Discovering Application Level Semantics for Data Compression using HCT

Deepa. R, K. John Peter

Abstract— Natural phenomena show that many creatures form large social groups and move in regular patterns. However, previous works focus on finding the movement patterns of each single object or all objects. I propose an efficient distributed mining algorithm to jointly identify a group of moving objects and discover their movement patterns in wireless sensor networks. This algorithm consists of the local mining phase and the cluster ensembling phase. The local mining phase adopts the VMM model together with Probabilistic Suffix Tree to find the moving patterns, as well as Highly Connected Component to partition the moving objects. The cluster ensembling phase utilizes Jaccard Similarity Coefficient and Normalized Mutual Information to combine and improve the local grouping results. The distributed mining algorithm achieves good grouping quality and robustness.

In this paper, I extend it further, and propose a technique called hybrid compression technique based on the location information of nodes in the sensor network. A hybrid compression technique problem is formulated to reduce the amount of energy consumption and increases the lifetime of network. The experimental result shows that the technique have good ability of approximation to manage the sensor network and have high data compression efficiency and leverages the group movement patterns to reduce the amount of delivered data effectively and efficiently.

Index Terms— clustering, hybrid, patterns, similarity

I. INTRODUCTION

Recent advances in location-acquisition technologies, such as global positioning systems (GPSs) and wireless sensor networks (WSNs), have fostered many novel applications like object tracking, environmental monitoring, and location-dependent service. These applications generate a large amount of location data, and thus, lead to transmission and storage challenges, especially in resource constrained environments like WSNs. To reduce the data volume, various algorithms have been proposed for data compression and data aggregation. However, the above works do not address application-level semantics, such as the group relationships and movement patterns, in the location data.

In object tracking applications, many natural phenomena show that objects often exhibit some degree of regularity in their movements. Biologists also have found that many creatures, such as elephant's zebra, whales, and birds, form large social groups when migrating to find food, or for breeding or wintering. These characteristics indicate that the trajectory data of multiple objects may be correlated for biological applications.

Manuscript received on July, 2012.

Deepa.R, Computer science and engineering, Vins Christian college of engineering/ Anna University, Nagercoil, India.

K. John peter, Computer science and engineering, Vins Christian college of engineering/ Anna University, Nagercoil, India.

This raises a new challenge of finding moving animals belonging to the same group and identifying their aggregated group movement patterns. Therefore, under the assumption that objects with similar movement patterns are regarded as a group, and define the moving object clustering problem as given the movement trajectories of objects, partitioning the objects into non-overlapped groups such that the number of groups is minimized. Then, group movement pattern discovery is to find the most representative movement patterns regarding each group of objects, which are further utilized to compress location data.

Discovering the group movement patterns is more difficult than finding the patterns of a single object or all objects, because we need to jointly identify a group of objects and discover their aggregated group movement patterns. The approaches that perform clustering among entire trajectories may not be able to identify the local group relationships. In addition, most of the above works are centralized algorithms, which need to collect all data to a server before processing. Thus, unnecessary and redundant data may be delivered, leading to much more power consumption because data transmission needs more power than data processing in WSNs.

The clustering algorithm itself is a centralized algorithm, and systematically combining multiple local clustering results into a consensus to improve the clustering quality and for use in the update-based tracking network. Thus the problem of compressing the location data of a group of moving objects as the group data compression problem is defined.

Therefore, in this paper, the distributed mining algorithm to approach the moving object clustering problem and discover group movement patterns. Then, based on the discovered group movement patterns, a novel compression algorithm is proposed to tackle the group data compression problem. The distributed mining algorithm comprises a Group Movement Pattern Mining (GMPMine) and Cluster Ensembling (CE) algorithms. It avoids transmitting unnecessary and redundant data by transmitting only the local grouping results to a base station, instead of all of the moving objects' location data.

