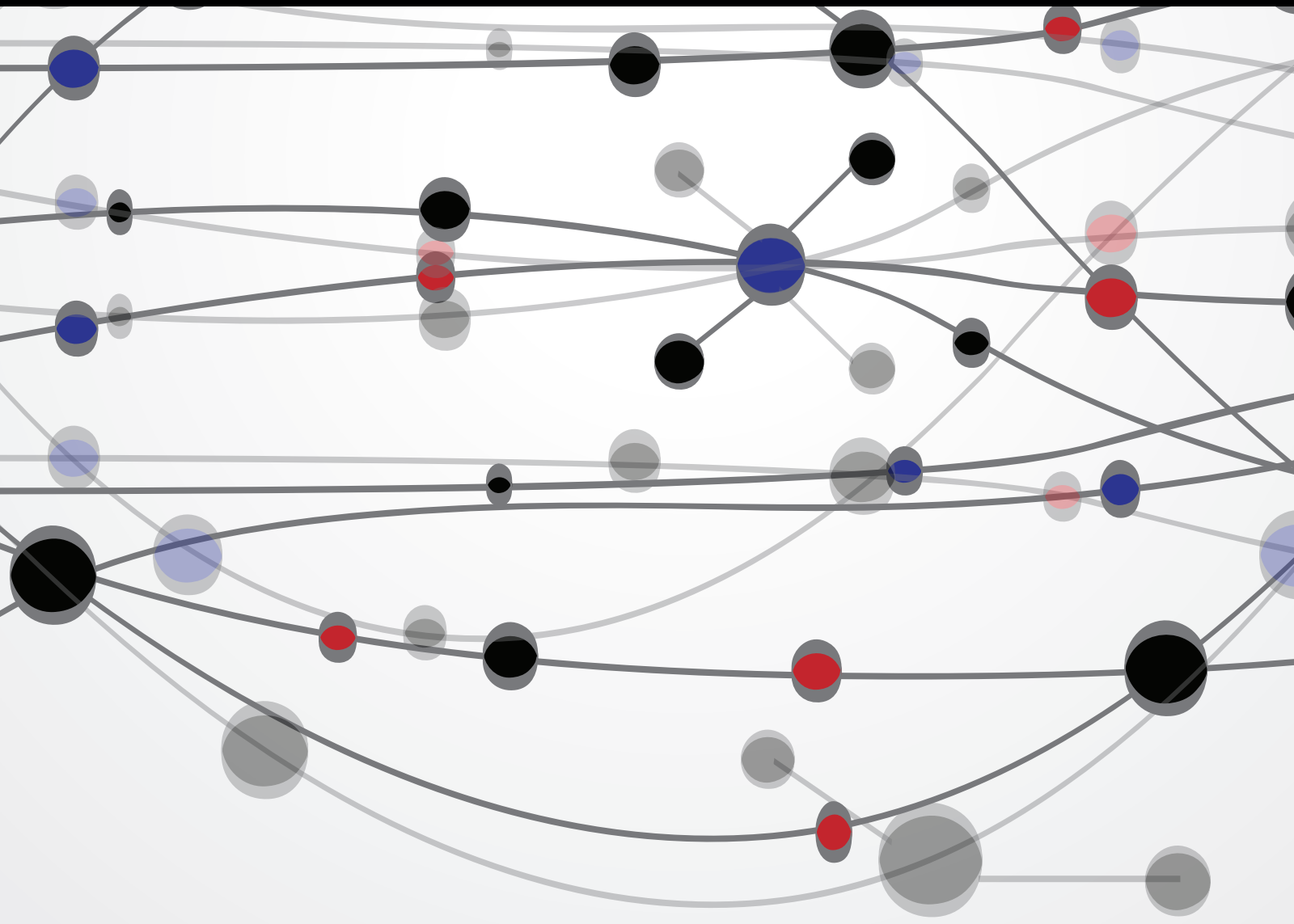
The Scientific World Journal

Recent Advances on Internet of Things

Guest Editors: Xiaoxuan Meng, Jaime Lloret, Xudong Zhu, and Zhongmei Zhou



Recent Advances on Internet of Things

The Scientific World Journal

Recent Advances on Internet of Things

Guest Editors: Xiaoxuan Meng, Jaime Lloret, Xudong Zhu, and Zhongmei Zhou

Contents

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 709345, 1 page
http://dx.doi.org/10.1155/2014/709345

