DATA COMPRESSION ALGORITHM FOR WSN BASED ON DISTRIBUTED CODING FOR ONE-DIMENSIONAL

AHMED JAMAL AHMED

A report submitted in partial fulfilment of the requirement for the award of the degree of Master of Electrical Engineering

Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia

Jan 201

ABSTRACT

One of the major challenges to design efficient Wireless Sensors Networks (WSN) is the scarcity of energy and computational resources. Distributed source coding schemes provide closed loop algorithms that exploit the source redundancy in the WSN to reduce the amount of information that each node transmits. In many wireless sensor network applications, the data obtained at sequential time points by the same node are time correlated, while, spatial correlation may exist in data obtained at the same time by adjacent nodes. A great deal of node energy will be wasted if data which include time and space correlation is transmitted. A distributed way of continuously exploiting existing correlations in sensor data based on adaptive signal processing and distributed source coding principles is proposed. Our approach enables sensor nodes to blindly compress their readings with respect to one another without the need for explicit and energy-expensive inter-sensor communication to effect this compression. Furthermore, the distributed algorithm used by each sensor node is extremely low in complexity and easy to implement (i.e., one modulo operation). A study is herein carried out along with an implementation of the aforementioned algorithms with the Matlab software. The stability of the closed loop algorithms is then tested by means of simulations. Sinks can obtain correlation parameters based on optimal order estimation by exploring time and space redundancy included in data which is obtained by sensors. Then the sink restores all data based on time and space correlation parameters and only a little necessary data needs to be transmitted by nodes. Because of the decrease of redundancy, the average energy cost per node is reduced and the life of the wireless sensor network is obviously extended as a result.

ABSTRAK

Satu daripadfa cabaran utama untuk merekabentuk Pengesan Rangkaian Tanpa Wayar (WSN) adalah kekurangan tenaga dan sumber perkomputeran. Skim sumber kod teragih menyediakan algoritma pusingan tertutup yang mengeksplotasi pertindihan sumber dalam WSN untuk megurangkan jumlah maklumat yang dihantar oleh setiap nod. Dalam kebanyakan aplikasi pengesan rangkaian tanpa wayar, data diperolehi pada titik masa berjujukan oleh nod-nod yang sama adalah berkolerasi dengan masa, sementara kolerasi spatial boleh wujud dalam pemerolehan data pada masa yang sama oleh nod –nod yang bersebelahan. Satu persetujuan yang penting pada tenaga di nod akan dibazirkan jika data yang terdiri daripada masa dan kolerai ruang dihantar. Satu kaedah pengagihan yang berterusan mengeksplotasi kolerasi yang telah wujud dalam pengesan data yang berasaskan pemperosesan isyarat yang sesuai dan prinsip sumber kod yang teragih di cadangkan dalam projek ini. Pendekatan ini membolehkan nod pengesan untuk memampatkan secara rambang bacaannya dengan berpandukan antara satu sama lain tanpa perlu untuk eksplisit dan tenaga kos tinggi antara pengesan komunikasi untuk member kesan kepada pemampatan ini. Selain itu, algoritma pengagihan ini telah digunakan oleh setiap pengesan nod adlah terlalu rendah dari segi kompleksiti dan mudah untuk diimplementasikan (contoh: operasi modul tunggal). Satu kajian telah dilaksanakan seiring dengan satu pemasangan seperti yang dinyatakan menggunakan perisian MATLAB. Kestabilan pusingan tertutup telah diuji secara simulasi. Pusat penyimpanan dapat memperolehi kolerasi parameter berdasarkan anggaran arahan yang optimum dengan eksplorasi masa dan tindanan ruang termasuk dalam data yang diperolehi pengesan. Pusat penyimpanan kemudiannya mengembalikan semua data berdasarkan kepada kolerasi parameter masa dan ruang dan hanya memerlukan sedikit data untuk dihantar ke nod-nod. Disebabkan pengurangan pertindihan, purata kos tenaga per nod dikurangkan dan jangka hayat pengesan rangkaian tanpa wayar dijelaskan pada keputusan ujikaji.

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADC	Analog to Digital Converter
AWGN	AdditiveWhite Gaussian Noise
BER	Bit Error Rate
dB	decibel
DS-CDMA	Direct Sequence - Coded Division Multiple Access
IM	Initialization Module
IS	International System
Kbps	Kilobits per second
LMS	Least-Mean-Squares
LSB	Least Significative Bit
MAI	Multiple Access Interference
ME	Minimum Energy
MSE	Mean Square Error
PDA	Personal Digital Assistant
PN	Pseudorandom Noise
RF	Radio Frequency
SINR	Signal to Interference + Noise Ratio
WSN	Wireless Sensor Networks
WSMN	Wireless Sensor Multimedia Networks

CONTENTS

CHAPTER 1

INTRODUCTION

The rapid growth of communication technology has resulted in many emerging technologies. One of these technologies is Wireless Sensor Networks (WSN), which are composed by a many of small nodes (sensors). These nodes are distributed on wide area to cooperate on sensing a physical phenomenon such as such as temperature which is part of an environmental monitoring process. The application space is huge and is comprised of, for example, industrial, smart life, monitoring and management scenarios. In WSNs, many constraints are imposed that obstruct this technology from performing its significant impact in larger scale systems.

Clearly, these constraints are related to energy consumption, bandwidth, scalability, reliability and cost-efficiency. The transmission and reception process consumes most of the nodes energy. The main task of WSNs is to provide channels between the virtual world of information technology and the real practical world. They create a change from traditional forms of inter- personal communications to automatic inter-device communications.Therefore, the main objective in this research is to minimize the amount of inter-node communication. Hence, energy-efficient compression technique for compressing the sensed data from individual nodes will be developed.

There is a wide variety of WSN applications ranging from agricultural soil management, seismic or geophysics activity monitoring, military coordination and business activities such as supply chain management processes.

In WSNs, many constraints are imposed that obstruct this technology from performing its significant impact in larger scale systems. Clearly, these constraints are related to energy consumption, bandwidth, scalability, reliability and cost-efficiency. Sensor nodes are usually battery-powered and should function without much supervision for a relatively long period of time. In most situations, it is very difficult and even impossible to replace or recharge batteries. The data transmission and reception process between WSN nodes and sink node consumes most of the nodes energy.

In summary, WSNs use wireless medium for inter-device communications among the sink and nodes. Moreover, the wireless nature of the channel and constant inter-node communications create an undesired phenomena such fast power consumption, noise disturbances, transmission errors and data reception delays. Considering these consequences, energy efficiency has a principal importance for the operational life time of a wireless sensor network. Therefore, the main objective in this research is to minimize the amount of inter-node communication energy. Hence, energy-efficient compression technique for compressing the sensed data from individual nodes will be developed.

