# Correlation-based Clustering for Energy Saving in Wireless Sensor Network

Thuật toán phân nhóm tiết kiệm năng lượng dựa vào đặc tính tương quan trong mạng cảm biến không dây

***Nguyen Thi Thanh Nga*, Nguyen Kim Khanh, Ngo Hong Son, Ngo Quynh Thu***

*Hanoi University of Science and Technology, No. 1, Dai Co Viet Str., Hai Ba Trung, Ha Noi, Viet Nam*

*Received: November 25, 2015; accepted: November 26, 2016*

**Abstract**

*The paper concentrates on the energy efficiency in Wireless Sensor Network, based on the correlation characteristics of environment. The sensor nodes are clustered into highly correlated regions (HCRs) to take advantage of correlation between sensor nodes in order to save energy. The determination of HCR is based on the evaluation of the correlation of each two sets of data. Because of highly correlated characteristics among sensed data of nodes in the same HCRs, some high correlated-nodes would be inactive for energy saving but still guarantee acceptable accuracy. Simulation results show that the network lifetime of proposed system is 1.75 times longer than that of the conventional protocol*

Keywords: Entropy; Highly Correlated Region; Energy efficiency; Correlation ratio

**Tóm tắt**

*Bài báo tập trung vào vấn đế tiết kiệm năng lượng trong mạng cảm biến không dây dựa vào đặc tính tương quan của môi trường. Các nút cảm biến được phân nhóm thành các vùng có độ tương quan cao (HCR) để có thể tận dụng được tính chất tương quan giữa dữ liệu đo tại các nút cho mục đích tiết kiệm năng lượng. Việc xác định các vùng HCR được dựa trên đánh giá mức độ tương quan của từng cặp nút. Do tính tương quan cao giữa các nút trong cùng một vùng HCR, một số nút trong vùng có thể được cho tạm nghỉ làm việc để tiết kiệm năng lượng mà vẫn đảm bảo độ chính xác cho phép. Các kết quả mô phỏng chỉ ra thời gian sống của mạng sử dụng thuật toán đề xuất cao hơn 1,75 lần so với các thuật toán thông thường.*

Từ khóa: Entropy, vùng có độ tương quan cao, Tiết kiệm năng lượng, tỉ số tương quan

## 1. Introduction

Wireless sensor networks (WSNs), consisting of a large number of small, inexpensive, battery-powered communication devices densely deployed throughout a physical space can fulfill the monitoring request for surrounding environment characteristics such as temperature, humidity, light, etc. [1-3]. In WSNs, energy conservation is commonly recognized as the key challenge in the design and operation [4-5]. In recent years, a considerable number of research on WSNs have deal with the energy conservation issue. Among these works, clustering is potentially viewed as the most energy-efficient and long-lived technique for sensor networks [6-7].

Numerous clustering algorithms for WSNs have been proposed in the literature [6-7]. However, these routing protocols have not yet considered the characteristics of environmental attributes.

In some special environments, the correlation characteristic of the environment could be used for the network energy efficiency. In [8, 9], a theoretical framework to model the spatial and temporal correlations in sensor networks was developed. In [10], the correlation characteristics for visual information was studied. However, this research considered the correlation based on position dependence and the environment is only single correlation region, i.e. all nodes observed the same phenomenon. Therefore, the problem of correlation grouping/clustering has not been considered yet.

The correlation clustering is considered in [11] but this correlation is described only as similarity between read values and this similarity has not been defined yet. In order to evaluate the correlation, entropy and joint entropy concepts was used [12-14]. Nodes in a same cluster must have high correlation in sensed data. Therefore, in a cluster, sensed data could be compressed smaller before being to be sent, and the cluster heads could save a considerable energy for transmission [12]. On the other hand, the determination of correlation level and correlation based clustering rules has not been referred in [12] and [13] but in [14]. In [14], the calculation of joint entropy is mentioned and there is an enomous number of combinations of picking sensor nodes to calculate the

* Corresponding author: Tel.: (+84) 904.567.424
Email: ngantt@soict.hust.edu.vn

joint entropy that causes a vast amount of computation time.