Editorial

Recent Advances on Internet of Things

Xiaoxuan Meng,[1] Jaime Lloret,[2] Xudong Zhu,[3] and Zhongmei Zhou[4]

[1] *VMware, 3401 Hillview Avenue, Palo Alto, CA 94304, USA*
[2] *Department of Communications, Universidad Politecnica de Valencia, Spain*
[3] *Department of Computer Science and Engineering, Minnan Normal University, Zhangzhou, China*
[4] *School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*

Correspondence should be addressed to Xudong Zhu; zhuxd.ieit@gmail.com

Received 18 November 2014; Accepted 18 November 2014; Published 22 December 2014

Recently, internet of things (IoT) has become increasingly ubiquitous. IoT enables a detailed characterization of the physical environment, as well as a rich set of interactions with the physical world. Therefore, IoT has the potential to revolutionize pervasive computing and its applications. The success of intelligent IoT highly depends on the system architectures, networks and communications, data processing, and ubiquitous computing technologies, which support efficient and reliable physical and cyber interconnections. Indeed, the realization of a ubiquitous IoT poses several challenges about seamless integration, heterogeneity, scalability, mobility, and many others.

In this special issue, we mainly focus on the latest advancements of IoT. We invite scientists and investigators to contribute to this special issue with original research articles and review articles on theories and key technologies for scientific and engineering problems in IoT, as well as their applications to conquer engineering problems.

Generally, the accepted papers in this special issue can be divided into several categories. Most of the papers focus on the key technologies for IoT, like wireless sensor networks (WSNs), video and image processing, distributed systems, and so forth. There are also papers related to algorithms and model for IoT, for example, bee colony algorithm, parallelized dilate algorithm, color gamut description algorithm, and so forth. The rest of the papers are talking about the applications of IoT in real life.

In all, we hope that readers will find in this special issue not only the new ideas, cutting-edge information, new technologies, and applications of IoT but also a special emphasis on how to solve various emerging problems by using new technologies and algorithms.

Xiaoxuan Meng
Jaime Lloret
Xudong Zhu
Zhongmei Zhou

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 476253, 13 pages
http://dx.doi.org/10.1155/2014/476253

Research Article

LPTA: Location Predictive and Time Adaptive Data Gathering Scheme with Mobile Sink for Wireless Sensor Networks

Chuan Zhu,[1] Yao Wang,[1] Guangjie Han,[1] Joel J. P. C. Rodrigues,[2,3] and Jaime Lloret[4]

[1] *College of Internet of Things Engineering, Hohai University, Changzhou 213022, China*
[2] *Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal*
[3] *University of ITMO, Saint Petersburg 197101, Russia*
[4] *Integrated Management Coastal Research Institute, Universidad Politecnica de Valencia, 46022 Valencia, Spain*

Correspondence should be addressed to Guangjie Han; hanguangjie@gmail.com

Received 24 April 2014; Accepted 24 July 2014; Published 3 September 2014

Academic Editor: Zhongmei Zhou

This paper exploits sink mobility to prolong the lifetime of sensor networks while maintaining the data transmission delay relatively low. A location predictive and time adaptive data gathering scheme is proposed. In this paper, we introduce a sink location prediction principle based on loose time synchronization and deduce the time-location formulas of the mobile sink. According to local clocks and the time-location formulas of the mobile sink, nodes in the network are able to calculate the current location of the mobile sink accurately and route data packets timely toward the mobile sink by multihop relay. Considering that data packets generating from different areas may be different greatly, an adaptive dwelling time adjustment method is also proposed to balance energy consumption among nodes in the network. Simulation results show that our data gathering scheme enables data routing with less data transmission time delay and balance energy consumption among nodes.