1.1 Research Background

The huge development of information technology in line with the development of sensing devices and wireless technology further motivated the innovation of wireless sensor networks [1]-[3]. As we know wireless sensor network is a network with large number of sensor nodes, which are deployed in a wide area by some rules or random ways. And within the wireless network, sensor nodes communicate mutually to build a network to collect important data and sent these data to the sink node.

The sensor nodes in the WSNs can be connected by radio frequency, act as without any wire connection. However, if two nodes can't communicate directly, other nodes which are located near those two nodes will transmit a data packet from the source node to the destination node. Then the data relays from sensor nodes to the sink, which then may be connect to the outside world through the internet gateway.

Energy is a primary constraint in the design of sensor networks. However, this fundamental energy constraint further limits everything from data sensing rates and link bandwidth, to node size and weight. Therefore, the transmitted packet must be small. And particularly, the large-scale sensor networks such as the WSNs in environmental monitoring, home automation, health facilities, military systems the data is so huge and not as small as temperature or humidity data [4]. The data for this large scale WSNs are rich and contains images, audio and videos data [5-7]. Reducing the size of the transmitted packet is one of the important tasks in WSNs.

Data compression techniques are emerging for such sensor networks. By compressing the data, the size is reduced and less bandwidth is required for transmitting data. There are two common solutions to conserve the energy of sensor nodes. One solution is to take advantage of node redundancy by selecting a subset of sensor nodes to be active while putting others to sleep to conserve their energy. The selected subsets of active sensor nodes have to cover the whole monitoring region and maintain the network connectivity. In other words, these active sensor nodes have to make sure that the network still functions as well as the case when all sensor nodes are active. By selecting different subsets of sensor nodes to be active by turns, it can be prevented some sensor nodes from consuming too much energy and thus extend the network lifetime. However, when node redundancy is not available (because of network deployment or sensor breakage, for example), such sleep-active mechanisms may not be applied.

Another solution is to reduce the amount or number of sensed data to be sent by sensor nodes, because transmission is one of the most energy-consuming operations of sensor nodes. This solution is especially useful when sensor nodes have to regularly report their sensing data to the sink(s) for a very long time. In order to reduce the amount of sensing data, it is needed to compress data inside the network. The data compression schemes can be classified into three categories: lossless, loss, and unrecoverable. A lossless compression is meant that after executing the decompression operation, we can obtain exactly the same data as those before executing the compression operation. Huffman coding is one of the representative examples. A loss compression means that some detailed (and usually minor) features of data may be lost due to the compression operation. Hence, in this research a lossless compression technique will be developed to avoid all the aforementioned energy based issues.

Due to the limitation of energy and transmission capacity, the transmitted packet must be small. However, in the large-scale sensor networks, for example, the WSNs in environmental monitoring, home automation, health, military, and other applications, the data is not the small data, such as temperature, humidity [5], and they are rich data, such as image, audio, especially, the video. Other-wise, authors in [6] indicate that the energy consumption of 1 bit information transfer between 100 meters distance is equivalent to the implementation of 3000 instructions. And the 80% of power is approximately used for data transmission [5, 7]. So, to reduce the size of the trans-mitted packet is one of the important tasks in WSNs.

Commonly, there are two solutions to conserve the energy of sensor nodes. The solution is to take advantage of node redundancy by selecting a subset of sensor nodes to be active while putting others to sleep to conserve their energy [7]–[9]. The selected subset of active sensor nodes have to cover the whole monitoring region and maintain the network connectivity. In other words, these active sensor nodes have to make sure that the network still functions as well as the case when all sensor nodes are active. By selecting different subsets of sensor nodes to be active by turns, we can prevent some sensor nodes from consuming too much energy and thus extend the network lifetime. However, when node redundancy is not available (because of network deployment [10], [11] or sensor breakage, for example), such sleep-active mechanisms may not be applied. Another solution is to reduce the amount of sensing data to be sent by sensor nodes, because transmission is one of the most energy-consuming operations of sensor nodes. Such a solution is especially useful when sensor nodes have to regularly report their sensing data to the sink(s) for a very long time. In order to reduce the amount of sensing data, we need to *compress* them inside the network.

Depending on the recoverability of data, we can classify the data compression schemes into three categories: *lossless*, *loss*, and *unrecoverable*. A lossless compression means that after executing the decompression operation, we can obtain exactly the same data as those before executing the compression operation. Huffman coding [12] is one of the representative examples. A loss compression means that some detailed (and usually minor) features of data may be lost due to the compression operation. Most of the image and video compression schemes such as JPEG2000 [13] belong to this category.

1.2 Problem Statement

The advances in sensor and communication technology have focused interest on using wireless sensor networks, which are formed by a set of small untethered sensor devices that are deployed in an ad hoc fashion to cooperate on sensing a physical phenomenon, making the inferences, and transmitting the data.

Recently, Wireless sensor network is one of the current major research areas in wireless systems. Although WSN doesn't desire lines to connect them, making them a viable solution that can be placed in various locations with minimal disturbance to the surroundings become a challenging problem.

WSN has limited impact on full scale systems due to the issue of power constraints. Energy is an important parameter constraint in the design of sensor networks. Due to the limitation of energy and transmission capacity, the transmitted packet must be small. Other-wise, authors in [6] indicate that the energy consumption of one bit information transfer between 100 meters distance is equivalent to the implementation of 3000 instructions.

One way to overcome this is to increase the amount of energy of the nodes. However, this is not practical option as the nodes are battery energized. The other solution is to reduce the consumed energy by the nodes. So, to reduce the size of the transmitted packet is one of the crucial activities in WSNs. In other words, the data need to be compressed using data compression technique, which can control energy consumption by making the big data into smaller one. Although there are many effective data compression algorithms, but these methods, which are called traditional data compression algorithms, are mostly based on the Nyquist sampling theorem. This theorem points out that sampling rate must be twice the highest frequency present in the signal of interest. And by using this theorem, some redundant data are collected firstly, but the redundant data will be compressed by some rules, such as the statistics theory, and so on. Obviously, the procedure of data collection and compression consume the energy so much that the life of the sensor nodes is shortened. The research questions which need to be answered are: whether the useful data can be acquired directly, not the useless data under the condition of power limitation? That is to say, whether changing the traditional data collection methods are effective and achievable? The compressed sensing gives us a new way to resolve these problems. This work explores energy consumption trade-offs associated with lossless data compression. This data compression technique extends the life time of sensor network.

1.3 Research Objectives

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.The energy conservation is the main factor that needs to be achieved. Therefore, an efficient data compression algorithm can reduce drastically the wasted energy then minimizing the power node cost and obtaining more reliability and scalability.

