Data Compression using Distributed Source Coding in Wireless Sensor Network

[1]Pradeep Kumar, [2]Vrinda Gupta

[1,2]Department of Electronics & Communication, NIT Kurukshetra, Haryana, India

Abstract

In this paper distributed source coding is used to compress data in a wireless sensor network in order to reduce communication energy in sensor nodes. Power is a precious resource in wireless sensor networks due to the limited battery capacity. Distributed source coding refers to the compression of the output of multiple correlated sensor that do not communicate with each other. In this paper, we first study about distributed source coding. Then we study how estimates the number of clusters needed to efficiently utilize data correlation of sensors for a general sensor network. We then present an approximation algorithm that can find an optimal rate allocation within each cluster to minimize the intra cluster communication cost. Then we evaluate the effect of spatial correlation and network size on optimum number of cluster, overall compression ratio and intra-cluster communication cost.

Keywords

Distributed Source Coding, Wireless Sensor Networks, data compression, power consumption.

I. Introduction

In a wireless sensor network (WSN), a number of sensor nodes are densely deployed in a field of interest with one or more data sinks located either at the center or out of the field [1]. The sensor nodes sense the phenomenon at different points in the field and process the data and then finally send the data to the sink(s). Power consumption in the sensor node is for Sensing, Data Processing and Communication. More energy is required for data communication in sensor node. Energy expenditure is less for sensing and data processing. As we know that power is a precious resource in wireless sensor networks due to the limited battery capacity. Once deployed it is often difficult to charge or replace the batteries for these nodes. The capacity of batteries is not expected to improve much in the future. This work explores energy consumption trade-offs associated with lossless data compression. The data compression techniques extend the life time of sensor network. Also by reducing data size less band width is required for sending and receiving data. The observed phenomenon is usually a spatially dependent continuous process, in which the observed data have a certain spatial correlation. In general, the degree of the spatial correlation in the data increases with the decrease of the separation between sensor nodes. Therefore, spatially proximal sensor observations are highly correlated, which leads to considerable data redundancy in the network [2].Slepian Wolf coding [4] is a Distributed source coding that can completely remove data redundancy without requiring inter-sensor communication and therefore a promising technique for data aggregation in WSNs. This technique is based on the assumption that each sensor node has a priori knowledge of the correlation structure, which depends on the distances between sensor nodes and the characteristics of the observed phenomenon. Next problem is how to optimally allocate a rate to each node within a cluster such that the intra-cluster communication cost, which is defined as the total energy consumed by all the sensor nodes in the cluster for sending data encoded with the allocated rates, is minimized. For this problem, we describe the procedures to perform Distributed Source coding locally within each cluster with the optimal rate allocation.

II. Distributed Source coding

Considering a wireless sensor network in which many sensor nodes sense a common event independently, and send their sensed data (readings) to a base station to do information processing. Since their readings usually are highly correlated, that is to say much redundancy exists in their readings. It is straightforward that they can avoid send many redundant information to base station if the sensor nodes can communicate each other. However, in some cases, it is difficult for all sensors to communicate each other; in the cases that sensors can communicate each other, and then communication among sensor nodes consumes energy, which usually is the most significant issue in designing wireless sensor networks. So here comes the idea: can the redundancy be reduced without the direct communication among sensor nodes in the networks? The answer is yes after Slepian and Wolf proposed the famous Slepian-Wolf coding theorem in 1973[5]. One of the enabling technologies for sensor networks is distributed source coding (DSC) [6 - 8], which refers to the compression of multiple correlated sensor outputs that do not communicate with each other. To explain the idea behind Distributed Source Coding (DSC) the concept of entropy is needed. The entropy of a discrete random variable X is denoted H(X) and could be seen as the minimum number of bits needed to encode X without any loss of information. However this is a theoretical bound and to achieve this bound the encoder may have to operate on blocks of infinite length. Similarly the joint entropy H(X,Y) of two discrete random variables X and Y can be seen as the minimum number of bits needed to encode X and Y jointly.

Suppose X and Y are a pair of correlated discrete random variables with the joint probability distribution of p_{xy} (x, y).Two correlated information sequences $X_1,X_2,.....,X_n$, denoted by X, and $Y_1,Y_2,........, Y_n$, denoted by Y, are composed of n independent sample pairs of the correlated variables. X and Y can also be regarded as two corresponding blocks of n-characters produced by two correlated information sources. For the independent decoding of both encoded sequences,[5] showed that, irrespective of whether the two correlated sources are encoded independently or not, the admissible rate region R is the same, bounded by the entropies of the two sources H(X) and H(Y), which are determined by the marginal distributions of X and Y , i.e. p_x(x) and p_y(y), respectively.

