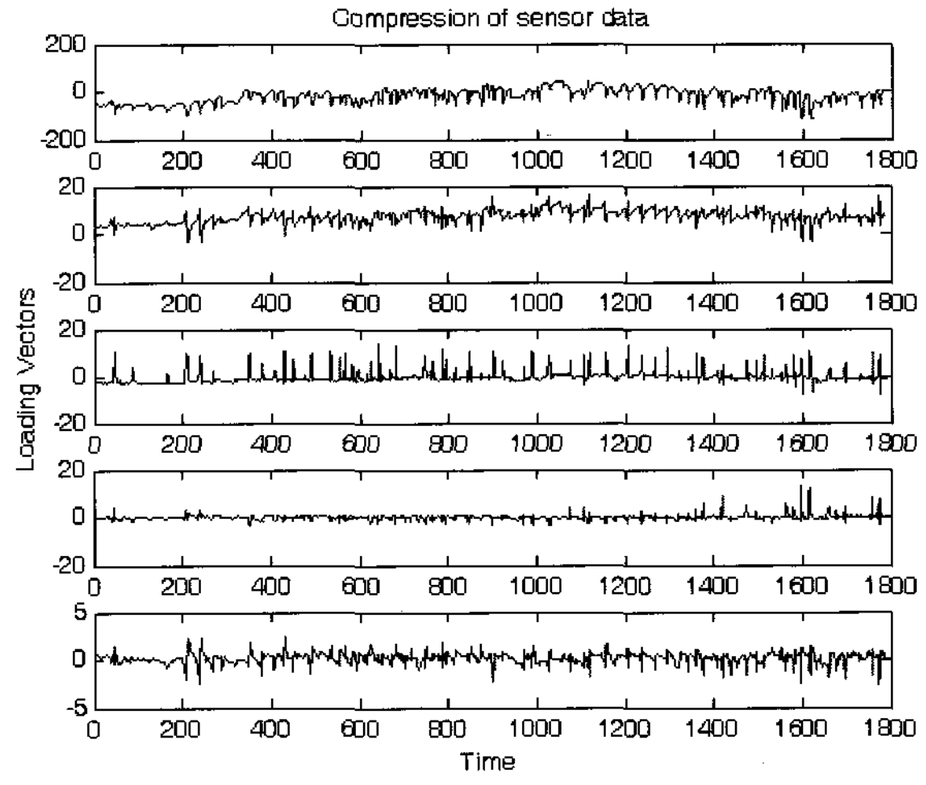


Fig. 7. Compressed data which should be transmitted to remote server computer

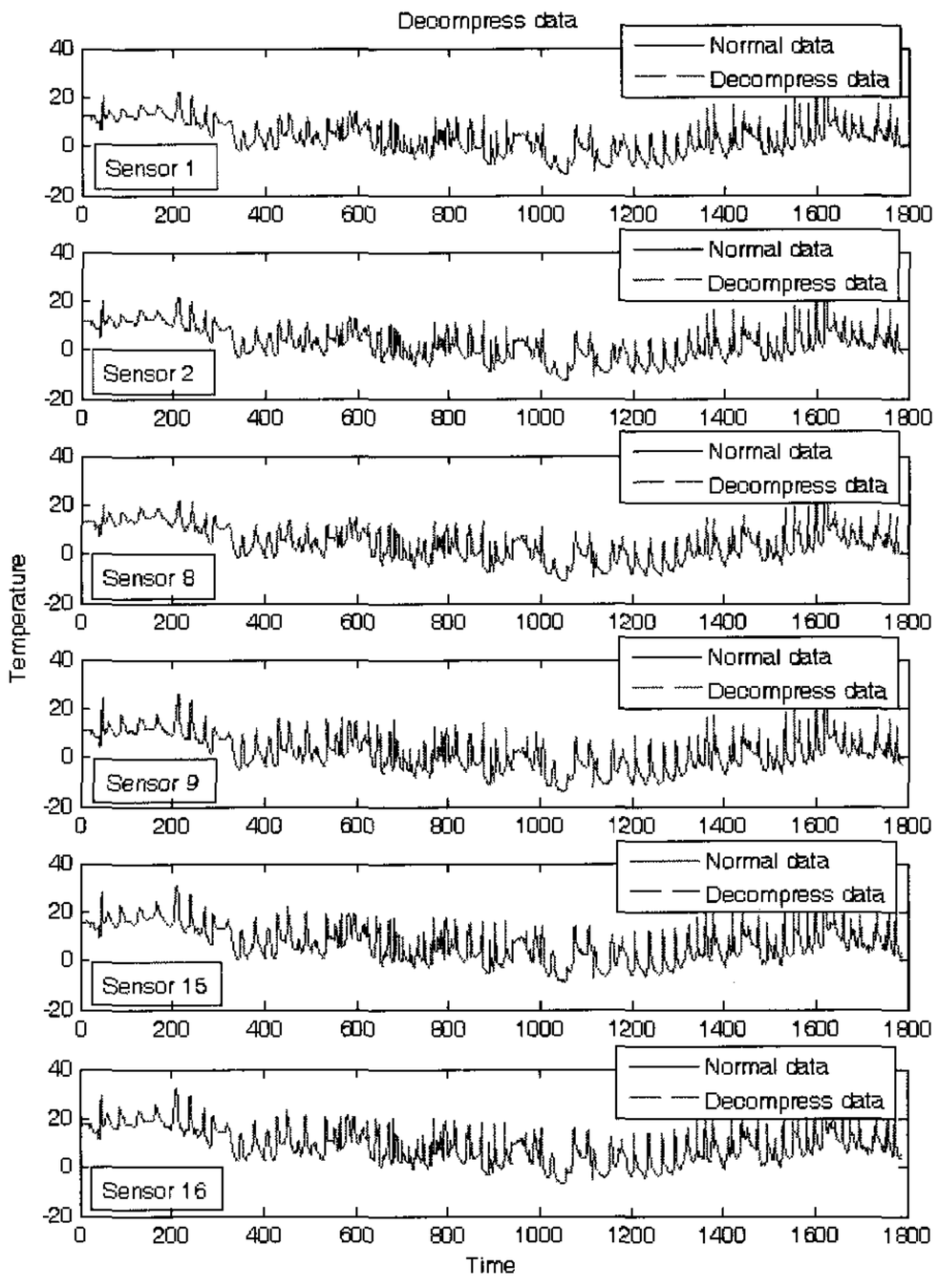


Fig. 8. Reconstructed original data in the server computer

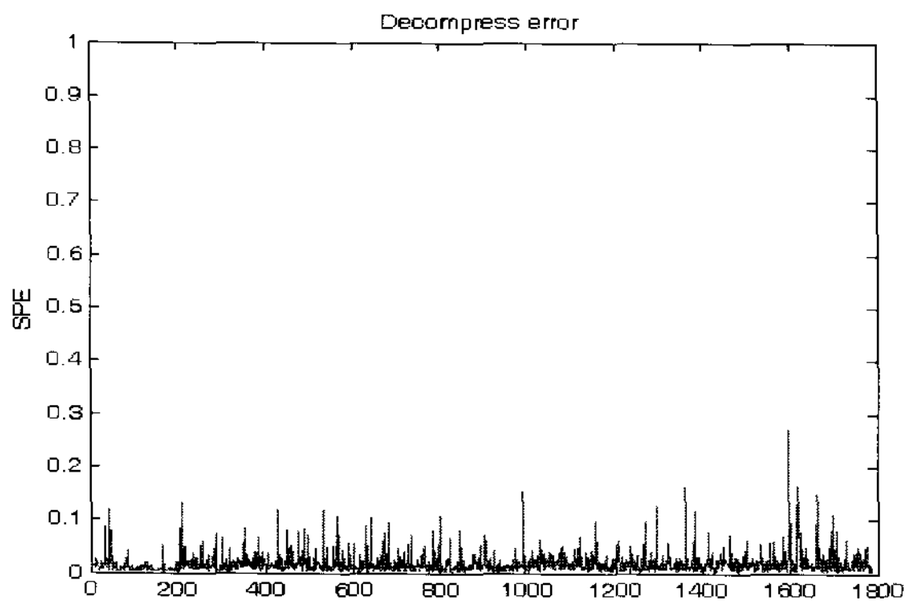


Fig. 9. Reconstruction error in the remote server computer

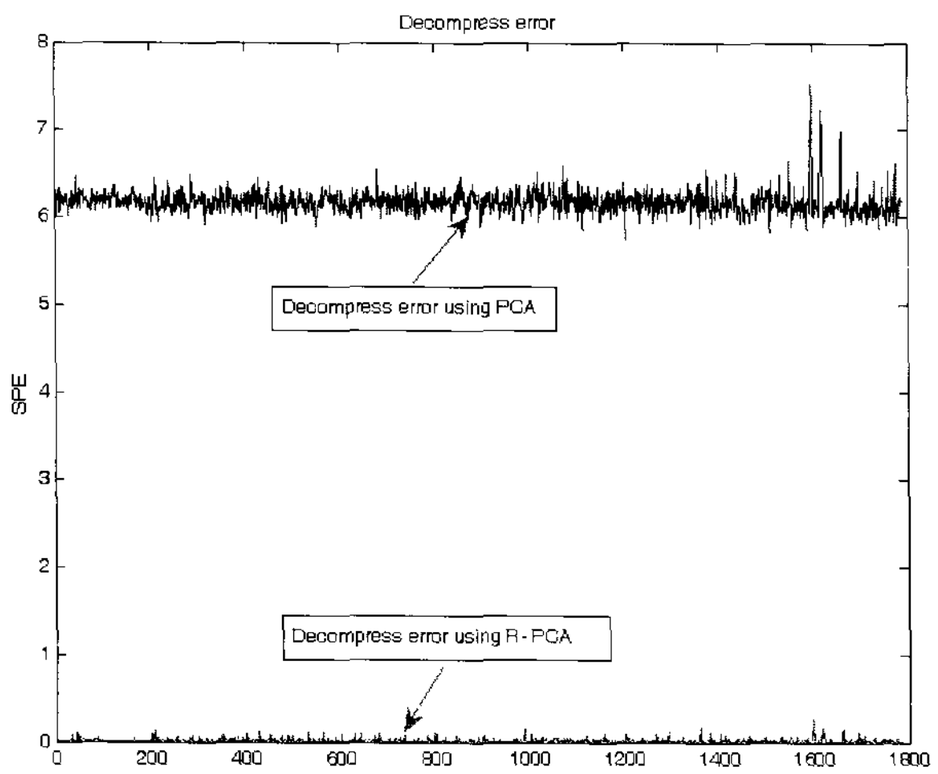


Fig. 10. Comparison of decompress error using PCA and R-PCA

6. Conclusions

In this paper, we proposed a recursive PCA- based fault detection and missing data reconstruction algorithm which can be applicable to sensor networks that may include frequent occurrence of faulty sensor nodes. The proposed scheme utilizes SPE obtained from PCA as a detection index of faulty sensor nodes. Furthermore, missing data are reconstructed by using Principle Components which are calculated from recursive PCA algorithm. To verify the applicability of the proposed scheme to sensor network, several simulation studies were carried out. With the simulation result, it is verified that the proposed scheme can successfully detect the occurrence of faulty sensor nodes and reconstruct the missing data caused by that fault.