In this project, mainly an algorithm for data compression of wireless sensor network which is capable to minimize the energy consumption of the nodes and resulting in overcoming the other obstacles will be developed. Precisely, the objectives of this project are as follow:-

i. To investigate data compression algorithms of wireless sensor networks
ii. To develop and enhance data compression algorithm based on the techniques of coding and information theory.
iii. To validate the developed algorithm using Matlab software.

1.4 Aims of the Research

Generally, this research is considered as a milestone for developing a reliable data compression algorithm that can encode the sensed data efficiently. This algorithm would help the wireless sensor network designers to enhance their designs and avoiding the other obstacles. Consequently, better transmission with higher reliability and minimum bandwidth can be obtained using the developed algorithm. The other evaluation parameters such as signal to noise ratio will be studied in details with its related effect to data compression.

1.5 Research Scope

The research area of this project is mainly on the compression of sensed data in wireless sensor network. However, more constraints are applied to limit the scope as follow:-

i. This research study is specified for the lossless data compression techniques for wireless sensor network.
ii. The developed algorithm is designed to be easy, enhanced and efficient.
iii. The algorithm will be simulated using MATLAB software.

1.6 Thesis Outline

This thesis is divided into five main chapters. Chapter I discusses the introduction, problem statement, objectives and scope of the project. In chapter II, literature review of wireless sensor network, data compression algorithms, power consumption reduction techniques and previous related literature of WSNs are discussed. For chapter III, the methodology of this project is discussed. The results of this project are highlighted and elaborated in Chapter IV. Chapter VI discusses the conclusion and recommendation for the future work.

CHAPTER 2

THEORY AND LITERATURE REVIEW

2.1 Introduction

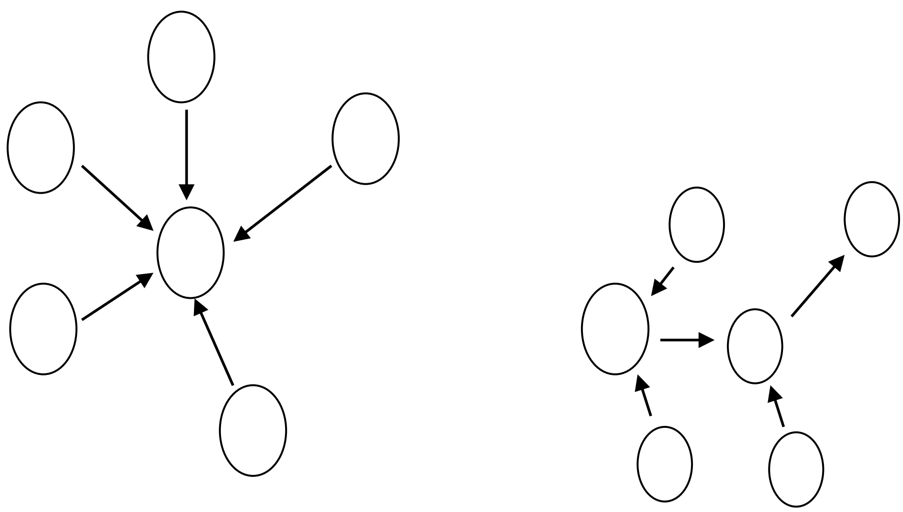
Wireless Sensor Network (WSN) has widely deployed in multidisciplinary applications of our life such as environmental monitoring and military surveillance. It composes of many small nodes that collect the data then send it to the sink node. It does not need a wire for connecting among the nodes. However, the wireless communication of WSN among these nodes has a lack of energy saving. Due to that small nodes are battery based power supply, the power is consumed quickly as they transmit data to the far away sink node.

Many researches had been conducted to overcome the issue of power consumption and its consequences. Obviously, the energy issue is considered the main subject that prohibits the WSN to be implemented in many large scale technology systems. The energy need to be minimized as possible to increase the performance, reliability and efficiency of WSN. One of the solutions is to reduce the amount of energy of the transmitted data then saving energy. In other words, the transmitted data can be compressed using compression technique. In this project, a data compression algorithm will be developed that can be used to encode the data in an efficient manner.

In WSNs, the distributed sensors in an outlined area observe the physical changes in the region. Although every sensor observes comparative physical changes, the signals received from every sensor have much correlation. To reduce the data size, these correlated signals can be compressed. The traditional way of data compression (e.g., Joint Entropy) for WSN involves communication between nodes and utilizes correlated data for data compression procedures. However, this conventional transmission procedure makes the system framework complex. Rather than the traditional schemes, Compressive Sensing (CS) is another and new decentralized data compression technology which attains a low network system complexity [12].

2.2 Architecture of Wireless Sensor Networks

The WSN systems can be designed in different ways. It is normally divided the choices in two groups; the fusion based and the ad hoc based design. The major difference between the two types of WSN architecture is that; in the ad hoc scheme each node needs to be equipped with both transmitting and receiving capabilities, while with a fusion centre the nodes only have to transmit their data. Fig. 2.1 shows the two types of WSN architectures.



(a) Fusion (b) Ad Hoc

Fig 2.1: WSN with different design architectures

Regardless of the different many versions of WSNs that have been designed and built during the last few years, a typical WSN schematic consists of five major components: sensors, memory, radio transceiver, microprocessor unit, and energy supply. An example wireless sensor node schematic is illustrated in Fig 2.2. in the following paragraphs, a detailed information of each component will be introduced.

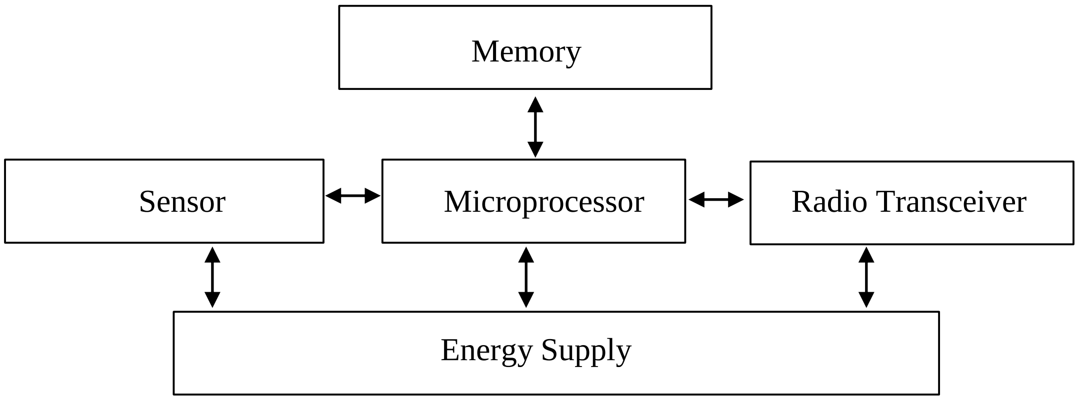


Fig 2.2: Wireless sensor node schematic diagram.