When the two correlated source are encoded independently while both encoded sequence are decoded jointly,[5] proved that the admissible rate region R is bounded by two conditional entropies, H(X/Y) and H(Y/X) , and a line giving RX + RY =H(X,Y) ,where H(X,Y) is the joint entropy of two sources X and Y shown in figure.

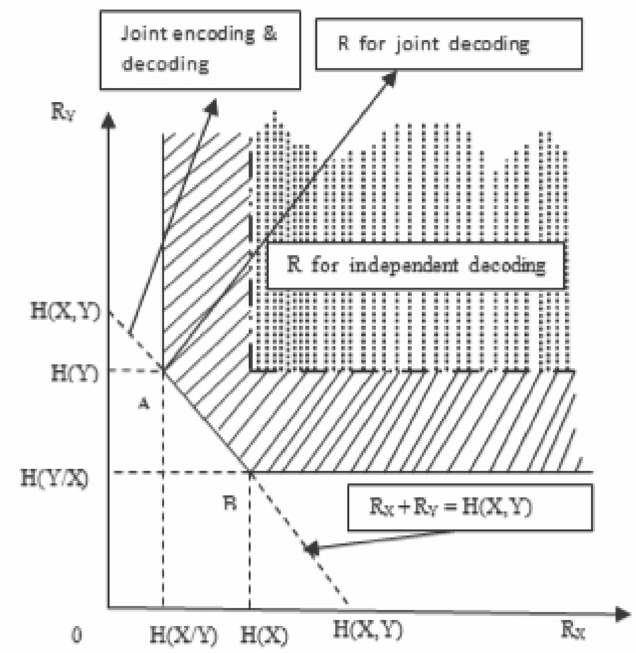


Fig. 1: Admissible rate region R for independent encoding and joint decoding.

Slepian and Wolf showed that two correlated sources can be coded with a total rate H(X, Y) even if they are not able to communicate with each other. Slepian –Wolf theorem says that the achievable region of DSC for discrete source X and Y is given by R_X= H (X|Y) , R_Y= H(Y|X) and R_X + RY ≥ H (X , Y) ,There is an important part of R in fig. 1 , the line segment AB connecting point A and point B. Point A represent R_X = H(X/Y) and R_Y = H(Y), whilst point B represents R_X = H(X) and R_Y = H(Y/X). Slepian and Wolf proved in their work [5] that this line segment is always a part of the boundary of admissible rate region for joint decoding cases, irrespective of whether the two correlated source are encoded independently or not. This shows that, for the joint decoding cases, when assigning coding rate R_X and R_Y using the points on the line AB, even if the two source are encoded separately, the same coding performance can be achieved as if they are encoded jointly. Among the points on AB, point A is often used for the application of DSC, as shown in fig. 1. This above result can be generalized to the N-dimensional case. Consider a network consisting of N sensor nodes uniformly distributed in a region of interest, where each node i produces reading X_i and all the readings constitute a set of jointly ergodic sources denoted by X = (X_1, X_2, .., X_N) with distribution p(X_1 , X_2 ,.............., X_N) which corresponds to the spatial correlation structure known by each node a priori. According to Slepian-Wolf Theorem, the nodes can jointly encode their data without inter-node communication, with a rate (in bits) lower-bounded by their joint entropy H (X_1, X_2............X_N) as long as their respective rates are under the constraints given by

$\sum_{i=0}^{N} \mathrm{Ri}$ = H(X_1 , X_2 , X_3 ...X_N) (1)

Where R_1 = H(X1)

R_i = H (X_i / X_{i-1}, X_{i-2},.., X_1), $2 \leq i \leq N$ (2)