Specifically, the GMPMine algorithm discovers the local group movement patterns by using a novel similarity measure, while the CE algorithm combines the local grouping results to remove inconsistency and improve the grouping quality by using the information theory. The moving object clustering problem is tackled by using a distributed mining algorithm, which comprises the GMPMine and CE algorithms. First, the GMPMine algorithm uses a PST to generate an object's significant movement patterns and computes the similarity of two objects by using simp to derive the local grouping results.

To combine multiple local grouping results into a consensus, the CE algorithm utilizes the Jaccard similarity coefficient to measure the similarity between a pair of objects, and normalized mutual information (NMI) to derive the final ensembling result.

Wireless sensor network consists of hundreds of inexpensive nodes that can be deployed in environments to collect useful information in a robust and autonomous manner. The key challenges in data gathering are energy efficiency and latency awareness. Hybrid compression technique is used to address how to deliver and process the data coming from all sensor nodes from the sink.

In data aggregation in wireless sensor network allows, set of nodes are selected randomly as cluster head. Each node joins a cluster depending upon the communication between the node and CH. The role of CH allows preserving energy.

Voronoi diagram is a special kind of decomposition of a given space, determined by distances to a specified family of objects in the space. These objects are usually called the sites or the generators and to each such object one associate a corresponding Voronoi cell, namely the set of all points in the given space whose distance to the given object is not greater than their distance to the other objects.

Compressive data sensing allows the approximation of the readings from the sensor field. It can be used to convert nodes data on the sub region and counted as array then process the matrix on to vector before calculating and sending data to the sink. Hybrid compression technique gives the forward vision of the spatial integration and minimizes the overall energy consumption which contributes better solutions on energy conservations.

The proposed HCT exploits the shapes divider on data compression at gateway before sending to sink. A new distributed data aggregation technique HCT based on voronoi diagram is proposed to address the problem of the distribution of WS field. The new techniques have the good ability of approximation and compress the data efficiently and reduce the amount of energy consumption and it increases the lifetime of the network.

II. RELATED WORKS

The energy consumption can be reduced in large-scale sensor networks which systematically sample a spatio-temporal field. A distributed compression problem is subject to aggregation costs to a single sink. It shows that the optimal solution is greedy and based on ordering sensors according to their aggregation costs.[1] A simplified hierarchical model for a sensor network including multiple sinks, compressors/aggregation nodes and sensors. This paper addresses arrangement of distributed compression subject to aggregation costs to a single sink and hierarchical architectures for aggregation/compression in large-scale sensor networks including multiple sinks.[2]

The aim is to minimize overall aggregation costs, associated with gathering sensor information. It maximizes the network lifetime. An optimal hierarchical organization of sensors, aggregation points/compressors, and sinks can be used to minimize the cost of gathering sensor data.[3]

A problem of broadcast communication in a multi-hop sensor network is a problem, in which samples of a random field are collected at each node of the network, and the goal is for all nodes to obtain an estimate of the entire field within a prescribed distortion value. [4]-[7].The main idea is that of jointly compressing the data generated by different nodes as this information travels over multiple hops, to eliminate correlations in the representation of the sampled field. The nodes compress all their data without exchanging any information .The goal is to provide a solid theoretical framework based on rate distortion theory which supports these intuitions. [8][9][10][11][12]

Sensor networks are fundamentally constrained by the difficulty and energy expense of delivering information from sensors to sink.[13] This paper focused on garnering significant energy improvements by devising computationally-efficient lossless compression algorithms on the source node. These reduce the amount of data that must be passed through the network and to the sink, and thus have energy benefits that are multiplicative with the number of hops the data travels through the network.