1. Introduction

With the enormous development in the field of embedded computing and wireless communication technology, wireless sensor networks (WSNs) nowadays are more applicable and practicable than the WSNs in the past were. A wireless sensor network is composed of hundreds or thousands of distributed sensors that monitor their interesting surroundings and report sensed data to a static base station or sink through multihop relay. Typical applications of WSN include target tracking, environment monitoring, military surveillance, health monitoring, natural disasters forecasting, and so forth [1–3]. Sensors with limited energy are usually left unattended after the initial deployment. And the environments of the deployment area are often harsh and involve obstacles, which makes sensors battery replacement infeasible. Hence, it is expected to minimize and balance energy consumption of sensors to prolong the network lifetime. It has been proved that in a static network, the sensors deployed near the sink exhaust their energy faster than those far apart due to their heavy overhead of relaying messages, and this is the so called "hot-spot" problem [4]. In addition, node failure or malfunction may cause energy holes in the deployed area, and the network connectivity and coverage around the sink may not be guaranteed [5]. Therefore, unbalanced energy consumption causes network performance to be degraded and network lifetime shortened.

Recently, various new strategies that use mobility attributes of elements in WSNs have been investigated to reduce and balance energy consumption of sensors [6, 7]. In this paper, we consider utilizing a mobile sink to proactively collect data generating from sensor network, and this strategy is also favored by many researchers [8–10]. When the sink moves in the network, the role of the "hot-spot" rotates among sensors [11], resulting in balanced energy consumption among nodes. The effectiveness of using mobile sinks to gather data has been demonstrated both by theoretical analysis and experimental study [12–15].

Generally, when a mobile sink is adopted to collect data all over the network, it should consider the following two requirements.

(i) *Low Data Delivery Latency*. Cheap sensors deployed in the monitored environment are often equipped with limited resources, that is, energy, memory space, and so forth. And the speed of a mobile sink is relatively slow compared with the speed of wireless wave. These features may cause a long data transmission delay, which will lead to sensor nodes out of memory and packets lost. Therefore, reducing data delivery latency is necessary, especially in time-sensitive applications.

(ii) *Less Control Message Overhead*. When a sink moves, it must broadcast its current location to the sensor nodes in the network repeatedly. The broadcasting among sensors consumes a large amount of energy, resulting in rapid energy depletion. Due to the limitation and the preciousness of the energy resource, it is particularly important to reduce the energy consumption of the control overhead among sensor nodes.

In this paper, we propose location predictive and time adaptive (LPTA) data gathering scheme with mobile sink for wireless sensor networks, which extends the lifetime of the network with less sink's location updating overhead and reduces the data transmission latency. The data gathering process is composed of multiple data gathering periods. Each data gathering period consists of three phases: loose time synchronization phase, date collection phase, and data gathering period ending declaration phase. The following primary features make our LPTA scheme differentiated from other existing data gathering algorithms [8, 9, 16].

(i) LPTA Needs No Location Updating Information for the Mobile Sink. Many existing data gathering algorithms, for example, ALURP [8, 9], need to broadcast the mobile sink location information during data gathering process. Our proposed scheme is based on a loose time synchronization mechanism among sensor nodes and the mobile sink. Therefore, each node in the network is able to calculate the position of the mobile sink based on their local clock. It significantly reduces the control messages overhead and makes the data reporting at any time possible for every sensor node.

(ii) LPTA Is an Adaptive Algorithm in Terms of Sink's Dwelling Time. In many scenarios, the deployed environments are not ideal, and often involve obstacles. The deployment of sensor nodes in such environments is nonuniform. Additionally, the frequentness and number of the events of interest may be different for different part of the deployment area during different periods. In LPTA, according to the quantity of the data generated from different area of the network, the mobile sink's dwelling time at different sojourn points can be adjusted adaptively, resulting in more efficient energy consumption of nodes.

The contributions of this paper can be summarized as follows.

(1) LPTA Utilizes Location Prediction to Reduce Data Transmission Delay and Location Broadcasting Overhead. The loose time synchronization at the beginning of each data gathering period and the mobile sink's fixed moving track with constant speed make the calculation of mobile sink location possible.

