

# Discovering Application Level Semantics for Data Compression using Hybrid Compression Technique

B. Sasikumar, R. Deepa

**Abstract**—Based on Natural phenomena many creatures form large social groups and move in regular patterns. Traditional works focus on finding the movement patterns of each single object or all objects. This paper propose an efficient distributed mining algorithm to jointly identify a group of moving objects and discover their movement patterns in Wireless Sensor Networks (WSN). The algorithm consists of a local mining phase and a cluster ensembling phase. The local mining phase adopts the Variable Length Markov (VMM) model together with Probabilistic Suffix Tree (PST) to find the moving patterns, as well as Highly Connected Component (HCC) 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. Moreover this paper extends a technique called Hybrid Compression Technique (HCT) based on the location information of nodes in the WSN. HCT is formulated to reduce the amount of energy consumption and increases the lifetime of the WSN. The experimental result shows that the technique have good ability of approximation to manage the WSN, 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, compression, hybrid, patterns, semantics, wireless sensor networks.

## I. INTRODUCTION

Recently advances in location-acquisition technologies, such as Global Positioning Systems (GPS) and WSN, have fostered many novel applications like object tracking, environmental monitoring, and location-dependent service. These applications generate a large amount of location data, and lead to transmission and storage challenges, especially in resource constrained environments like WSNs. To reduce the data volume, different algorithms have been proposed for data compression and data aggregation. But, the conventional 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 many creatures, such as elephant, 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.

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. 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 of the 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 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. The problem of compressing the location data of a group of moving objects as the group data compression problem is defined. In this paper, the distributed mining algorithm to approach the moving object clustering problem and discover group movement patterns is discussed. Based on the group movement patterns, a compression algorithm is proposed to tackle the group data. The distributed mining algorithm comprises a Group Movement Pattern Mining (GMPMine) and Cluster Ensembling (CE) algorithms, which 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.

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

**Manuscript received on July, 2012.**

**B. Sasikumar**, Professor, Department of CSE, The Rajaas Engineering College, Vadakkankulam, Thirunelveli District,

**R. Deepa**, Computer Science and Engineering, Vins Christian College of Engineering, Nagercoil.

GMPMine algorithm uses a PST to generate an objects significant movement patterns and computes the similarity of two objects 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. WSN 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. HCT addresses how to deliver and process the data coming from all sensor nodes from the sink. In data aggregation, set of nodes are selected randomly as Cluster Head (CH). Each node joins a cluster depending upon the communication between the node and CH to preserve 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 and is used to convert data on the sub region and counted as array then process the matrix on to vector before calculating and sending data to the sink. HCT gives the forward vision of the spatial integration and minimizes the overall energy consumption which contributes better solutions on energy conservations. 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 in WSN. The new technique has good ability of approximation and compresses the data efficiently and reduces the amount of energy consumption and 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. This 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 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. This 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]-[12].

Sensor networks are fundamentally constrained by the difficulty and energy expense of delivering information from sensors to sink [13]. This 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 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. To the expectation 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 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  $O = \{o_1, o_2, \dots, o_n\}$  and their associated location sequence dataset  $S = \{S_1, S_2, \dots, S_n\}$ .

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

**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  $n \times n$  sensor nodes handled by a CH and communicate with each other by using multihop routing. Assume that each node in a sensor cluster has a locally unique ID and denote the sensor IDs by an alphabet  $\Sigma$ . 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 objects location based on the ID of the sensor node with the strongest signal, 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 objects location and reporting the location data to the sink periodically at a time interval. An observation on an object is defined by the obtained location data assuming that sensor nodes wake up periodically to detect objects.

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

The VMM, an objects movement is expressed by a conditional probability distribution over  $\Sigma$ . Let  $s$  denote a pattern which is a subsequence of a location sequence  $S$  and  $\sigma$  denote a symbol in  $\Sigma$ . The conditional probability  $P(\sigma|s)$  is the occurrence probability that  $\sigma$  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. PSTs low complexity and makes it more attractive especially for streaming or resource-constrained environments. The PST building algorithm learns from a location sequence and generates a PST whose height is limited by a specified parameter  $L_{max}$ . Each node in the tree represents a significant movement pattern  $s$  whose occurrence probability is above a specified minimal threshold  $P_{min}$ . This also carries the conditional empirical probabilities  $P(\sigma|s)$  for each  $\sigma$  in  $\Sigma$  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 non-overlapped groups, denoted by  $G = \{g_1, g_2, \dots, g_i\}$ , such that the number of groups is minimized. This is known as object clustering problem. 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 objects significant movement patterns and computes the similarity of two objects by using  $sim_p$  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 NMI to derive the final ensembling result.

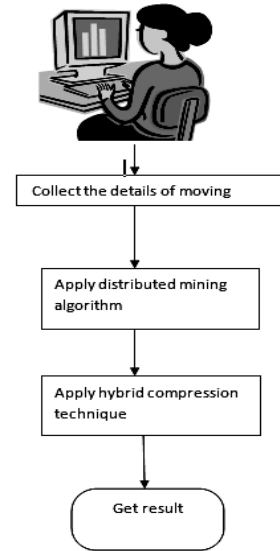


Fig. 1 Diagram for Techniques

It is 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. Some research domains, such as the study of animals 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 is used to compress the location data of a group of moving objects with or without loss of information. Formulate the HCT to minimize the entropy of location data and explore the HCT based on Voronoi diagram to prolong the lifetime of the whole network to a greater degree and is proved that the proposed algorithm obtains the optimal solution efficiently.