The distributed nature of the sensor network architecture introduces unique challenges and opportunities for collaborative networked signal processing techniques that can potentially lead to significant performance gains. The low-power sensor network scenarios need to have high spatial density to enable reliable operation in the face of component node failures as well as to facilitate high spatial localization of events. This induces a high level of network data redundancy, where spatially proximal sensor readings are highly correlated. This redundancy can be reduced by a new approach in a completely distributed manner.[14]

Moving object representation and computing have received a fair share of attention over recent years in the spatial database community. This is understandable as positioning technology is rapidly making its way into the consumer market, not only through the already ubiquitous cell phone but soon also through small, on-board devices in many means of transport and in types of portable equipment. It is thus to be expected that all these devices will start to generate an unprecedented data stream of time-stamped positions.[15] Compression techniques aim at substantial reductions in the amount of data without serious information loss. Lossless compressions have no information loss and are often based on optimizing the information capacity per byte used. Lossy compressions do suffer from information loss, and are often based on discarding the least informative data. The central theme of this paper concerns a lossy compression technique for moving object data streams that attempts to preserve the major characteristics of the original trajectory.

III. PROBLEM DESCRIPTION

A. Problem Description:

The problem is formulated as exploring the group movement patterns to compress the location sequences of a group of moving objects for transmission efficiency. Consider a set of moving objects $O=\{o_1,o_2,\ldots o_n\}$ and their associated location sequence dataset $S=\{S_1,S_2,\ldots S_n\}$.

Definition 1: Two objects are similar to each other if their movement patterns are similar. Given the similarity measure function sim_p[2] and a minimal threshold sim_{min}, o_i and o_j are similar if their similarity score sim_p (O_i,O_j) is above sim_{min}.The set of objects that are similar to o_i is denoted by $so(oi)=\{o_i|\forall\ o_j\in O, sim_p(o_i,o_j)\geq sim_{min}\}$.

Definition 2: A set of objects is recognized as a group if they are highly similar to one another. Let g denote a set of objects. G is a group if every object in g is similar to atleast a threshold of objects in g.

B. Network and Location Models:

A sensor cluster is a mesh network of n ×n sensor nodes handled by a CH and communicate with each other by using multihop routing. We assume that each node in a sensor cluster has a locally unique ID and denote the sensor IDs by an alphabet ∑.An object is defined as a target, such as an animal or a bird that is recognizable and trackable by the tracking network. To represent the location of an object, geometric models and symbolic models are widely used. A geometric location denotes precise two-dimension or three-dimension coordinates; while a symbolic location represents an area, such as the sensing area of a sensor node or a cluster of sensor nodes, defined by the application. Since the accurate geometric location is not easy to obtain and techniques like the Received Signal Strength (RSS) can simply estimate an object's location based on the ID of the sensor node with the strongest signal, we employ a symbolic model and describe the location of an object by using the ID of a nearby sensor node.

Object tracking is defined as a task of detecting a moving object's location and reporting the location data to the sink periodically at a time interval. Hence, an observation on an object is defined by the obtained location data. Assume that sensor nodes wake up periodically to detect objects.

C. Variable Length Markov Model (VMM) and Probabilistic Suffix Tree (PST):

The VMM, an object's movement is expressed by a conditional probability distribution over∑. Let s denote a pattern which is a subsequence of a location sequence S and σ denote a symbol in∑. The conditional probability P(σ|s) is the occurrence probability that σ will follow s in S. Since the length of s is floating, the VMM provides flexibility to adapt to the variable length of movement patterns.

Probabilistic Suffix Tree (PST) can be used for learning the significant movement patterns and it has the lowest storage requirement among many VMM implementations. PST's low complexity and makes it more attractive especially for streaming or resource-constrained environments. The PST building algorithm1 learns from a location sequence and generates a PST whose height is limited by a specified parameter L_{max}. Each node of the tree represents a significant movement pattern s whose occurrence probability is above a specified minimal threshold P_{min}. It also carries the conditional empirical probabilities P(σ|s) for each σ in ∑ that use in location prediction. PST is frequently used in predicting the occurrence probability of a given sequence, which provides us important information in similarity comparison. The occurrence probability of a sequence s regarding to a PST T, denoted by $P^T(s)$, is the prediction of the occurrence probability of s based on T.