Sensors: These are devices which has the ability to detect and respond to input signals from the physical surrounding environment. Wireless sensors nodes may have more than one sensors on board depending on their applications. Some examples of sensors in wireless sensor nodes are: Light sensors, temperature sensors, humidity sensors, magnetometers and resolution cameras. The main function of sensors are generally related to analog – to – digital convertions (ADC). The analog input signals are converted to digital signals, in a way that microprocessors can undertdand and process the data.

Microprocessor: Computational data processing of wireless sensor nodes are done by an installed mircoprocessors. It processes both close proximity data sensed and remote data for communication to other sensor nodes. Because of the size and energy consumption limitations of WSN, low-power implanted microprocessors are utilized. Because of the limited computational speed of microprocessors, the operating system, software and data processing capability of WSN are very limited.

Memory : WSNs has a small device on board which stores data. This storage device is in the form of Random access memory (RAM) and read –only memory (ROM). It incorporates both information memory, to store crude and processed sensor estimations and other neighborhood data and system memory, from which implanted microprocessor reads and executes guidelines to complete the assigned sensing assignments.

Radio transceiver: To create a wireless communications among wireless sensor nodes, a short –range radio transcievers are used. Due to the data processing capability and power usage limitations, the data rate of current time transceivers on sensor nodes is limited to about 200 Kbps and the communication range is restricted to tens of meters. As radio transceiver is the major power comsumer component in the node, it must employ power efficient modes such as sleep and waking up approaches.

Power supply: A standout and very crucial component in WSN is power supply. For wireless sensor nodes to be adaptable and deployable, batteries are used as a source of energy supply. Most of the energy supplies in wireless sensor nodes are consumed by radio transmission and reception processes as well as complex computational tasks of the microprocessor.

2.3 Demand for Source Coding

Since the power consumption problem of wireless sensor nodes is the major constraints of the system, many efforts have been made to suitably address this issue and to solve the problem. One of the involved aspects to address this matter is source coding. As a consequence of the high deployment density, nodes sense highly correlated data containing both spatial and temporal repetition. Hence, given the high correlation exhibit in the data, creating a suitable protocol and source coding became very important. The source coding techniques utilizes this specific characteristic to decrease the overall transmitted information (i.e. to lower the power consumption). However, algorithms must be simple and be able to occupy the limited memory and lower capability of the sensor’s processing unit.

2.4 Data Compression Algorithms in WSN

Wireless sensor networks (WSNs) had opened a new research field for continuous monitoring of physical environments. Many WSN applications are deployed to do at long-term environmental monitoring. In these applications, energy consumption is the major concern because sensor nodes have to persistently report their sensing data to the remote sink(s) for a very long time. Therefore, many research efforts focus on reducing the energy amount of data transmissions using data compression techniques.

Commonly, the data compression techniques can be divided into two different types; lossless and lossy compression. Lossless compression technique, as the name implies involve no loss of information. In other words, the original data can be recovered exactly from the compressed data. This can be obtained by employing the statistical redundancy to represent the sender’s data more with fewer errors. In contrast, lossy compression techniques involve some loss of information and data that have been compressed using lossy technique generally cannot be recovered or reconstructed exactly. From other view point, the data compression techniques in WSNs can be classified into five categories: (1) the string-based compression techniques treat sensing data as a chain of characters and then adopt the text data compression schemes to compress them. (2) The image-based compression techniques hierarchically organize WSNs and then employ the idea from the image compression solutions to handle sensing data. (3) The distributed source coding techniques extend the Slepian-Wolf theorem to encode multiple correlated data streams independently at sensor nodes and then jointly decode them at the sink. (4) The compressed sensing techniques adopt a small number of non-adaptive and randomized linear projection samples to compress sensing data.(5) The data aggregation techniques select a subset of sensor nodes in the network to be responsible for fusing the sensing data from other sensor nodes to reduce the amount of data transmissions.

The data compression process can be performed based on the spatial or transform domain. The spatial domain algorithms are simpler than the others in transform domain [4-17] and they use the correlation of the nodes to construct a new data set. However, algorithms in transform domain use some transformation to get the new type of the collected data. Then discover the redundancy of the data. That is to say, these algorithms find the correlation of data, not the nodes, in the transform domain, because the correlation of data is not obvious in spatial to temporal domain sometimes.

2.4.1 Algorithm in Spatial Domain

Although the algorithms in spatial domain are simple and effective, the compressed rate of these algorithms isn't very high, because they don't consider the relationship between the collected data. In the next section, more details are provided about the spatial algorithms based on the correlation of inter-nodes [7], the order on the nodes [8], the pipelined process [9], and the quantization method [10-16].

A. Algorithm based on nodes correlation of inter-nodes

The correlation of the nodes can be used to construct a new data set to compress the collected data. For the nodes in WSNs are distributed in a wide area, so the distributed source coding is very effective. Then the data collected by the nodes can be use distributed source coding to compress. The main idea of these algorithms [7] is described in Figure 2.3. It shows that the sensor node ***B*** and node ***A*** both have an analog to digital device, but the sensor node ***B*** has an encoder, too. That is to say, the analog signal in sensor ***A*** is only transformed into digital signal ***Y***. And the analog signal in sensor ***B*** must be transformed into digital one then is encoded into the new signal ***P***, which is as the check information. When the signal ***Y*** and ***P*** send to the data center, the signal ***Y*** and ***P*** is as the input of the encoder. The reconstruction information ***X*** is built by the data sent by sensor ***A***, which is recognized as the raw data, and the check information ***P***. The compression procedure is realized by sensor node ***B***. However, from this procedure, we can't find the compressed data of sensor node ***A***. And the data of sensor node ***A*** is as the auxiliary part of sensor node ***B***. The goal of the WSNs is to realize the energy save for the entire sensor nodes. When the sensor ***A*** is power off, the information of sensor node ***B*** can't be reconstructed, which make these algorithms lack many applications when the number of sensor nodes is big.