This paper presents a new cluster-based routing scheme that can provide energy efficiency by proposing a correlation ratio to evaluate the correlation level in order to create cluster (called CCES – Correlation ratio-based Clustering in WSN for Energy Saving Protocol). Instead of calculating the joint entropy of all possible combinations of some nodes in the system, our new algorithm considers the correlation between two nodes alternatively to simplify the computation.

## 2. Entropy and Correlation Ratio

In order to measure the correlation among sets of data, we first consider the concept of entropy and mutual information [15].

If a random variable *X* is defined by a probability distribution *P(X)*, then we will write the entropy of the random variable as:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \quad (1)$$

Joint entropy is the entropy of a joint probability distribution, or a multi-valued random variable. If *X* and *Y* are jointly distributed according to *P(x, y)*, then the joint entropy *H(X,Y)* is:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x,y) \quad (2)$$

The relation between entropy and joint entropy is

$$H(X,Y) \le H(X) + H(Y) \quad (3)$$

with equality if *X* and *Y* are independent.

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. The formal definition of the mutual information of two random variables *X* and *Y*, whose joint distribution is defined by *P(X, Y)* is given by:

$$I(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \quad (4)$$

The relation between mutual information and entropy is given by:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (5)$$

It is difficult to compare the correlation level between two pairs using mutual information or joint entropy, because their values depend on the value of entropy of each individual data in the pair. To overcome this problem, we propose a concept of correlation ratio that is given as follows:

$$\alpha = \frac{H(X,Y)}{H(X) + H(Y)} = 1 - \frac{I(X,Y)}{H(X) + H(Y)} \quad (6)$$

From Eq. (6), it is found that correlation ratio presents the correlation level of a pair of data, that is independent to entropy of each individual data, and therefore, it can be used to compare correlation level of two pairs of data.

From Eq. (3), the joint entropy of a pair of nodes is maximize if *X* and *Y* are independent, which means that joint entropy of *X* and *Y* equals to the sum of entropy of *X* and *Y*. Then, the value of correlation ratio $\alpha$ varies from 0.5 to 1, depending on the correlation between two nodes. The smaller the value of $\alpha$, the higher the correlation is. If $\alpha=0.5$, (in case *H(X)=H(Y)=H(X,Y)*), two sets of data totally depend on each other. If $\alpha=1$ (in case *H(X,Y)=H(X)+H(Y)*), they are indepedent.

## 3. Correlation based clustering in WSN for Energy Saving Protocol (CCES)

From the analyzing in the previous section, it is found that the correlation ratio can measure the correlation among sets of data. In other words, this ratio can be used to determine high correlated region. In this section, we propose CCES, a new clustering based routing protocol for WSN that calculates and uses correlation ratio in order to create different high-correlated cluster, each represents by a Cluster Head (CH). Obviously, data in each cluster is correlated and that is why CH does not need to collect data from every member for saving energy. In other words, within a cluster, nodes that have high correlation ratio needs to send data to the CH alternatively. The operation of CCES is divided into three periods and is described as follows:

### *3.1. Initializing phase*

This period includes a number of rounds that collects sample data for determine the correlation relationship between pairs of nodes. The clustering algorithm in this phase is based on LEACH [16]. In each round, *k* clusters will be created with a CH for each cluster where *k* is the optimum number of cluster Number of cluster *k* is determined based on computation and communication model and is about 5%-10% of total number of nodes in the network. Each node elects itself to be CH by creates a random number and compares to a threshold. The threshold is chosen so that every node will be the cluster head with the same number of time.

In each round, CHs will collect data from their cluster members and send them to base station (BS).

### *3.2. Determination of HCRs*

After receiving data from all nodes in the first period, BS begins to calculate correlation ratios of each pair of nodes for HCRs determination. The HCR forming mechanism is described as follows:

At first, initial nodes for each HCR would be determined. All nodes in the network are divided into *k* groups as same as in the last round of the first period. In each group, the node closest to the group center is chosen to be the initial node of a HCR.