IV. MINING GROUP MOVEMENT PATTERNS

A set of moving objects O together with their associated location sequence data set S and a minimal similarity threshold sim_{min}, the moving object clustering problem is to partition O into nonoverlapped groups, denoted by $G=\{g_1,g_2,\ldots g_i\}$, such that the number of groups is minimized. This is known as object clustering problem.

The moving object clustering problemis tackled by using a distributed mining algorithm, which comprises the GMPMine and CE algorithms. First, the GMPMine algorithm uses a PST to generate an object's significant movement patterns and computes the similarity of two objects by using simp to derive the local grouping results.

To combine multiple local grouping results into a consensus, the CE algorithm utilizes the Jaccard similarity coefficient to measure the similarity between a pair of objects, and normalized mutual information (NMI) to derive the final ensembling result.

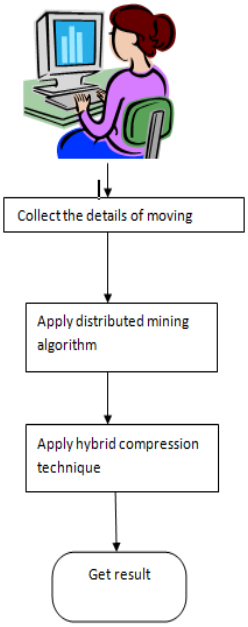

Fig: Diagram for Techniques

We have found that many creatures, such as elephants, zebra, whales, and birds, form large social groups when migrating to find food, or for breeding or wintering. These characteristics indicate that the trajectory data of multiple objects may be correlated for biological applications. Moreover, some research domains, such as the study of animal's social behavior and wildlife migration, are more concerned with the movement patterns of groups of animals. These details are given as an input data. A novel compression algorithm to compress the location data of a group of moving objects with or without loss of information. Formulate the Hybrid compression technique to minimize the entropy of location data and explore the Hybrid compression technique (HCT) based on voronoi diagram to prolong the lifetime of the whole network to a greater degree. We also prove that the proposed compression algorithm obtains the optimal solution of the HCT problem efficiently.

A. The Group Movement Pattern Mining Algorithm

The GMPMine algorithm uses a PST to generate an object's significant movement patterns and computes the similarity of two objects by using sim_p to derive the local grouping results. simp can compare the similarity of two data streams only on the changed mature nodes of emission trees instead of all nodes. The similarity score simp of oi and oj based on their respective PSTs, Ti and Tj, is defined as follows:

$$sim_p(o_i, o_j) = -\log \frac{\sum_{s \in \tilde{S}} \sqrt{\sum_{\sigma \in \Sigma} \left(P^{T_i}(s\sigma) - P^{T_j}(s\sigma)\right)^2}}{2L_{max} + \sqrt{2}}, \quad \text{------ (1)}$$

The similarity score simp includes the distance associated with a pattern s, defined as,

Four steps:

Step 1. Learning movement patterns for each object.

Step 2. Computing the pair-wise similarity core to constructing a similarity graph.

Step 3. Partitioning the similarity graph for highly-connected sub graphs.

Step 4. Choosing representative movement patterns for each group.

B . Cluster Ensembling Algorithm

Objects scattered in grassland may not be identified as a group. Furthermore, in the case where a group of objects moves across the margin of a sensor cluster, it is difficult to find their group relationships. Therefore, the CE algorithm is proposed to combine multiple local grouping results. The algorithm solves the inconsistency problem and improves the grouping quality.

To combine multiple local grouping results into a consensus, the CE algorithm utilizes the Jaccard similarity coefficient to measure the similarity between a pair of objects, and normalized mutual information (NMI) to derive the final ensembling result. It trades off the grouping quality against the computation cost by adjusting a partition parameter. The ensembling problem involves finding the partition of all moving objects O that contains the most information about the local grouping results. NMI can be used to evaluate the grouping quality.