(2) LPTA Avoids Getting Sensor Nodes out of Memory. When there is data to report or relay to the mobile sink, sensor nodes can calculate the location of mobile sink and send out the data immediately. Therefore, sensor nodes need not cache the data for a long time and wait for the sink moving to neighboring scope.

(3) LPTA Takes Obstacles into Accoun, and Can Adjust the Sink Dwelling Time at Sojourn Points during Data Gathering. The criteria for the time adaptation is based on the quantity of history data generated in each area. When gathering data, the mobile sink will dwell longer time in the area with more data to send than those with less, resulting in the average relay hops decreased and the energy efficient for data transmission.

As the implication of of LPTA abbreviation, our LPTA mainly focuses on two strategies:

(1) mobile sink location prediction,

(2) dwelling time adjustment of mobile sink.

During the process of data reporting from nodes to the mobile sink, shortest data routing protocol or other existing routing protocols can be adopted, as long as data can be relayed to the mobile sink based on location information. To simplify our scheme and focus on the two strategies of LTPA, we use shortest path routing protocol as the data routing protocol.

The rest of this paper is organized as follows. We present related work in Section 2. The network model and problem statement of our scheme are given in Section 3. Then, the moving strategy of the mobile sink is presented in Section 4, and the data reporting process of nodes towards the mobile sink is described in Section 5. And in Section 6, the performance of our protocol is analyzed and compared with that of the adaptive sink mobility scheme (*Adaptive*) [16] in terms of data transmission latency and energy consumption through simulations. Finally, the conclusion is given in Section 7.

2. Related Work

Here, we briefly summarize some of the related works on data gathering mechanisms in WSNs. According to the main purpose of these works, the existing protocols can be classified into three categories: *extension of network lifetime*, *reduction of sink's location updating overhead,* and *reduction of time delay of data reporting*.

2.1. Extension of Network Lifetime. To alleviate the influence of "hot-spot," a number of research works on prolonging the lifetime of WSNs by using mobile sinks have been proposed. Wang et al. [17] explored sink mobility to prolong the lifetime of sensor networks. They gave a linear programming formulation for the joint problems of determining the movement schedule of the sink and the sojourn time at different points in

the network. Their proposed routing scheme can work only in a grid network topology. Luo and Hubaux [15] proved that in a circle topology, to achieve the maximum network lifetime, a mobile sink should rotate on the periphery of the network. A joint mobility and routing strategy with a combination of periphery moving and round routing was proposed. Then, by assuming Manhattan routing, Lee et al. [10] obtained the similar conclusion. They also proposed a heuristic algorithm for sink mobility to achieve near-optimal network lifetime. Liu et al. [18] proposed a density adjustment algorithm in order to increase the network lifetime and coverage by appropriately adjusting node density. A biased adaptive sink mobility scheme (*Adaptive*) was proposed in [16]. In order to achieve accelerated coverage of the network and fairness of service time of each region, the sink moves probabilistically, favoring less visited areas and adaptively staying longer in network regions that tend to produce more data. *Adaptive* balances the energy consumption among nodes and prolongs the network lifetime. However, because the mobile sink has to traverse all vertexes in the graph, it may cause a rigorous time delay problem in large scale networks.