Then, the initial node will choose its neighbors to establish the HCR. It is noted that two nodes are neighbors if the distance between two nodes is less than or equal to communication range $d_{max}$ of the predetermined node. The correlation ratio $\alpha$ of the initial node and its neighbor nodes is calculated by the BS. If $\alpha \leq \alpha_{HCR}$, where $\alpha_{HCR}$ is a threshold that is determined based on sensed data and application, the corresponding neighbor node is set to belong to this HCR. The threshold $\alpha_{HCR}$ is represented for the correlation level of the environment. The smaller the value of $\alpha_{HCR}$ the more correlation of the environment, and the data can be compressed with higher rate. However, this threshold should be chosen very carefully, because, if $\alpha_{HCR}$ is chosen to be too small, the number of HCRs will increase, and the number of nodes in a HCR also decreases. The increment of HCR number reduces the efficiency of the correlation characteristic and even wastes more energy. The optimal value of $\alpha_{HCR}$ will be considered in the next research.

For nodes that do not belong to any HCR established in the previous step, one of them is chosen to be the initial node of a new HCR and the HCR forming process is the same as the previous step.

The forming process will finish when all nodes are arranged in HCRs. In each HCR, the node with largest remaining energy will be chosen to be the CH. The forming process of determination period is described in Fig. 1.

### *3.3. On-off node mechanism*

In a HCR, some pairs of node in which members have dominant correlation, i.e. small enough value of correlation ratio, are calculated. For the purpose of energy efficiency, a node that has higher energy could be active while the other can be turned off. Because of very high correlation among them, sensed data of one node could be the representation for the other . BS will choose a pair of any two nodes in the cluster and calculate their joint entropy. If the result $\alpha \leq \alpha_{dmnt}$ where $\alpha_{dmnt}$ is a ratio threshold that is determined based on required precision of the application and satisfies $\alpha_{dmnt} \leq \alpha_{HCR}$ (*dmnt* means dominant), then in these two nodes, one which has higher energy would be chosen to be an active node, the other will go to sleeping period. When CHs collect data from their members, data from a slept node is treated to be equal to its active node. After each round, the role of pair active node and slept node will be changed to each other.

It is also noted that when $\alpha_{dmnt}$ is used, the collected data is not from all nodes in the network. Such user must accept a distortion. The determination of $\alpha_{dmnt}$ depends on the required distortion level. The relation between distortion level and $\alpha_{dmnt}$ will be considered in the next research.

```
BEGIN
    FOR each group among k groups
        FOR each node in a group
            Set the closest node to the group center
            to be the init_node
            Call HCR_Forming with init_node
        END FOR
    END FOR
    FOR each node in the network
        IF a node is not in any exist HCRs
            Set the node the init_node
            Call HCR_Forming with init_node
        END IF
    END FOR
    FOR each HCR
        Find the node which has the largest remaining energy
        Set the node to be the Cluster_Head
    END FOR
END
BEGIN SUBPROGRAM HCR_Forming with init_node
    Estabish a HCR with init_node
    FOR each node which distance to the init-node ≤ d_max
        IF (the node is not in any exist HCRs) and
            (correlation ratio α_node-init_node ≤ α_HCR)
            Set node into the HCR
        END IF
    END FOR
END SUBPROGRAM
```

**Fig.1** Correlation-based clustering algorithm.

### *3.4. Data transmission*

After dividing the nodes into HCRs, the network will turn to data transmission period where the nodes are clustered according to HCR. Each HCR corresponds to a cluster. Because of high correlation among sensed data at a HCR, sensed data could be compressed in higher rate before being sent to the base station, thus can save energy. In this data transmission period, each cluster has three types of member: CH, active nodes and slept nodes. The active nodes will transmit data to their cluster head. The cluster heads compress received data and send to the base station. The transmitting operation in CCES is also broken into frames as LEACH [16] where nodes send their data to the cluster head at most once per frame during their allocated transmission slot. The slept nodes do not transmit data. They just listen to the waking up command from their cluster head. If there was a significant change of environment attribute at a node, the cluster head will wake all the slept nodes to sense the environment. When there is no change or little change in the environment attributes, after a number of rounds $\Delta r$, BS will calculate joint entropy of the nodes based on new collected data to rearrange the new HCRs.