A finer-grained configuration of D achieves a better grouping quality but in a higher computation cost. Therefore, for a set of thresholds D, we rewrite our objective function as G,

$$G_{\delta'} = arg\max_{G_\delta \delta \in D} \sum_{i=1}^{K} NMI(G_i, G_\delta). \quad \text{-----(2)}$$

Three steps:

1. Measuring the pair-wise similarity to construct a similarity matrix by using Jaccard coefficient.
2. Generating the partitioning results for a set of thresholds based on the similarity matrix.
3. Selecting the final ensembling result.

V. HYBRID COMPRESSION TECHNIQUE

Wireless sensor network consists a collection of inexpensible nodes called sensor nodes that is responsible for monitoring and collecting useful information in robust manner. Compressive sensing is a new method in data gathering and processing before being sent to the base. The key challenges in data gathering are energy efficiency and latency awareness.HCT addresses how to deliver and process the data from all sensor nodes to the sink and communicate between all nodes on sink through cluster head.

A. Hybrid compression problem

Assume sensor network containing thousands of nodes and these nodes are deployed randomly. Each node will start routing with neighbor nodes after deployment. The sink will generate a query message to all nodes to know the location for nodes. The GPS enabled to allow sensor nodes to get and broadcast their locations. The next process is to implement voronoi tessellation to divide the field into small shape. Each shape contains different number of nodes. Gateway is used to collect the information from all nodes on its own sub regions then compress the data and sent to sink.

Voronoi diagram is a special kind of decomposition of a given space, determined by distances to a specified family of objects in the space. Voronoi tessellations of regular lattices of points in two or three dimensions give rise to many familiar tessellations. Higher-order Voronoi diagrams can be generated recursively. The process of constructing the Voronoi diagram for *n* point sites can be seen as an assignment of a planar convex region to each of the sites, according to the nearest neighbor rule.

Slepian-wolf coding is distributed source coding techniques that can be used to remove data redundancy without requiring inter sensor communication. It deals with lossless compression of two or more correlated data streams.

Wireless sensor node counted as most important technologies, it is consist of a collection of nodes called sensor nodes, and these nodes are responsible for monitoring and collecting information about different physical phenomena. In HCT the sub-region node will represented as matrix M each node will take one value of M, then M will be converted in gateway G to vector, in order to be send to sink.

$$\begin{bmatrix} X1 & X2 & \cdots & \\ \vdots & & \ddots & \vdots \\ & & \cdots & Xn \end{bmatrix} \text{------ (3)}$$

Where each sensor node is represented by one element of the matrix and this matrix is converted into vector,

$$\tilde{v}=\{x_1, x_2, x_3 \ldots x_n\} \text{------ (4)}$$

The HCT can use number for active sensor node less than other technique; more over inter-region transmission cost is much less than normal transmission that each node in sub-region connected directly to sink. Since the number of sensor on each sub region compared with sub- region area, thus the number of active node is reduced. HCT assumption the values it get from sensor is 1 or 0 based on that the amount of data transferred from gateway to sink is reduced.

VI. RESULT

The distributed mining algorithm is used to locate the objects and their discovered patterns. The GMP mime algorithm can measure the similarity between two objects based on the probabilistic suffix tree. The clustering algorithm uses the jaccard similarity coefficient and normalized mutual information (NMI).It improves the grouping quality. The existing algorithms such as FP growth and apriori algorithms still focus on discovering frequent patterns of all objects and may suffer from computing efficiency or memory problems.