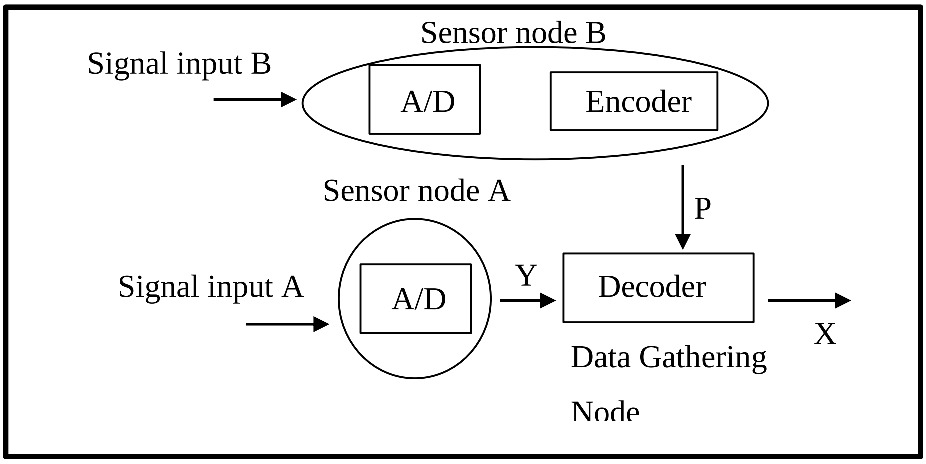


Fig 2.3: Distributed coding algorithm

B. Algorithm based on order mapping

The WSN nodes can have simple order. If the order is known for all the nodes, simply the collected data can be compressed based on the nodes orders. The basic procedure of this algorithm is as follows: data collected by the four nodes is aggregated by the aggregate node. Then the order of the other three nodes represents the value of the node4 as shown in Table 2.1. This algorithm is simple but it requires known ordering of the nodes and this is not easy for huge number of nodes. And the mapping value of sensor nodes may have too much that we should construct a lookup table, which is used to encode and recover the original information. But the energy consumption on finding the lookup table and compute the mapping value may be more than the energy on transmitting the data directly.

Table 2.1 : Ordering Map algorithm

Ordering	**Mapping value**
P1,P2,P3	0
P1,P3,P2	1
P2,P1,P3	2

P2,P3,P3	3
P3,P1,P2	4
P3,P2,P1	5

C. Algorithm based on pipelined

The algorithm based on pipelined is similar to the algorithm based on the order mapping. The key point of this algorithm is to aggregate the different nodes into a data set, and then transmits the new data set [9]. The redundancy is considered and removed. Fig 2.4 shows how the data of three nodes are aggregated and encoded. The format of the data packet is as follows: <measured value, node ID, timestamp>. And we can get the aggregated packet like: <shared prefix, suffix list, node ID list, timestamp list>. The number of the sensor nodes also influences the usability of this algorithm. The aggregated data set may be large enough that transmission procedure needs too much time. And the data centre must know the time clearly. Otherwise, the received data may be confused easily.

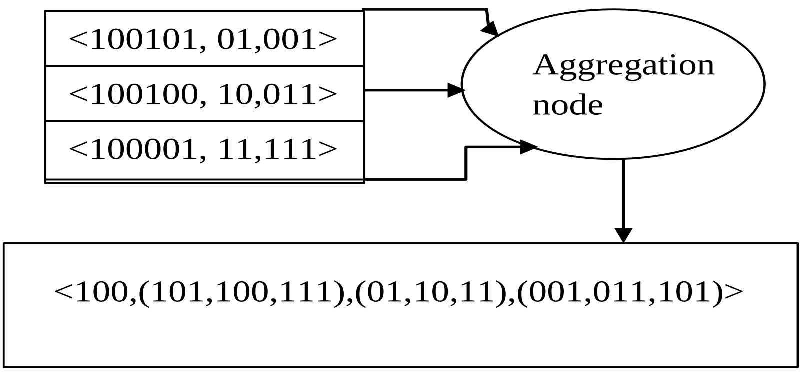


Fig 2.4: Pipelined compression algorithm

D. Algorithm based on quantization

The data collected by single node can be quantitated of the differences between consecutive samples with Differential Pulse Code Modulation (DPCM) [10]. And different trade-offs between compression performance and information loss is determined. And this method can be used in the applications with variable compression rate. The codebook is also used in [12], and the authors in [12] use the Learning Vector Quantization (LVQ) to construct the codebook in the Dictionary Lookup Scheme.

Data compressions algorithms in spatial domain are simple, and can be used in the small number of nodes in WSNs. The correlation of the collected data is not considered, because it is not clear in the spatial domain. The distributed wavelet compression and distributed source coding are introduced in the data compression algorithms.

2.4.2 Algorithms in Transform Domain

In WSNs, the congestion causes an increase in the amount of data loss and delays in data transmission. So an adaptive compression-based congestion control technique (ACT) in [13] for packet reduction based was developed. The detail of the procedure is showed in Fig 2.5. The collected data is transformed into Discrete Wavelet Transform (DWT) firstly. And the range of the data is reduced with the help of Adaptive Differential Pulse Code Modulation (ADPCM). Then the number of packets is reduced by employing Run-Length Coding (RLC) before transfer of data in source node. Then, the transformed data is classified into four groups and assigned priorities for DWT different frequency. The data is defined different quantization steps for different group. In the relaying node, the ACT reduces the number of packets by increasing the quantization step size of ADPCM in case of congestion. The destination node (usually a sink node) reverses the compression procedure. A sink node should apply inverse RLC, inverse ADPCM, and then inverse DWT. The inherent correlation existing between sensor readings was explored.

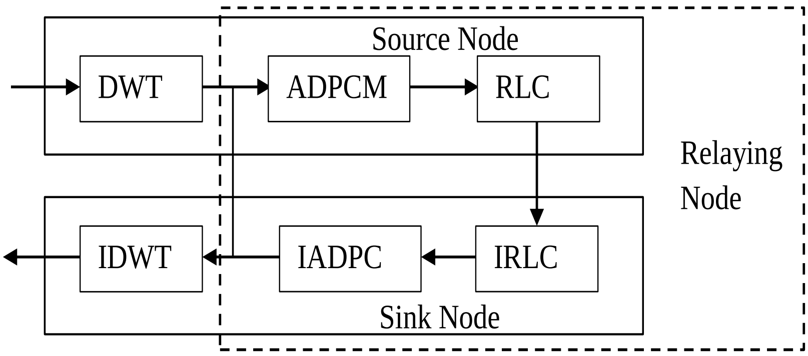


Fig 2.5: Adaptive compression scheme

2.5 Distributed Source Coding

The distributed source coding techniques compress sensing data inside the network according to the Slepian-Wolf theorem, which proves that two or more correlated data streams can be encoded independently and then be decoded jointly at a receiver with a rate equal to their joint entropy. Therefore, the distributed source coding techniques can support lossless compression. One foundation of the distributed source coding techniques is the Slepian-Wolf theorem [18]. Given two or more correlated data streams, each being encoded independently, and then decoded jointly at one receiver, the Slepian-Wolf theorem shows that it is feasible to achieve lossless encoding of these two data streams at a rate (that is, the of bits used to encode each character) equal to their joint entropy. Fig 2.6 (a) gives an example, where there are two correlated data streams X and Y generated by making n independent drawings from a joint probability distribution $P(X = x, Y = y)$. Encoder ***1*** receives data stream X and then transmits a coded message to the decoder, where each character of X is encoded by a number of R_X bits. Similarly, encoder ***2*** receives data stream Y and then transmits a coded message to the decoder, where each character of Y is encoded by a number of R_Y bits. On receiving these two coded messages, the decoder will generate two n-vectors X* and Y *, which are the estimations of the original data streams X and Y, respectively.