In [14], in order to calculate the correlation of some node, it is necessary to calculate the sum of a huge number of combinations between the variables and their probability distributions, causes a vast amount of computation time, especially when the number of variables is large. CCES in contrary proposed the evaluation of HCR by calculating the correlation of a pair of sensed data, thus simplifies the calculation as well as reduces the computation time significantly.

## 4. Performance Evaluation

In order to evaluate the advantages of CCES with correlation ratio, simulation is done using sample data given by Intel Berkeley Research Lab [17]. Temperature data are extracted in the collected data including temperature, humidity, light and voltage values for each 31 seconds, and those data are used to be the input data for the sensors in simulation. These data then would be conducted as the collected sensing data of sensors and be used to calculate the correlation ratio between pairs of nodes.

### *4.1. Simulation conditions*

Simulation is done in an area 42m x 30m with 52 nodes which are distributed as in Fig. 2. The radio hardware energy dissipation model of wireless sensor network and compression model are the same as in [12]. The simulation parameters are expressed in Table I. In this simulation, the BS is placed outside the monitoring area at [20, 35]. The initial energy of each sensor node is assumed to be equal and be 0.1[J]. The simulation environment is developed in OMNET++. Based on the analysis of sample data, two correlation thresholds are chosen as follows: $\alpha_{HCR} = 0.62$ and $\alpha_{dmnt} = 0.54$. However, in practice, these thresholds could be chosen based on the need of accuracy of the application.

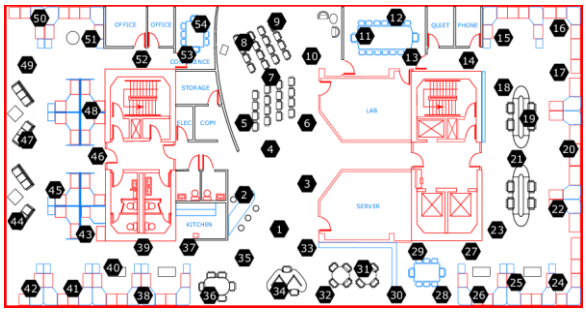

**Fig.2** Sensor arrangement in Intel Berkeley Lab.

In this simulation, with the number of sensor nodes in the network, the network is divided into $k = 4$ groups in the initializing phase, as described in [12].

**Table 1**: Simulation parameters

| Parameter | Value |
|---|---|
| Sensing area [*mxm*] | 42x30 |
| Base station position [*x, y*] | [20, 35] |
| Node number | 52 |
| Initial Energy [*J*] | 0.1 |
| Energy dissipated per bit *Eelec* [*J/bit*] | $50x10^{-9}$ |
| Free space loss $\varepsilon_{fs}$ [*J/bit/m2*] | $10^{-11}$ |
| Aggregated energy *EDA* [*J/bit*] | $5x10^{-9}$ |
| Packet size [*bits*] | 4000 |
| Sensing range[*m*] | 70 |
| Communication range[*m*] | 150 |
| $d_{max}$[*m*] | 75 |
| $\alpha_{HCR}$ | 0.62 |
| $\alpha_{dmnt}$ | 0.54 |
| $\Delta r$[*rounds*] | 60 |

### *4.2 Simulation results*

After first 60 rounds for data collection with LEACH, the BS will calculate to determine HCRs. In the initializing phase, the network is divided into 4 group, but in Fig. 3, after the calculation of correlated data, the network is divided into 8 HCRs. Nodes in the same HCRs have the same colors.

Let's consider the HCR including 12 nodes numbered 3, 4, 5, 6, 7, 8, 45, 46, 47, 48, 49, 51. The correlation of their temperature data is shown in Fig. 4. The variation of their sensed data is quite similar and match with the calculation of correlation ratio value. Thus, evaluation the correlation among data using correlation ratio is appropriate.