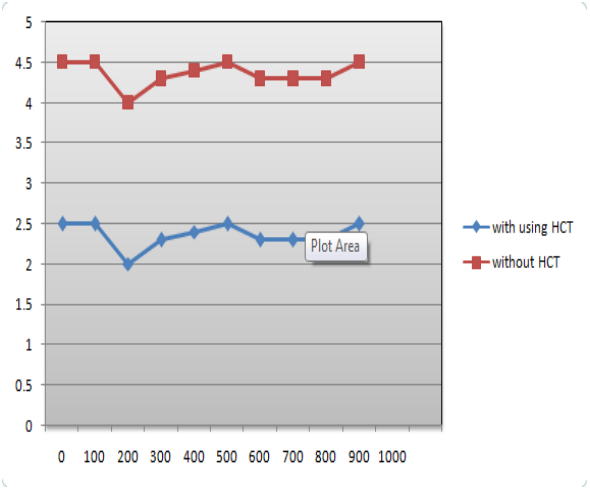

Fig: Average energy consumed with HCT

The distributed mining algorithm focuses on discovering group of objects and their movement patterns. The batch-based approach is compared with an online approach for the overall system performance evaluation and study the impact of the group size (n), as well as the group dispersion radius (GDR), the batch period (D), and the error bound of accuracy (eb). The GDR is used to control the dispersion degree of the objects. Smaller GDR implies stronger group relationships. The compression ratio is defined as the ratio between the uncompressed data size and the compressed data size.

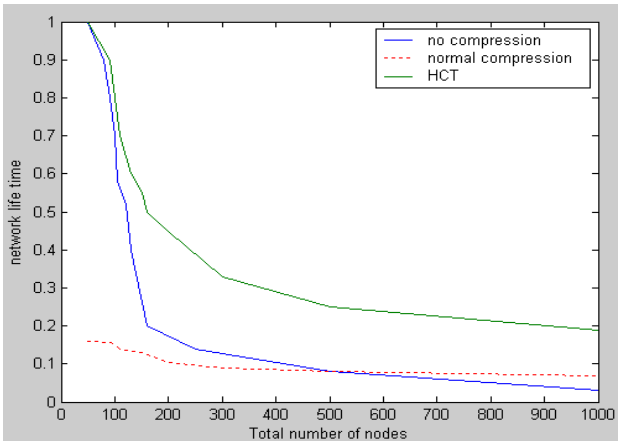

Fig: Networks life time with HCT techniques

The compression ratio is defined as the ratio between the uncompressed data size and the compressed data size. .The amount of data of our batch based approach is relatively low and stable as the GDR increases. The simulations try to study the cases transmission without compression, transmission with usual compression and HCT. network are divided according to Voronoi tessellation into sub-region the small regions considered to be one group and to have one Gateway to collect and transmit data to sink which is assumed to be located at origin. Where in HCT the inter-region communication is achieved then the data will be sent to sink through gate way after converting and compression into vector, and inner communication cost is less than normal transmission.

The quality of precision in the data received at end that obtained by HCT not less than other techniques. In order to test the performance of the data compression, provides a comparison of the average energy cost of the node in the network . After running the transform, the average energy cost of node change from 4.48 to 2.36. So the transform can effectively reduce the average energy cost of node and prolong the lifetime of the whole network to a greater degree.

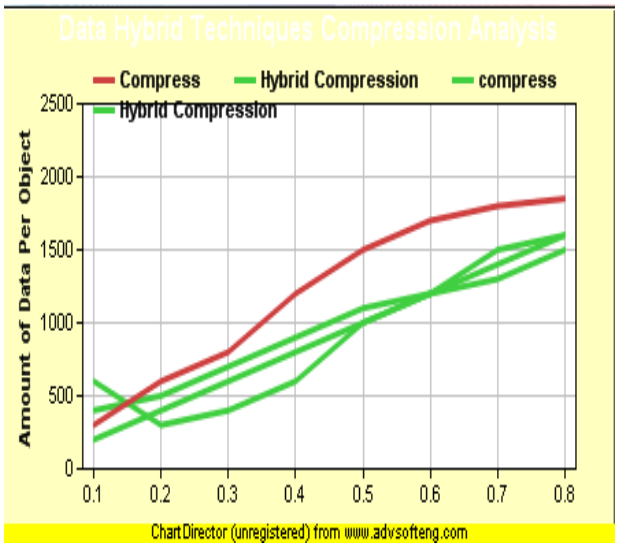

Fig: comparison of Hybrid compression Technique

In the above figure defines the energy consumption of replace and Huffman algorithms compared with hybrid compression technique. The hybrid compression technique reduces the energy consumption. The simulation results shows that the energy consumption due to the compression algorithm is reduced by using the hybrid compression technique.