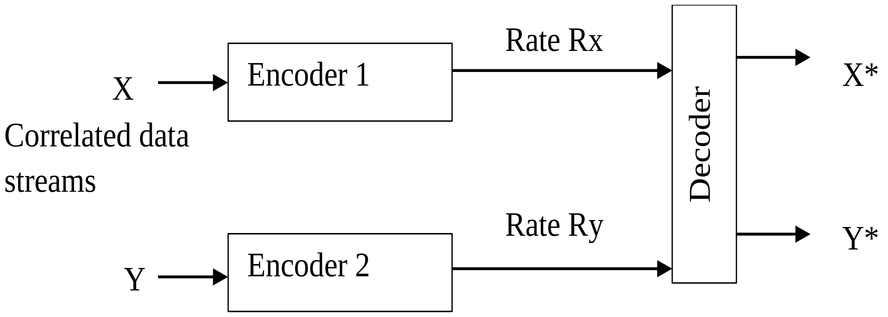


Fig 2.6 (a): Distributed source coding using Slepian-Wolf theorem

When n is sufficiently large, the probability that $X^* \neq X$ or $Y^* \neq Y$ can approximate to zero. That is, we can achieve lossless data compression of X and Y. In this case, the system is called an admissible system. The pair of rates (R_X, R_Y) for an admissible system is called an admissible rate pair. The closure of the set of all possible admissible rate pairs is called the admissible rate region as shown in Fig 2.6 (b). The admissible rate region can be calculated by measuring the entropies of random variables X and Y with joint probability distribution P(X = x, Y = y):

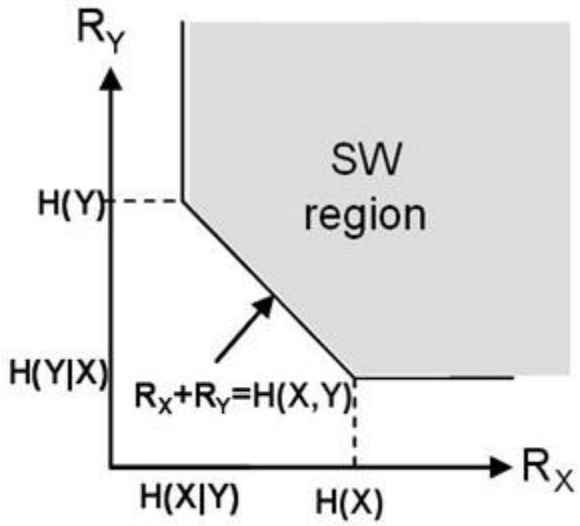


Fig 2.6 (b): Distributed source coding showing the admissible rate region

By the Slepian-wolf theorm, the admissible rate region for the pair of rates (R_X, R_Y) is the set of points that satisfy the following three inequalities:

$$R_X \geq H(X|Y), \quad (1a)$$

$$R_Y \geq H(Y|X), \quad (1b)$$

$$R_X + R_Y \geq H(X,Y), \quad (1c)$$

Where

$$H(X,Y) = -\sum_x \sum_y P(X=x,Y=y) \cdot \log P(X=x,Y=y), \quad (2a)$$

$$H(X) = -\sum_x P(X=x) \cdot \log P(X=x), \quad (2b)$$

$$H(Y) = -\sum_x P(Y=y) \cdot \log P(Y=y), \quad (2c)$$

$$H(Y|X) = -\sum_x P(X=x) \sum_y P(Y=y|X=x) \cdot \log P(Y=y|X=x), \quad (2d)$$

$$H(X|Y) = -\sum_y P(Y=y) \sum_x P(X=x|Y=y) \cdot \log P(X=x|Y=y), \quad (2e)$$

The advantage of the Slepian-Wolf theorem can be observed by comparing it with the entropy bound for compression of single sources. In particular, separate encoders that ignore the source correlation can achieve rates of only $RX + RY \geq H(X) + H(Y)$. However, by adopting the Slepian-Wolf coding, the separate encoders can exploit their knowledge of the correlation to achieve the same rates as an optimal joint encoder, that is, $RX + RY \geq H(X, Y)$.

The Slepian-Wolf theorem provides a theoretical tool to characterize the amount of communications required for the distributed source coding in a network where correlated data streams are physically separated or each encoder has limited computation capability. The studies of [19], [20] give an example of applying the Slepian-Wolf theorem to compress sensing data in WSNs. Specifically, suppose that X and Y are the sensing readings of two sensor nodes, where X and Y are equiprobable binary triplets with X, Y $\{0, 1\}3$ and the Hamming distance between X and Y is no more than one. In this case, we have $H(X) = H(Y) = 3$ bits. Since X and Y differ at most in one position, for any given Y, there are four equiprobable choices of X. For example, suppose that Y is 111, and then X belongs to the set {111, 011, 101, and 110}. Thus, we can obtain that $H(X|Y) = 2$ bits. In other words, to jointly encode X and Y , it takes three bits to represent Y and two additional bits to index these four possible choices of X associated with Y . Therefore, at least $H(X, Y) = H(Y) + H(X|Y) = 5$ bits are required. In fact, the information Y is perfectly known at the decoder (for example, the sink) but not at the encoder (that is, the sensor that generates X). However, according to the Slepian-Wolf theorem, it is still possible to send only $H(X|Y) = 2$ bits rather than $H(X) = 3$ bits to decode X without any loss at the joint decoder.

One solution is to first divide the set of all possible outcomes of X into four subsets $X00 = \{000, 111\}$, $X01 = \{001, 110\}$, $X10 = \{010, 101\}$, and $X11 = \{011, 100\}$ and then send two bits for the index i of the subset Xi that X belongs to. When generating the subsets Xi's, we should guarantee that each of these subsets has two elements with a Hamming distance of 3. Then, to jointly decode with i (and thus Xi) and information Y , we choose the X with $dH(X, Y) \leq 1$ in subset Xi, where $dH(X, Y)$ is the Hamming distance between X and Y . In this case, we can make sure of unique decoding because the two elements in each subset Xi have a Hamming distance of 3.