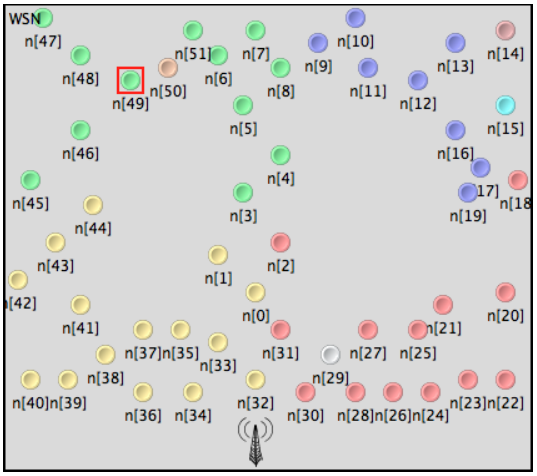


**Fig. 3** Grouping of sensor nodes into HCRs after 60 rounds.

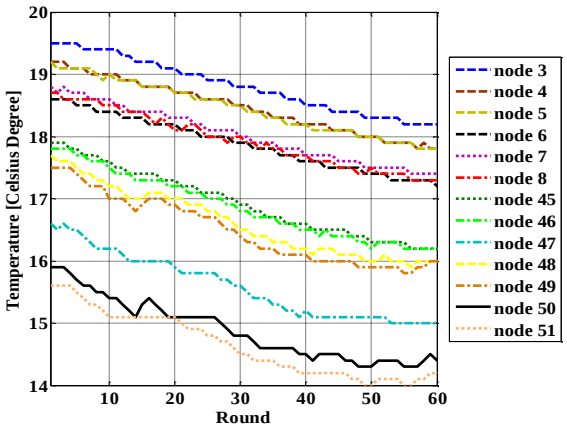


**Fig.4** Variation of sensed data in the considered HCR after 60 rounds.

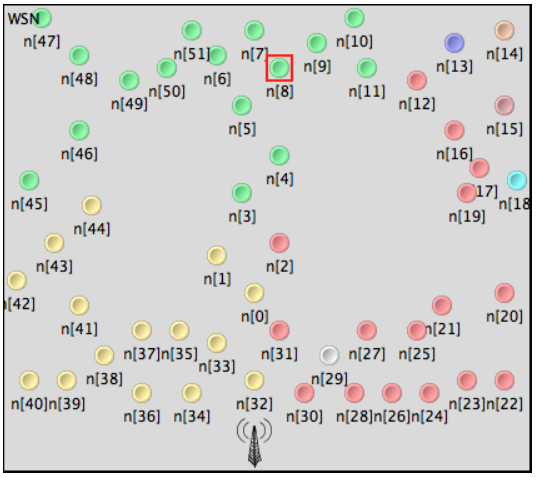


**Fig. 5** Grouping of sensor nodes into HCRs after 120.

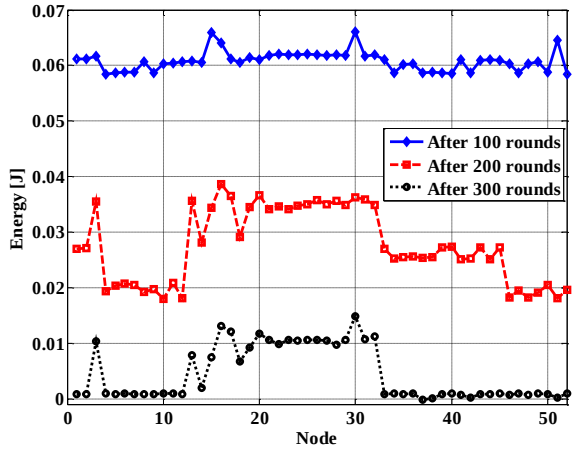


**Fig. 6** Energy distribution of CCES at round 100, 200 and 300.

It is noted that node 50 is located close to nodes in the above HCR, however it does not belong to this HCR. At the round of 120, node 50 joins the HCR. The reason is that the variation of sensed data at node 50 is quite similar to the other nodes in the HCR. At this round, the network is arranged into HCRs as shown in Fig. 5.

Fig. 6 shows the comparison of remaining energy of nodes in CCES at round 100, 200 and 300 with the initial energy for each node is 0.1[J]. After 100 rounds, the remaining energy of nodes reduces about 40% compared with the initial energy. After 200 rounds, remaining energy reduces nearly half compared with remained energy at round 100 and half of the nodes dead after 300 rounds.