VII. CONCLUSION

In this work, exploit the characteristics of group movements to discover the information about groups of moving objects in tracking applications. I proposed a novel distributed mining algorithm that consists of the GMPMine algorithm and the Cluster Ensembling algorithm to leverage the object moving patterns in grouping objects. With the discovered information, devise the Hybrid compression technique, which have the good ability of approximation to manage the sensor field and have high ability and efficiency on data compression. The experimental results show that the proposed compression algorithm effectively reduces the amount of delivered data and enhances compressibility and reduces the amount of energy consumption. The compression algorithm increases the life time of the sensor network. The advantage realized by HCT in resulting in a minimum overall energy consumption reducing energy consumption which contributes better solutions on energy conservation. The HCT have the good ability of approximation to manage the sensor field and have the high ability and efficiency on data compressing and reduce the amount of energy consumption and increases the lifetime of the network.

REFERENCES

1. S.S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed Compression in a Dense Micro sensor Network," IEEE Signal Processing Magazine, vol. 19, no. 2, pp. 51-60, Mar. 2002.
2. A. Scaglione and S.D. Servetto, "On the Interdependence of Routing and Data Compression in Multi-Hop Sensor Networks,"Proc. Eighth Ann. Int'l Conf. Mobile Computing and Networking, pp. 140-147, 2002.
3. N. Meratnia and R.A. de by, "A New Perspective on Trajectory Compression Techniques," Proc. ISPRS Commission II and IV, WG II/5, II/6, IV/1 and IV/2 Joint Workshop Spatial, Temporal and Multi-Dimensional Data Modeling and Analysis, Oct. 2003.
4. S. Baek, G. de Veciana, and X. Su, "Minimizing Energy Consumption in Large-Scale Sensor Networks through Distributed Data Compression and Hierarchical Aggregation," IEEE J. Selected Areas in Comm., vol. 22, no. 6, pp. 1130-1140, Aug. 2004.
5. C.M. Sadler and M. Martonosi, "Data Compression Algorithms for Energy-Constrained Devices in Delay Tolerant Networks," Proc. ACM Conf. Embedded Networked Sensor Systems, Nov. 2006.
6. I.F. Akyldiz, W. Su, Y. Sankarasubermanian, and E. Cayirici, "A survey on sensor networks," IEEE Communications Magazine, vol. 40, no. 8, pp. 102–114, August 2002.
7. A.J. Goldsmith and S.B Wicker, "Design challenges for energy constrained ad hoc wireless networks," IEEE Wireless Communications, vol. 9, no. 4, 2002.
8. S. S. Pradhan, J. Kusuma, and K. Ramachandran, "Distributed compression in a dense mircosensor network," IEEE Signal Processing Magazine, pp. 51–60, March 2002.
9. A. Scaglione and S. D. Servetto, "On the interdependence of routing and data compression in multi-hop sensor networks," in Proc. ACM Mobicom, 2002.
10. J. Chou, D. Petrovic, and K. Ramchandran, "A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks," in Proc. IEEE Infocom, 2003.
11. W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless micro sensor networks," in Hawaii International Conference on System Sciences, 2000.
12. S.S. Pradhan and K. Ramchandran, "Distributed source coding: Symmetric rates and applications to sensor networks," in *Proc.* IEEE Data Compression *Conf.*, Snowbird, UT, Mar. 2000, pp. 363-372.
13. T. M. Cover and J. Thomas. Elements of Information Theory. John Wiley and Sons, Inc., 1991.
14. A. Gersho and R. M. Gray. Vector Quantization and Signal Compression. Kluwer Academic Publishers, 1992.
15. Q. Zhao and M. Effros. Optimal Code Design for Lossless and Near Lossless Source Coding in Multiple Access Networks. In Proc. Data Compression Conf., Snowbird, UT, 2001