Therefore, we can achieve the Slepian-Wolf limit of $H(X, Y) = H(Y) + H(X|Y) = 3+2 = 5$ bits in the above example with lossless decoding.

A distributed spatial-temporal data compression algorithm based on a ring topology wavelet transformation is proposed by Zhou and Lin [21], who supports a broad scope of wavelet transformations and is able to efficiently decrease spatial-temporal redundancy, but the algorithm may be too complex for nodes because of the wavelet transformations. An online information compression algorithm is proposed [22], by van which first divides data obtained by nodes into different lengths of shorter data, and then a dictionary can be composed according to the shorter data and updated with the increased data obtained by nodes. The algorithm can reduce the average energy cost per node and improve the accuracy of the restored data, however, the course of updating the dictionary makes the algorithm more complex than the self-based regression algorithm discussed above. The distributed structure tree depression algorithm is proposed by in [24]. It explores the spatial-temporal relativity that exists in data and computes the relativity parameters in the sink. Then the sink restores data according to relativity parameters and part of the original data transmitted from the nodes, so the nodes' energy cost can be reduced to some extent.

In this project, both 1-dimentional and 2-D are xeplored. However, only 1-D data compression algorithm based on the concept of distributed coding and hamming distance will be introduced. This algorithm will employ the feature of correlation among the data of different nodes. Although this algorithm is more difficult than others, it can provide good compression comparing with its original data.

2.6 Lossless Distributed Source Coding

Source coding technique can be divided into two: Lossless source coding and lossy source coding. Lossless source coding is a method where the signal is compressed and the decompressed signal gives back an exact copy of the original signal. The Slepian Wolf coding method is lossless source coding. Modems use lossless mode of data compression. Moreover, a lossless algorithm may be used as a building block in designing a lossy compressor. Some practical examples of lossless source coding are shown below:

1. **Fixed-length to fixed-length (FF) coding**: Suppose you have a program that uses files that contain only decimal digits, spaces, tabs, carriage

returns, dots, and commas. On most computers these files would be stored in eight-bit ASCII. It is easy to write a special-purpose compressor that will compress these files by a factor of two just by encoding each ASCII character into a four-bit codeword.

2. **Fixed-length to variable-length (FV) coding**: Even if all 256 possible characters occur in the file, it is still possible to compress it if some characters are more frequent than others. Suppose the files described in the previous example are not guaranteed to contain only the fifteen characters mentioned, but the remaining 241 characters occur very infrequently. The fifteen frequent characters could be encoded using codewords 0001 through 1111, while 0000 would be reserved as an escape code to indicate that the following eight bits should be interpreted as a single character.
3. **Variable-length to fixed-length (VF) coding**: Suppose you have some files representing black and white images, with black pixels represented by 1-bits and white pixels by 0-bits. Often such files will have long runs of identical bits. A simple run-length compressor might encode runs of 1 through 128 0-bits using codewords 00000000 through 01111111 and encode runs of 1 through 128 1-bits using codewords 10000000 through 11111111. If there are many more long runs than short runs this would give good compression.
4. **Variable-length to variable-length (VV) coding**: The compressor in the previous example might be improved by using short codewords for frequently-occurring run lengths at the expense of lengthening the codewords for less frequent run lengths.

Some applications of lossless image compression:

- Images meant for further analysis and processing (as opposed to just human perception)
  - Medical, space
- Images where loss might have legal implications
  - Medical

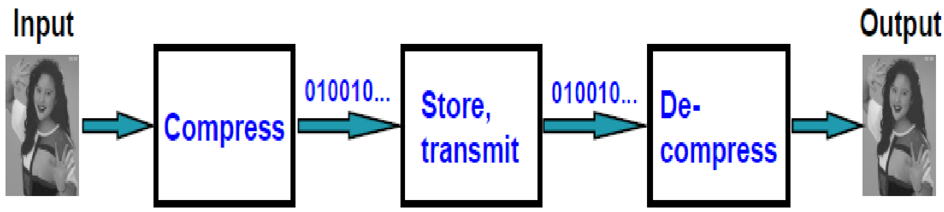


Fig 2.7: Lossless image compression

2.7 Lossy Distributed Source Coding

Lossy distributed source coding accomplishes data compression by losing some parts of the original signal that don’t have much importance. Wyner-Ziv source coding method is an example of lossy source coding. The source statistics play an important role in successful source coding.

The advantage of lossy methods over lossless methods is that in some cases a lossy method can produce a much smaller compressed file than any lossless method, while still meeting the requirements of the application. Lossy methods are most often used for compressing sound, images or videos. This is because these types of data are intended for human interpretation where the mind can easily "fill in the blanks" or see past very minor errors or inconsistencies – ideally lossy compression is transparent (imperceptible), which can be verified via an ABX test. Fig 2.8 shows comparison of lossless and lossy data compression methods.

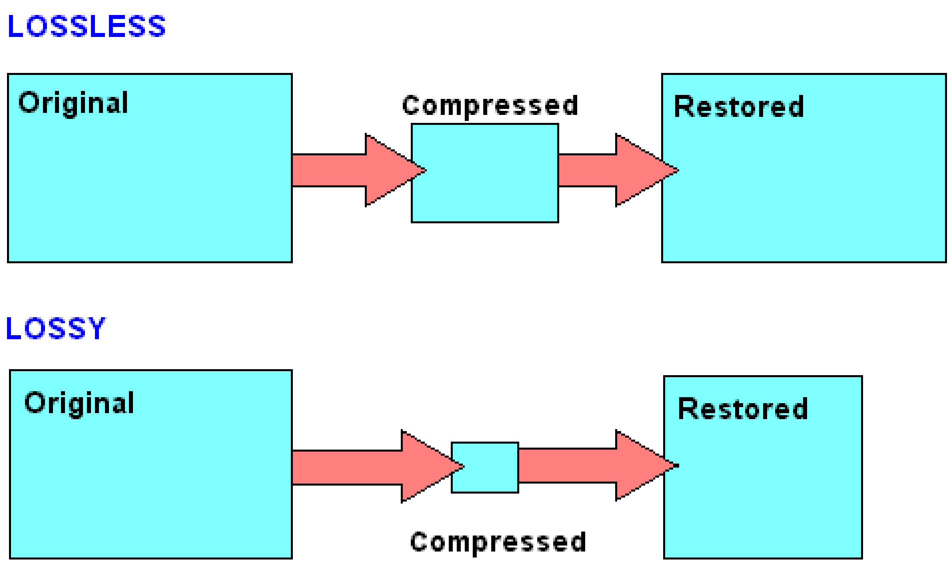


Fig 2.8: Lossless vs Lossy data compression

2.8 Distributed Source Coding using Syndromes

Both distributed compression and non-distributed settings have continuous sampling which needs to be quantized so as to find finite entropy. The data correlation between the sources in a WSN may be designed as a correlation channel. This concept can be used for a channel coding method to further compress the data from the source. For instance, the quantized codeword U is correlated to X. If X is correlated to the side information Y, in this case U is also correlated to Y. Hence, the correlation channel can be defined by the conditional distribution P(Y |U). The side information carries the information I (U; Y) about U which can be utilized on the decoder side to approximate X. Now the correlation distribution may differ considerably from case to case, but it is often modelled in literature as a Binary Symmetric Channel (BSC) or a channel with Additive White Gaussian Noise (AWGN).