The remaining energy of nodes after 100 rounds in CCES shows the balancing between nodes. It is can realized that each node dissipated almost the same energy. After 200 rounds, the balancing in dissipated energy has changed. Some nodes have more than energy than the others. However, from the sensor arrangement, nodes in the same HCR have almost the same remaining energy. These results show the balancing energy could be achieved in each HCR, but not in whole network.

Fig.7 shows the comparison in detail between remaining energy of nodes after 200 rounds of proposed system and conventional system LEACH [16]. It is found that the remaining energies of LEACH are always smaller than those in CCES. In addition, remaining energy of nodes in LEACH is more unbalanced than CCES. The difference in the lowest and the highest energy dissipated in CCES is smaller than in LEACH.

Fig. 8 shows the comparison of network life time between two routing protocols based on number of dead nodes. It is found that the first node lost its all energy after 241 rounds in CCES, but only 186 rounds

in LEACH. Haft of nodes died after 278 rounds in CCES, but after 212 rounds in LEACH. Last node died after 392 rounds in CCES, but after 226 rounds in LEACH alternatively.

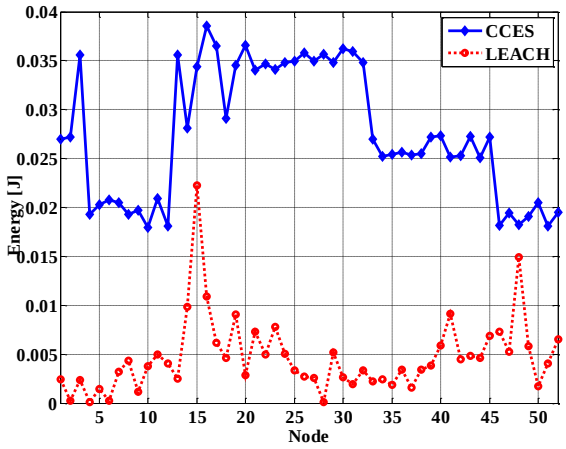


**Fig. 7** Energy distribution at round 200 of CCES and LEACH.

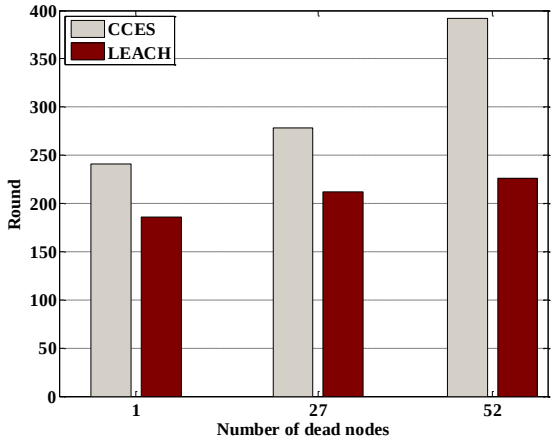


**Fig. 8** Comparison of network life time based on number of dead nodes.

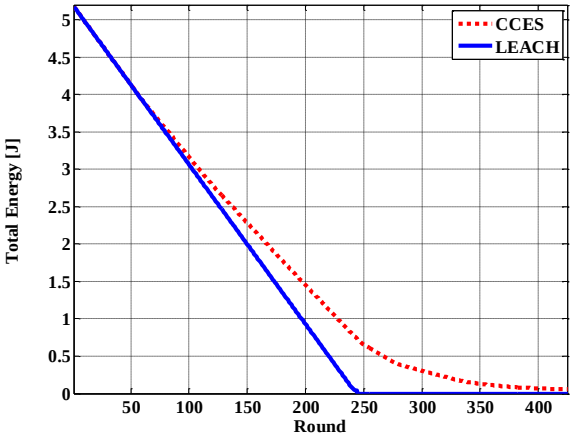


**Fig. 9** Comparison of network life time based on total energy

Fig. 9 shows the comparison of network life time between two routing protocols based on total energy. CCES has longer life time than LEACH about 1.75 times (427 versus 231 rounds). It means that the proposed routing protocol has better efficient dissipated energy than LEACH by turning off some nodes which have dominant correlated data.. As metioned in the simulation conditions, the advantage of grouping correlation nodes in to a HCR to increase the compression rate is not considered in this simulation, but the proposed protocol is still better than LEACH.