The proof of the Slepian-Wolf theorem is based on typical sequence [9] and random-binning. Following is a simple example of random-binning Suppose X and Y is equiprobable 3-bit binary word correlated in the sense the Hamming distance between X and Y is no more than one. Then H(X)=H(Y) = 3 bits. Because the Hamming distance between X and Y is dH(X,Y)≤1, for a given Y , there are four equiprobable choices of X. For example, when Y = 000, X∈{000,100,010,001}. Hence H(X|Y) = 2 bits. For joint encoding of X and Y, three bits are needed to convey Y and two additional bits to index the four possible choices of X associated with Y, thus a total of H(X,Y) = H(Y) + H(X|Y) =5 bits suffice. This can be done by first partitioning the set of all possible outcomes of X into four sets Z_{00},Z_{01},Z_{10},Z_{11} with Z_{00}={000,111}, Z_{01}={001,110}, Z_{10}={010,101} and Z_{11}={011,100} and then sending two bits for the index s of the Zs that X belongs to. In forming the sets Zs's, we make sure that each of them has two elements with Hamming distance dH=3. For joint decoding with s (hence Z_s) and side information Y, we pick in set Z_s the X with d_H(X, Y) ≤1. Unique decoding is guaranteed because the two elements in each sets Z_s have Hamming distance d_H=3. Thus we indeed achieve the Slepian-Wolf limit of H(X,Y) = H(Y) +H(X/Y) = 3 + 2 = 5 bits in this example with lossless decoding.

III. Clustering

Applying Distributed source coding globally in the whole network is difficult and would incur significant additional costs because each sensor node needs the knowledge of global correlation structure to encode its own data and more energy to transmit data to sink. Moreover, Distributed source coding is not tolerant to relay and node failures because the data from one node may affect the decoding of the data from other nodes [10]. For these reasons, it is unsuitable to apply Distributed source coding globally in a large network. In a cluster-based network, however, each cluster covers a smaller number of sensor nodes within a small local range of the network. This makes it feasible to apply Distributed source coding locally within each cluster because in this case a sensor node only needs the knowledge of local correlation structure to perform coding. Meanwhile, it will not obviously compromise the compression performance because in the real world the spatial correlation usually decreases with distance [2,11].

Consider the sensor network with N randomly located sensors [12]. Suppose the average distance from a sensor S_i to the other sensors which it can directly communicate with, is di. Entropy of the sensor S_i is H_i. The information that is only provided by the sensor S_i can be expressed by entropy

H(S) – H(S ∩ $\hat{S}_i$) (3)

where H(S) is entropy of total sensor network; H(S ∩ $\hat{S}_i$) is entropy of total sensor network excluding sensor S_i. The coefficient c_i is defined as the percentage of unique information of sensor S_i compared to H(S_i), the complete information provided by S_i:

$$ci = \frac{H(S)-H(S\cap\hat{S}i)}{H(Si)}, \quad 0 \leq c_i \leq 1 \qquad (4)$$

The coefficient c_i can express the degree of correlation between a sensor and its neighborhood sensors. In real sensor networks, the data from two sensors will be almost decorrelative when both sensors are located far away from each other. Each sensor S_i can estimate c_i according to the correlation between current sensor and its nearby sensors.

We consider K=N/s clusters each consisting of s sensors. The cluster head for each cluster is located at the center of cluster. The total number of bits-hop cost for the whole sensor network is expressed as

$$E_{whole} = \sum_{i=1}^{K}(E_{intra}(i)+ E_{extra}(i))$$
$$=K(E_{intra} + E_{extra}) \qquad (5)$$

Where $E_{intra}(i)$ and $E_{extra}(i)$ are the bit-hop cost within cluster i and the bit-hop cost for cluster i to the sink respectively. E_{intra} and E_{extra} are the average bit-hop cost within the clusters and the average bit-hop cost from the clusters to the sinks. We can obtain expressions for each of these:

$$E_{intra} \; \alpha \; ((s-1)Hc)(d\sqrt{s}) \qquad (6)$$

$$E_{extra} \; \alpha \; (H + (s-1)Hc)(d\sqrt{N}) \qquad (7)$$

where H, c and d is the average of $H(S_i)$, c_i and d_i . (s −1)Hc and $d\sqrt{s}$ are average number of bits of all sensors except the head of cluster in a cluster, and average distance from the sensors to the head of cluster respectively. H + (s −1)Hc and $d\sqrt{N}$ are average number of bits of the head of cluster after data fusion and average distance from the head of cluster to the root. The number of sensors in each cluster is much larger than one (s-1≈s when s>>1) in most case. So, we have

$$E_{whole} = K\,E_{intra} + K\,E_{extra} \qquad (8)$$

$$\alpha \; K(sHcd\sqrt{s})+K(H\,d\sqrt{N} + (s-1)Hc\,d\sqrt{N})$$

$$\propto \; \sqrt{s}\,HcdN + s\text{-}1HdN\sqrt{N}\,(1\text{-}c) + HcdN\sqrt{N}$$