2.2. Reduction of Sink's Location Updating Overhead. Sink's location updating may cause great energy overhead of nodes. Therefore, some schemes have been suggested to reduce location updating control messages. Ye et al. [19] presented a Two-Tier Data Dissemination (TTDD) protocol in which each data source proactively constructs a grid structure enabling mobile sinks to continuously receive data on the move by flooding queries within a local cell only. The sink confines the destination area as it moves in order to broadcast its location information within the destination area only rather than to the entire network so as to reduce energy consumption of location updating message. Similarly, in [8], Wang et al. proposed Adaptive Local Update-based Routing Protocol (ALURP), which uses local flooding method to effectively update the location information of a mobile sink. However, there is substantial overhead for sink's location updating, especially when the sink moves at high speed. Shin and Kim [9] proposed a milestone-based predictive routing protocol that improves energy efficiency and prolongs the lifetime of networks. By introducing milestone node, the estimated sink's future location information is spread towards the nodes located in the vicinity of the recent trail of the sink by multihop relay by milestone node. The neighbors of these relay nodes can update their own "routing information" by overhearing, as a result, all local nodes can acquire the latest location information of the mobile sink. This protocol improves energy consumption and data packet delivery ratios. However, it still needs substantial overhead when sinks change their moving direction frequently. Shi et al. [20] proposed an efficient data-driven routing protocol with mobile sinks (DDRP). In order to reduce the protocol overhead for route discovery and maintenance caused by sink mobility while keeping high packet delivery, DDRP integrates data-driven routing and random walk routing in its implementation. Exploiting the broadcast feature of wireless medium, nodes overhear the data packets transmitted by their neighbors to learn fresh route information towards the sink. When no route to the mobile sink is known, random walk routing is adopted for data packet forwarding. DDRP can achieve lower protocol overhead and longer network lifetime. Fodor and Vidács [21] reduced communication overheads by proposing a restricted flooding method. Routes are updated only when topology changes. In [22], the authors utilize a logical coordinate system to infer distances and establish data reporting routing by greedily selecting the shortest path to the destination reference. It effectively reduces energy consumption. However, when changing its location, the mobile sink still needs to reestablish logical coordinate system.

2.3. Reduction of Time Delay of Data Reporting. The introduction of mobile sink may cause serious time delay problem; therefore, many researchers seek solutions to this kind of problems under time-sensitive application scenarios. Liang et al. [14] studied the network lifetime maximization problem for time-sensitive data gathering using a mobile sink with several constraints, such as the total travel distance and maximum distance between the sink's two sojourning locations. They presented a mixed integer linear programming solution to this multiple-constrained problem and proposed a heuristic solution. Xing et al. [6] proposed a rendezvous-based approach in which a subset of nodes serve as the rendezvous points (RPs) that buffer data originated from sources and transfer to MEs when they arrive. Taking data delivery deadline into account, RP-CP and RP-UG were proposed to facilitate reliable data transfers from RPs to MEs under the condition of significant unexpected delays in ME movement and network communication. Aioffi et al. [23] proposed the Minimum Wiener index Spanning Tree (MWST) as a routing topology for multiple base stations. A branch and bound algorithm for small-scale WSNs and a simulated annealing algorithm for large-scale WSNs are designed alternatively. The energy efficiency and packet delay attributes performance better than that of traditional minimum spanning tree.

In this paper, a location predictive and time adaptive data gathering scheme with mobile sink is proposed for wireless sensor networks. The trajectory of mobile sink can be a predefined circle, rectangle, or other geometric shapes depending on the deployed area, and the moving velocity of the sink is a constant. These two conditions make the mobile sink location predictable, which reduces the energy overhead for broadcasting location update messages of the mobile sink while maintaining low data transmission delay. When reporting or forwarding data to the mobile sink, sensors calculate the location of the mobile sink based on a loose time synchronization mechanism among sensor nodes and the mobile sink. Different from [9], in LPTA, the mobile sink needs no location updating message to inform sensor nodes of its latest location, which saves a lot of control overhead. The sink collects data from sensors only when it is dwelling at sojourn points. The sink dwelling time at sojourn points is dynamically adjustable, but unlike depending on local nodes density in [16], time adjustment method in our scheme is

based on the number of historical data generated in each area, which is more applicable to real environments.

3. Network Model and Problem Statement

The network model is shown in Figure 1. N sensor nodes are deployed randomly in the network and one mobile sink gathers data from the whole network. All sensor nodes are quasi-stationary and location-aware (i.e., equipped with GPS-capable antennae). The mobile sink is not constrained by energy and can move at constant velocity. The whole network area is a $W \times L$ rectangle, and the mobile sink moves along a predefined trajectory. A predefined fixed trajectory is the base of location prediction, and many researchers have investigated the performance of different trajectory for data gathering protocols [10, 15, 24]. We use a rectangle of $w \times l$ and a circle with radius R as the examples of trajectories in our network as shown in Figures 1 and 2, respectively. The prediction of the location of mobile sink based on these two trajectories is simple and effective.