## 5. Conclusion and Future works

The paper proposed a new correlation-based clustering scheme CCES that provides energy efficiency by using correlation ratio. Correlation ratio then is used to divided nodes in a wireless sensor network into HCRs and then an energy saving protocol that uses correlation ratio concept is proposed. The simulation results show that our scheme achieves better energy efficiency than the traditional clustering scheme which did not consider the environment characteristics.

In the future, by utilizing the advantages of correlation characteristics, it is expected that the sensed data could be compressed in better to reduce the aggregated messages, thus can save more energy for the transmission from CH to BS. On the other hand, the relation between the correlation ratio threshold $\alpha_{HCR}$ and compression rate should be evaluated.

## References

[1] I. F. Akyldiz, W. Su, Y. Sankarasubramanian and E. Cayirci, "A survey of sensor networks," IEEE Communications Magazine, (2002), pp.102-114.

[2] D. Culler, D. E. M. Srivastava, “Overview of Sensor Network”, IEEE Computer Magazine, vol. 37, no. 8, (2004), pp. 41-49.

[3] C. Siva Ram Murthy and B. Manoj, “Ad Hoc Wireless Networks: Architectures and Protocols”, Prentice Hall, 2004.

[4] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," IEEE Wireless Comm., vol. 11 (2004) pp. 6-28, 2004.

[5] Ignacio Solis and Katia Obraczka, “Isolines: Energy efficient mapping in Sensor Networks”, Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC'05), Cartagena, Spain, (2005).

[6] A. Abbasi and M. Younis, “A Survey on Clustering Algorithms for Wireless Sensor Networks,” Computer Communications, vol. 30, no. 14-15, (2007), pp. 2826–2841.

[7] N. Vlajic and D. Xia, “Wireless Sensor Networks: To Cluster or Not To Cluster?” in Proc. International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), (2006).

[8] Akyildiz, Ian F., Mehmet C. Vuran, and Ozgür B. Akan. "On exploiting spatial and temporal correlation in wireless sensor networks." Proceedings of WiOpt. Vol. 4. 2004.

[9] Shakya, Rajeev K., Yatindra N. Singh, and Nishchal K. Verma. "Generic correlation model for wireless sensor network applications." IET Wireless Sensor Systems 3.4 (2013): 266-276.

[10] Rui Dai, Ian F. Akyildiz, A Spatial Correlation Model for Visual Information in Wireless Multimedia Sensor Networks, IEEE transaction on multimedia, vol.11, No.6, 10. 2009

[11] Myung Ho Yeo, Mi Sook Lee, Seok Jae Lee, Jae Soo Yoo, "Data Correlation-Based Clustering in Sensor Networks", CSA, 2008, Computer Science and its Applications, International Symposium on, Computer Science and its Applications, International Symposium on 2008, pp. 332-337.

[12] D. Maeda, H. Uehara, and M. Yokoyama, Efficient Clustering Scheme Considering Non-uniform Correlation Distribution for Ubiquitous Sensor Networks, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (2007) E90-A (7):1344-1352.

[13] N. T. T. Nga, H. Uehara, T. Ohira, “Attribute change adaptation routing protocol for energy efficiency of wireless sensor networks,” ICITA 2009 (2009).

[14] Taka, H., Uehara, H. and Ohira, T., Intermittent Transmission Method based on Aggregation Model for Clustering Scheme, Third International Conference on Ubiquitous and Future Networks (ICUFN), 2011, pp.107-111, Print ISBN 978-1-4577-1176-3, (2011),pp. 15-17

[15] Thomas M. Cover, Joy A. Thomas, “Elements of Information Theory,” John Wiley & Sons, Inc. Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1, (1991), Chapter2 pp.13-49.

[16] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, An application-specific protocol architecture for wireless microsensor networks, IEEE Trans. on Wireless Communication, vol.1, no.4, (2002), pp.660-670.

[17] Intel Berkeley Research Lab http://db.csail.mit.edu/labdata/labdata.html