The optimum value of the cluster size s_{opt} can be determined by setting the derivative of the above expression equal to zero:

$$\frac{dEwhole}{ds} = 0$$

$$s_{opt} = \left(\frac{c}{2(1-c)\sqrt{N}}\right)\text{^}(-2/3) \qquad (9)$$

The optimum number of clusters can be expressed as:

$$K_{opt} = \frac{N}{Sopt} = \left(\frac{cN}{2(1-c)}\right)\text{^}(2/3) \qquad (10)$$

The optimum number of clusters K_{opt} depends on the number of sensors in the entire sensor network and the degree of correlation c. As expected low correlation levels require small cluster sizes while high correlation levels require larger cluster sizes. Also the number of clusters is relative to sensor network size. In other words, the larger the sensor network (larger N) the more clusters are required to apply optimal data aggregation.

IV. Optimal Rate allocation for Distributed source coding

Consider a cluster A with |A| sensor nodes shown in Fig. 2. where the node in black represent the cluster head and nodes in white represent cluster members. The cluster head produces reading X_1. From left to right, the first cluster member is closest to the cluster head and produces reading X_2,and so on. Each node i produces reading X_i and uses the shortest distance d(i, v) to reach cluster head. The rate allocation problem is to find an optimal rate vector $\{R_i\}$, i = {1, 2,..................., |A|), for all |A| nodes so that the total flow cost = $\sum_{i=1}^{|A|} d(i,v)\,Ri$, is minimized. The rate allocation for each cluster must satisfy the condition that the total rate of the coded sensor readings in a cluster is equal to their joint entropy. Where d (i , v) is defined as the distance from node i to the cluster head v. Thus, for the optimal rate allocation for distributed source coding within this cluster arrange cluster members in descending order of their distance to the cluster head, as shown in fig. 2. Thus optimal intra-cluster rate given below minimizes total flow cost.

$$R_1 = H(X_1) \qquad (11)$$

$$R_i = H(X_i \mid \{X_j \mid d(j,1) \leq d(i,1),\, j \in A\}) \;,\; 2 \leq i \leq |A| \qquad (12)$$

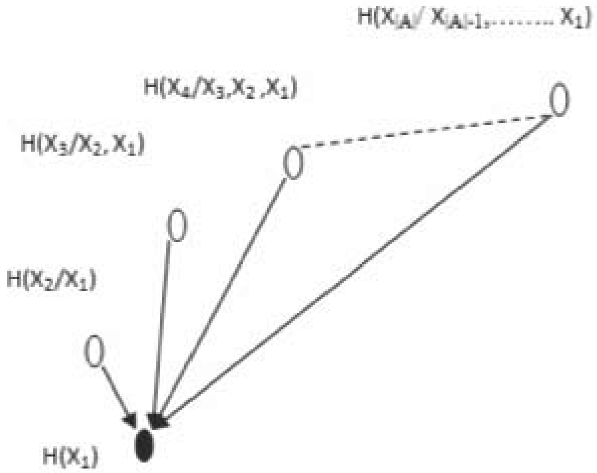


Fig. 2 : Distributed source coding with optimal rate allocation

According to chain theory, we can easily prove validity of the solution given by (1),

$$\sum_{i=1}^{|A|} Ri = H(X1) + \sum_{i=2}^{|A|} H(Xi \mid Xi-1, Xi-2, \ldots\ldots, X1) \qquad (13)$$

$$\sum_{i=1}^{|A|} Ri = H(X_1, X_2, \ldots\ldots\ldots, X_N) \qquad (14)$$

Here the cluster head with zero distance to itself is encoded with a rate equal to its unconditional entropy and each of the cluster members in the cluster is encoded with a rate equal to its respective entropy conditioned on all the other nodes in the cluster which are closer to the cluster head than itself.