As described in Section 1, the frequentness and the number of the events of interest may be different in different part of the deployment area. And the network is deployed into two-dimensional Cartesian coordinate system. To simplify the related formulas and expressions, focus on the core algorithm of LPTA's dwelling time adjustment, and make it easily understood; we divide the network into four quadrants and its center is denoted as origin point O. The mobile sink turns off its radio transceiver while moving between two sojourn points and collects data from sensors only when it is dwelling at sojourn points. The number of sojourn points n is a multiple of four, and they are evenly distributed on the trajectory. Because the dwelling time of the mobile sink is adjusted based on the data generation portion of different region rather than different sojourn point, the number of sojourn points has no effect on the overhead of control message, and is only used for mobile gathering data in a predictive discrete manner. An anticlockwise rule is used to determine which quadrant a sojourn point belongs to; for example, point A belongs to quadrant I as shown in Figure 1. In Section 6, the simulation results under different sink's moving trajectory show that when sink moves along the rectangle trajectory, there is a better performance. Therefore, the location formulas of the mobile sink in our paper will be given based on a rectangle trajectory and other conditions such as circle trajectory can be acquired in a similar way.

Sensor nodes are able to communicate with the mobile sink by multihop relay. The nodes that can communicate directly with the sink within their communication radius r are one-hop neighbors of the mobile sink. For the sake of convenience, the main symbols used in this paper are listed in Notations Section.

There are two core problems to be solved in this paper. The first one is the moving strategy of the mobile sink. It includes the behavior of the mobile sink when the sink moves along the predefined trajectory and the dwelling time adjusting method in each quadrant. As illustrated in Figure 1, there may be obstacles in the monitoring area, for example, pools or swamps; we assume nodes deployed in these area are disabled to monitor the surroundings, which will cause the difference of data amount generating from each quadrant. Therefore, it is necessary to adjust the dwelling time of mobile sink to reside longer in the quadrant which generates more data packets. The mobile sink can be a quadcopter or an aircraft which is not influenced by these obstacles when moving along the trajectory. The second one is the data routing method for sensor nodes reporting data towards the mobile sink. Multihop relay communication method among nodes is adopted, as shown in Figure 1. Source node uploads data

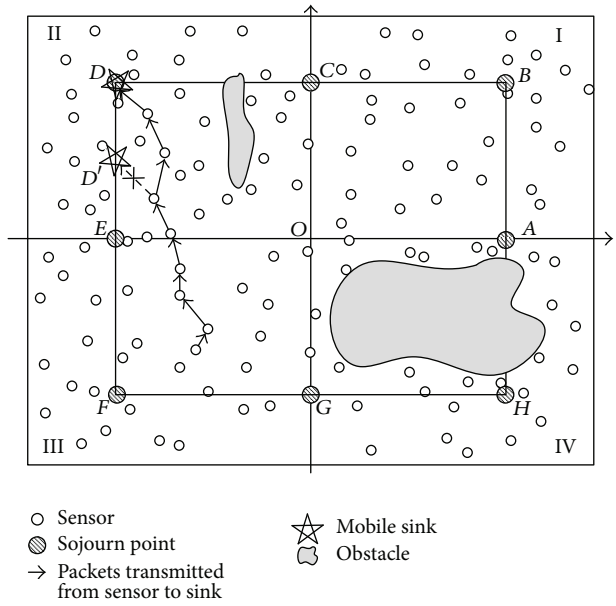


Figure 1: Network model of rectangle trajectory.

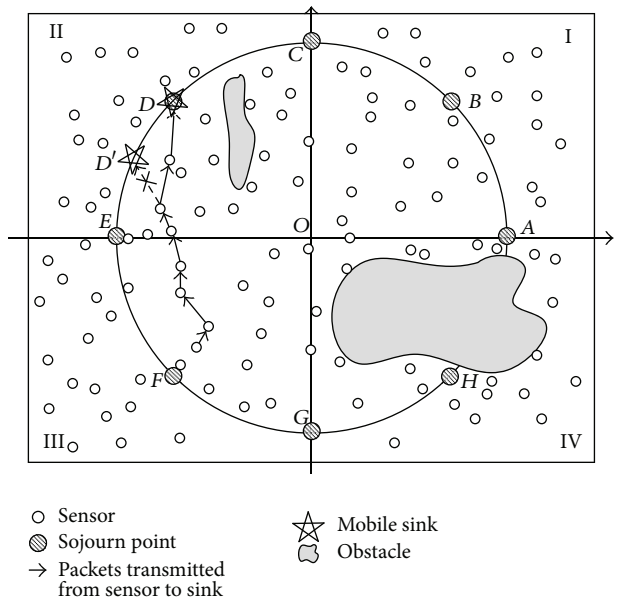


Figure 2: Network model of circle trajectory.