In the following example is illustrated a way that the amount of a bits (e.g. X) be reduced without knowing the equivalent sample of Y. For instance, assume that X and Y are two equiprobable 3-bit words where the correlation is given by a Hamming distance not more than one. So, If Y was known both at the decoder and at the encoder, there would be no point in representing X with more than 2 bits. (Given Y, $X \oplus Y$ is in the set {000, 001, 010, 100} while $\oplus$ is the modulo-two sum). With Y known only at the decoder, is this still possible? Yes, there is no point sending both X = 000 and X = 111 because the Hamming distance between them is 3. With Y available, one of the codewords in the set is uniquely chosen. Now all the possible codeword representations of X can be sorted into similar sets giving the additional sets of X: {001, 110}, {010, 101} and {100, 011}. Hence X needs only be transmitted with 2 bits instead of 3.

REFERENCES

[1] He NH, Li HS, Gao J. "Energy-saving Routing Algorithm Based on Cluster inWSN". *TELKOMNIKA*.2013; 11(2): 839-847.

[2] Nasri M, Helali A, Sghaier H, Maaref H. "Efficient JPEG 2000 Image Compression Scheme for Multihop Wireless Networks". TELKOMNIKA. 2011; 9(2): 311-318.

[3] Chen X, Li F, Wang L. "Fault-Tolerant Method for Detecting Coverage Holes of Wireless Sensor Networks".*TELKOMNIKA*. 2012; 10(4): 876-882.

[4] Marcelloni F, Vecchio M. "A Simple Algorithm for Data Compression in Wireless Sensor Networks". *IEEE Communications Letters*. 2008; 12(6): 411-413.

[5] Pottie J, Kaiser WJ. "Wireless Integrated Network Sensors". *Communications of the AC*. 2000; 43: 51-58.

[6] Kimura N, Latifi S." *A Survey on Data Compression in Wireless Sensor Networks"*. In *Proceedings of International Conference on Information Technology*: Coding and Computing. LasVegas. 2005; 2; 8-13.

[7] Yin YG, Ding GG. "Study on Data Compression in Wireless Sensor Networks". *Computer Applications and Software (In Chinese)*. 2010; 27(7): 1-4.

[8] Petrovie D, Shah RC, Ramchandran K, et al. "*Data Funneling: Routing with Aggregation and Compression for WirelessSensor Networks"*. In Proceedings of First IEEE International Workshop on Sensor Network Protocols and Applications. Anchorage. 2003; 156-162.

[9] Ariei T, Gedik B, Ahunbasak Y, et al. *PINCO: "a Pipelined In Network Compression Scheme for Data Collection in Wireless Sensor Networks"*. In Proceedings of 12th International Conference on Computer Communications and Networks. Dallas. 2003; 539-544.

[10] Marcelloni F, Vecchio M. "Enabling Energy-efficient and Lossy-aware Data Compression in Wireless Sensor Networks by Multi-objective Evolutionary Optimization". *Journal of Information Sciences*. 2010; 180(10): 1924-1941.

[11] Schoellhammer T, Greenstein B, Osterweil E, Wimbrow M, Estrin D. "*Lightweight Temporal Compression of Microclimate Datasets"*. in 29th Annual IEEE International Conference on Local Computer Networks. San Francisco. 2004; 516-524.

[12] Song L, Gunopulos D, Kalogeraki V, Lonardi S." *A Data Compression Technique for Sensor Networks with Dynamic Bandwidth Allocation". In TIME '05 Proceedings of the 12th International* Symposium on Temporal Representation and Reasoning. Burlington. 2005; 186-188.

[13] Lee JH, Jung IB. "Adaptive-Compression Based Congestion Control Technique for Wireless Sensor Networks". *Sensors (Basel)*. 2010; 10(4): 2919–2945.

[14] Dong H, Lu JG, Sun YX. “Distributed Wavelet Compression in Wireless Sensor Networks. *Chinese* journal *of sensor and actuators”*. 2007; 20(11): 2481-2486.

[15] Zhu TJ, Lin YP, Zhou SW, Xu XL. “Adaptive Multiple-modalities Data Compression Algorithm Using Wavelet for Wireless Sensor Networks”. *Journal on Communications*. 2009; 30(3): 48-53.

[16] Zhang JM, Lin YP, Zhou SW, Ouyang JC. Harr “Wavelet Data Compression Algorithm with Error Bound for Wireless Sensor Networks”. *Journal of Software*. 2010; 21(6): 1364-1377.

[17] Donoho D. “Compressed Sensing”. *IEEE Transactions on Information Theory*. 2006; 52(4): 1289-1306.

[18] D. Slepian and J.K. Wolf, “Noiseless coding of correlated information sources,’’ *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[20] S.S. Pradhan, J. Kusuma, and K. Ramchandran, “Distributed compression in a dense microsensor network,” *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 51–60, 2002.

[21] Z. Xiong, A.D. Liveris, and S. Cheng, “Distributed source coding for sensor networks,” *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, 2004.

[22] Zhou, S.W.; Lin, Y.P.; Zhang J.M.; Ouyang, J.C.; Lu, X.G. “A wavelet data compression algorithm using ring topology for wireless sensor networks”. *J. Softw.* 2007, *18*, 669-680.

[23] Song, L.; Kalogeraki, V.; Gunopulos, D.; Lonardi, S. “Online Information Compression in Sensor Networks”. In *Proceedings of IEEE International Conference on Communications (ICC’06)*, Istanbul, Turkey, 11–15 June 2006; Volume 7, pp. 3371-3376.

[24] Jim, C.; Dragan, P.; Kannan, R. “A distributed and adaptive signal processing approach to exploiting correlation in sensor networks”, *Ad hoc* Networks. *J. Elsevier.* 2004, *2*, 387-403.