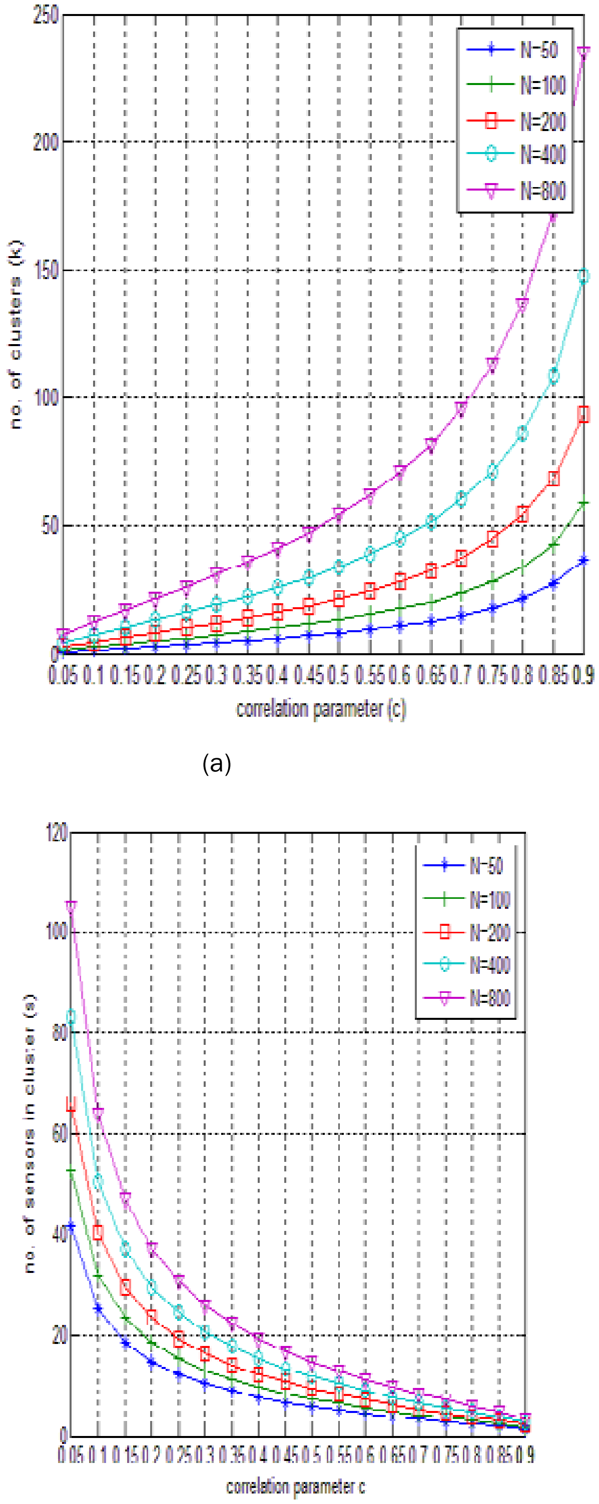
V. Simulation Results

In this section, we evaluate the effects of the extent of spatial correlation and the network size on the optimal number of the cluster needed to efficiently utilize data correlation, number of sensor in a cluster and compression performance of Slepian-Wolf coding through simulations based MATLAB 7.04 simulation tool. Also, we investigate the performance of optimal intra-cluster rate allocation with respect to the intra-cluster communication cost of a cluster of different network size.

The optimal number of clusters can be expressed as :

$$K_{opt} = \frac{N}{Sopt} = \left(\frac{cN}{2(1-c)}\right)\text{^}(2/3) \qquad (15)$$

The optimal number of cluster depends on the number of sensors in the entire sensor network and the degree of correlation. Fig. 3 shows how different number of clusters and cluster size perform across a range of correlation levels and sensor network sizes.



(a)

(b)

Fig. 3 : Analytical curves for number of cluster (K) and sizes of clusters (s) with respect to degree of correlation (c) and size of the network (N).

As expected low correlation levels require small cluster sizes while high correlation levels require larger cluster sizes. Also the number of clusters is relative to sensor network size. In other words, the larger the sensor network (N) the more clusters are required to apply optimal data aggregation.