Reference

[1] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, "Instrumenting the World with Wireless Sensor Networks," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2001)*, Salt Lake City, Utah, May 2001.

[2] Ia.F, Akyildiz, W. Su, Y. Sankarasubramaniam, E.Cayirci , A Survey on Sensor Networks, *IEEEComunications Magazine*, vol.40, no.8, pp.102-114, August 2002

[3] G.J.Pottie and W.J.Kaiser,"Wireless integrated network sensors," *Communications of the ACM*, vol.43, no.5, pp.51-58, 2000

[4] A.Mainwarning, J.Polastre, R.Szewczyk, and D.Culler,"Wireless sensor networks for habitat monitoring," in *First ACM International Workshop on Wireless Sensor Networks and Applications*,(Atlanta, GA), Sept. 2002.

[5] C.Aubrun, C.Leick, "Sensor Fault Accommodation to an Activated Sludge Process", 2005

[6] Deniz Erdogmus, Yadunandana N. Rao, Hemanth Peddaneni, "RECURSIVE PRINCIPAL COMPONENTS ANALYSIS USING EIGENVECTOR MATRIX PERTURBATION"

[7] Dunia, R. "Identification of faulty sensors using principle component analysis", *AIChE* J., 42(10), pp. 2797-2812, 1996.

[8] C. Chih-Chen, K. Sze, and Z. Sun, "Structural damage assessment using principal components analysis." *Proceedings of the SPIE: Health Monitoring and Smart Nondestructive Evaluation of Structural and Biological Systems III.* Volume 5394, pp. 438-445, 2004

[9] S. Costa and S. Fiori, "Image compression using principal component neural networks." *Image and Vision Computing*,Vol. 19, Issues 9-10, 649-668, 2001

[10] J. Li and Y. Zhang, "Interactive sensor network data retrieval and management using principal components analysis transform."*Journal of Smart Materials and Structures*. Vol. 15, No. 6, pp. 1747-1757, 2006

Sung Ho Kim
Professor, Department of Electrical Engineering and Information Technology Kunsan National University

Phone : +82 16 610 1224
E_mail : shkim@kunsan.ac.kr

Yui-Su Youk
Ph.d Student, Department of Electrical Engineering and Information Technology Kunsan National University

E-mail : sixofnum@hotmail.com