packets towards the mobile sink by multihop relay. Both of them will be stated in detail in the following section.

4. Sink Moving Strategy

In this section, we describe the data gathering process of mobile sink and introduce the dwelling time adaptive criteria. The data gathering process is periodically carried out by the mobile sink. During the periodical data gathering process, the mobile sink circles along the trajectory, stops at a sojourn point to collect sensed data, and then moves to the next sojourn point. To simplify our data gathering scheme, we divide the whole data gathering process into several time intervals, each of which is corresponding to a data gathering period (DGP). The mobile sink circles along the trajectory over and over in one data gathering period, and we further divide one data gathering period into multiple data gathering circles (DGPs). The detailed definition of data gathering period and data gathering circle are as follows.

(i) *Data Gathering Period (DGP)*. A DGP is defined as the process from the beginning of the sink entering the network to its leaving the network. During one DGP, the mobile sink carries out the data gathering algorithm which includes loose time synchronization phase, time adaptive data collection phase, and DGP ending declaration phase. One DGP is composed of multiple DGCs.

(ii) *Data Gathering Circle (DGC)*. A DGC refers to the process of the sink moving along the trajectory, and backing to the initial point. For example, the sink starts from sojourn point A, moves along the trajectory, passes through sojourn points from B to H, and comes back to point A again, as shown in Figure 1. This process is a DGC.

The number of DGCs contained in one DGP depends on the application requirements and is limited by the available energy carried by the mobile sink. The higher the value, the better it is. Because at the beginning and the ending of one DGP, the mobile sink needs to broadcast a *HELLO* message and a *BYE* message, respectively, which will consume the energy of nodes. Considering realistic situation, it is set 10 in our simulation. Additionally, as the relevant formulas of sink's location is related to sink's moving trajectory, we only give the formulas hereinafter according to the rectangle moving trajectory; similar methods can be used to acquire the expression of other shapes.

4.1. Loose Time Synchronization. The mobile sink as well as every node in the network owns its own clock. At the beginning of one DGP, the mobile sink broadcasts a time synchronization message, *HELLO*, to achieve loose time synchronization among the mobile sink and all nodes in the network. Based on this synchronization, every node in the network can calculate the location of the mobile sink according to its local time information when uploading data packets to the mobile sink.

The loose time synchronization phase is the first phase during one DGP. When entering into the network, the mobile sink broadcasts a *HELLO* message to the whole network. The *HELLO* message consists of the starting location information $S(x, y)$, current time t_0, the moving velocity V of the mobile sink, the number of sojourn points n on the trajectory, the dwelling time at each sojourn point in quadrant i ($i \in \{1, 2, 3, 4\}$) during the first DGC $T_s(i, 1)$, and the width and length of rectangle trajectory w and l. Every node changes its clock to t_0 when it receives the *HELLO* message for the first time and then retransmits this message to its neighbors. Note that the parameters $T_s(i, 1)$ in the *HELLO* message are equal to each other, that is $T_s(1, 1) = T_s(2, 1) = T_s(3, 1) = T_s(4, 1) = T_s$.

After the network achieved loose time synchronization, the mobile sink starts to collect data packets from the network. The time for loose time synchronization can be ignored since it is quite small compared with the time for one DGP.

4.2. Time Adaptive Data Collection. During the time adaptive data collection phase, the dwelling time is adjusted dynamically. In some application scenarios, there may exist obstacles in the network area; hence the data generated from each quadrant can be different greatly. According to the variation degree of $P_{\text{data}}(i, k-1)$ and $P_{\text{data}}(i, k)$, the dwelling time $T_s(i, k+1)$ in the $(k+1)$th DGC at sojourn points is adjusted dynamically. In this way, the energy consumption of entire network can be further balanced.