Now for evaluating compression performance of Distributed Source coding, we consider a cluster with N sensor nodes uniformly deployed in a 100m×100m sensing region and the data sink located at the center of the region. The simulation results are based on the average of 2 experiments and each experiment uses a different randomly-generated topology. For the correlation structure, we assume that the observations $X_1,X_2,......,X_N$ at N sensor nodes are modeled as an N-dimensional random vector $X = [X_1,X_2 ....., X_N\ ]T$, which has a multivariate normal distribution with mean zero and covariance matrix K and the differential entropy of $H(X_1,X_2 ....., X_N)$ is

$$H(X_1,X_2 ....., X_N)=\frac{1}{2}\log_2(2\pi e)^N \det(K) \qquad (16)$$

where det(K) denotes determinant of the matrix K [4]. In the simulation, we use an exponential model of the covariance $K_{ij} = \exp(-dij\ c)^2$ to model the observed physical event, where dij denotes the distance between the nodes measuring X_i and X_j respectively. The parameter c controls the relation between the nodes and it can be set different values to indicate different levels of correlation within a given distance. For the sake of simplicity and without loss of generality, we use differential entropy instead of discrete entropy because we assume that the sensor reading from different nodes are quantized with an identical and sufficient small quantization step, in which case the differential entropy differs from discrete entropy by only a constant [4].

Fig. 4 shows the effects of the extent of correlation and the network size on the compression performance. The compression performance is measured in an overall compression ratio, which is the total amount of data produced in the network after clustered Distributed Source Coding is applied over the total number of bits generated by all nodes without using this distributed source coding scheme.

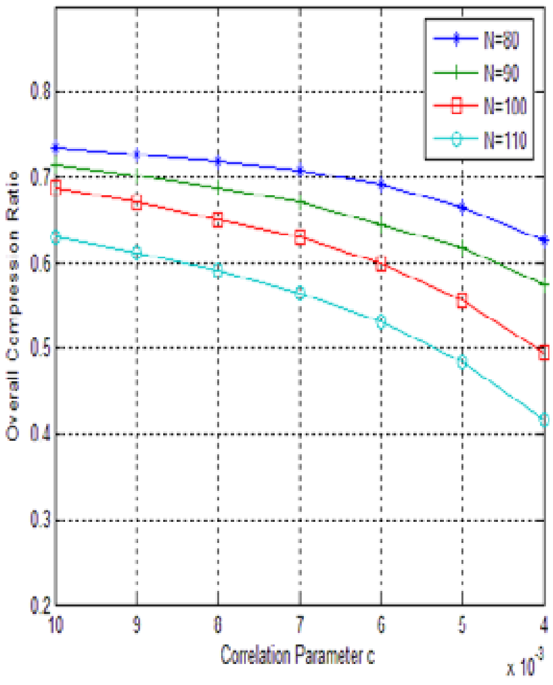


Fig. 4: Impact of the extent of correlation and the network size on overall compression ratio

In the case of high correlation, a better compression performance is achieved because the Distributed Source Coding can remove more redundancy caused by the high spatial correlation among the readings of different sensor nodes. In addition, the compression performance is improved as the network size or the density of sensor nodes increases. This behavior is due to the fact that the denser sensor deployment results in more sensor nodes residing within cluster while Slepian-Wolf coding can completely get rid of the highly redundant data generated by these sensor nodes in closer proximity to each other.

Fig.5 shows the intra-cluster communication cost with the optimal rate allocation and rate without using this distributed source coding scheme, respectively. The optimal rate allocation scheme first arranges nodes in cluster A in ascending order of their distance to the cluster head, thus the rate assigned to the node i can be expressed by

$$R_i = H(Xi \mid \{X_j \mid d(j,1) \le d(i,1), j Є A\}) \qquad (17)$$

Intra-cluster communication cost with the optimal rate allocation

$$= \sum_{i=1}^{|A|} Ri\ d(i,1) \qquad (18)$$

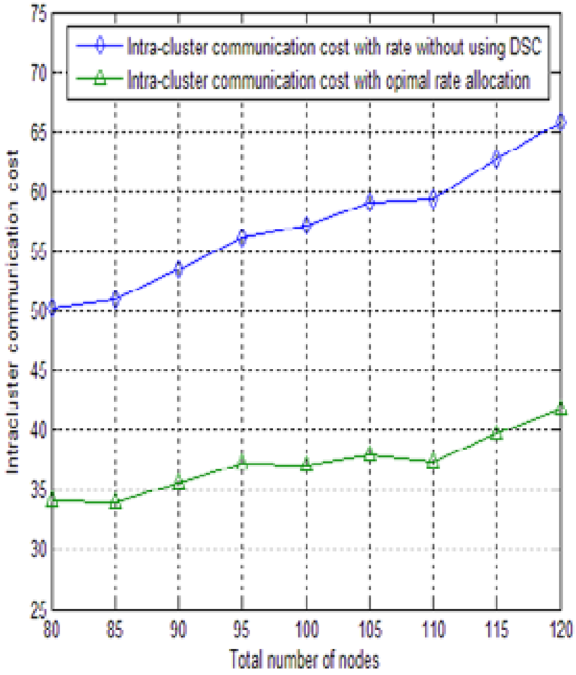


Fig. 5: Intra-cluster communication cost with optimal rate allocation and rate without using distributed source coding