### A. The Group Movement Pattern Mining Algorithm

The GMPMine algorithm uses a PST to generate an objects significant movement patterns and computes the similarity of two objects by using  $sim_p$  to derive the local grouping results.  $sim_p$  can compare the similarity of two data streams only on the changed mature nodes of emission trees instead of all nodes. The similarity score  $sim_p$  of  $o_i$  and  $o_j$  based on their respective PSTs,  $T_i$  and  $T_j$ , is defined as follows:

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

The similarity score  $sim_p$  includes the distance associated with a pattern  $s$ , defined in the following 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 group relationships. 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 NMI to derive the final ensembling result. This 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. For a set of thresholds  $D$ , we rewrite our objective function as  $G$ ,

$$G_D = \operatorname{argmax}_{G_i \in D} \sum_{i=1}^K \operatorname{NMI}(G_i, G_\delta). \quad \text{-----(2)}$$

Three steps:

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

### V. HYBRID COMPRESSION TECHNIQUE

WSN consists a collection of inexpensive 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

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.

$$\begin{bmatrix} X_1 & X_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & X_n \end{bmatrix} \quad \text{-----(3)}$$

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

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

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 from sensor is 1 or 0 based on that the amount of data transferred from gateway to sink is reduced.

### VI. RESULTS

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 NMI. This 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.

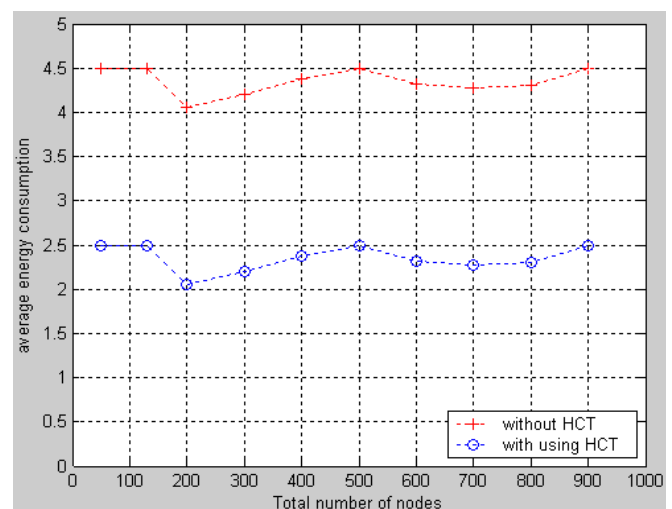


Fig. 2 Average energy consumed with HCT

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.

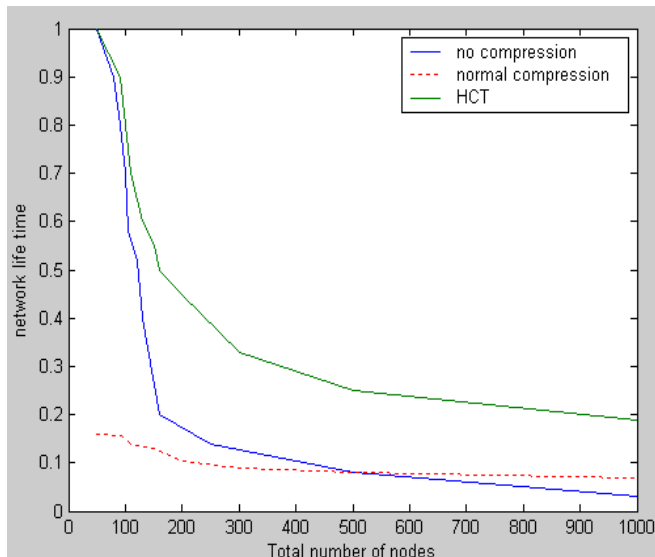


Fig. 3 Networks life time with HCT techniques

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 study the cases transmission without compression, transmission with usual compression and HCT. Networks 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 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.

## VII. CONCLUSION

This work, exploit the characteristics of group movements to discover the information about groups of moving objects in tracking applications. The distributed mining algorithm 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 HCT, 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," *Proceedings of 8th Annual International Conference on Mobile Computing and Networking*, pp. 140-147, 2002.
- [3] N. Meratnia and R. A. De By, "A New Perspective on Trajectory Compression Techniques," *Proceedings of ISPRS Commission II and IV, WG II/5, II/6, IV/1 and IV/2 Joint Workshop Spatial, Temporal and Multi-Dimensional Data Modelling 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 Journal on Selected Areas in Communications*, 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," *Proceedings of ACM Conference on 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. Ramchandran, "Distributed Compression in a Dense Microsensor 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," *Proceedings of 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," *Proceedings of 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," *Proceedings of IEEE Data Compression Conference, Snowbird, UT*, pp. 363-372, Mar. 2000.
- [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," *Proceedings of Data Compression Conference, Snowbird, UT*, 2001.