The result shown is an intra-cluster communication cost of a cluster with different network sizes. As expected, the optimal intra-cluster rate allocation results in less communication cost compared with rate without using this distributed source coding scheme because former scheme jointly considers rate assignments and transmission distances between the cluster members and the cluster head.

VI. Conclusion & Future Scope

The most important factor of designing WSN is how to improve the energy efficiency. The distributed source coding scheme will significantly decrease the whole network energy consumption by decrease the complexity of sensor encoder and compress the amount of transmitted data. The simulation results demonstrate that the clustered Distributed Source Coding can significantly reduce the total amount of data in the whole network while the transmission cost within cluster can be remarkably reduced by performing the optimal intra-cluster rate allocation. There are several interesting and challenging questions that yet remain to be answered here for example, correlation model and measurement noise. Correlation statistics is mainly a function of the location of the sensor nodes. In some applications, sensor network has a varying network topology. This requires a time-varying correlation model and adaptive or universal DSC that could achieve gains for time-varying correlation. It is still an open and very challenging DSC problem. The existence of measurement noise, if not taken into account, causes some mismatch between the actual correlation between the sources and the noiseless correlation statistics used to do the decoding, which means worse performance.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci,''Wireless sensor networks: A survey,'' Computer Networks (Elsevier)Journal, vol. 4, no. 12, Mar. 2002, pp. 393–422.

[2] M. C. Vuran, I. F. Akyildiz, "Spatial correlation-based collaborative medium access control in wireless sensor networks," Networking, IEEE/ACM Transactions on, vol. 14, no. 2, 2006, pp. 316-329.

[3] Linguo Yin, Changmain Wang, Geir E.Qien, "On the Minimization of Communication Energy Consumption of Correlated sensor Nodes." Springer Science + Business Media,LLC.2008

[4] T.M.Cover, J.A.Thomas, Element of information Theory,New York, NY, USA:John Wiley and Sons,Inc.,1991..

[5] Slepian, J.K.Wolf, "Noiseless coding of correlated information source,"IEEE Trans. Of Information Theory, pp. 471-480, Jul.1973.

[6] S.S. Pradhan, K. Ramchandran. "Distributed Source coding: Symmetric rates and application to sensor networks." IEEE Data Compression Conference(DDC),March 2000.

[7] S.S Pradhan, J. Kusuma, K. Ramchandran. "Distributed Compression in a Dense Microsensor Network." IEEE Signal Processing Magazine, March 2002.

[8] Zixiang Xiong, et. Al., "Distributed Source Coding For Sensor Networks",IEEE Signal Processing Magazine, Sep. 2004.

[9] T. M. Cover, J. A. Thomas, Elements of Information Theory. New York: John Wiley & Sons, 1991..

[10] D. Marco, D.L. Neuhoff, "Reliability vs. efficiency in Distributed source coding for field-gathering sensor networks," in Proc. Of the Third International Symposium on Information Processing in Sensor Networks(IPSN'04), Berkeley, CA Apr. 26-27, 2004, pp. 161-168.

[11] R. Cristescu, B. Beferull-Lozano, M. Vetterli, ''Networked slepianwolf: theory, algorithms, and scaling laws,'' Information Theory, IEEE Transactions on, vol. 51, no. 12, 2005, pp. 4057– 4073.

[12] Hao chen, Seapahn Magerian, "Cluster Sizing and Head Selection for Efficient Data Aggregation and Routing in Sensor Networks." IEEE Wireless Communication and Networking Conference (WCNC), April 2006.

[13] M. C. Vuran, I. F. Akyildiz, "Spatial correlation-based collaborative medium access control in wireless sensor networks," Networking IEEE/ACM Transactions on , vol. 14, no. 2, 2006, pp. 316-329.

[14] R. Cristescu, B. Befferull-Lozano, M . Vetterli, " On network correlated data gathering," in Proc. Of INFOCOM'04, Vol.4, Hong kong,March.2004,pp. 2571-2582.