As the data packets in each quadrant are generated in a random manner and transmitted by multihop relay to the mobile sink, the routing path for these packets generated in quadrant i will be longer than those in quadrant j ($i \neq j$) where the mobile sink locates; therefore, the former will consume much more energy than the latter. To reduce the energy consumption caused by long distance data packets routing, after finishing each DGC, the mobile sink statistics the number of packets received from each quadrant and then calculates the proportion of these packets to the entire network data packets P_{data1}, P_{data2}, P_{data3}, and P_{data4}, accordingly. Depending on these proportions, the dwelling time of the mobile sink at sojourn points in each quadrant is adjusted dynamically, which makes the energy consumption in the network more balanced and the network lifetime extended.

The principle of adjusting the dwelling time $T_s(i, k+1)$ in the $(k+1)$th DGC is described in detail as follows.

In order to distinguish the quadrant in which a data packet is generated, source nodes append a 2-bit quadrant information to the head of the data packet before sending it to its next hop. The quadrant information can be calculated based on the sensor node location $\text{loc}(x_i, y_i)$ relative to the origin point O's location information. Note that only the source nodes need to add their own quadrant information to the head of the data packets.

During one DGP, the mobile sink calculates the $(k+1)$th DGC's dwelling time in each quadrant according to the proportions P_{data1}, P_{data2}, P_{data3}, and P_{data4} in the $(k-1)$th DGC and the proportions in the kth DGC. When the value

$P_{\text{change}}(k, k-1)$ is greater than the threshold value T_h, the dwelling time in corresponding quadrant will be adjusted as $T_s(i, k+1) = 4P_{\text{data}i}T_s$. The value of $P_{\text{change}}(k, k-1)$ is calculated according to the following formula:

$$P_{\text{change}}(k, k-1) = \sqrt{\sum_{i=1}^{4}\left(P_{\text{data}}(i,k) - P_{\text{data}}(i,k-1)\right)^2}. \quad (1)$$

$P_{\text{change}}(k, k-1)$ represents the variation degree of data generation proportion in different quadrant between two adjacent DGC. When this value is greater than the threshold T_h, it means that the quantity of data packets generated in each quadrant has changed significantly, and the dwelling time needs to be adjusted. Under this circumstance, the mobile sink will broadcast a *UPDATE* message to all nodes in the network, which includes the adjusted dwelling time in each quadrant $T_s(i, k+1)$. Otherwise, there is no necessity to modify the dwelling time, and the mobile sink maintains the dwelling time in each quadrant the same as the previous DGC.

4.3. DGP Ending Declaration. DPG ending declaration phase is the last phase in one DGP. At the beginning of the last DGC of one DGP, the mobile sink broadcasts a *BYE* message to all nodes in the network. The broadcasting of *BYE* message means that the current DGP is coming to an end, and the mobile sink will stop gathering data and leave the network. The *BYE* message consists of the time T_{bl}, which is the time interval between current time and the mobile sink finishing current DGP. Instead of routing the data to the mobile sink, when receiving *BYE* messages, nodes will buffer the data sensed from surroundings after time T_{bl}:

$$T_{\text{bl}} = nT_s + \frac{2(w+l)}{V} - 2T_{\text{syn}}. \quad (2)$$

T_{syn} is the time needed for the network to achieve loose time synchronization and, as to T_{syn}, there is

$$T_{\text{syn}} \ll nT_s + \frac{2(w+l)}{V}. \quad (3)$$

Therefore T_{syn} has little effect on T_{bl} and can be ignored in practical applications.

5. Data Reporting Process

In the network, source nodes transmit data packets to the mobile sink by multiple hops. The principle of selecting next hop is to make the path between a source node and the mobile sink approximately shortest. When nodes have data to report or relay, they need to calculate the mobile sinks current location based on their own clocks, which have been loosely synchronized to the mobile sink, and then choose one of its neighbors as the next hop. Using other existing routing protocols for data routing is also feasible. Our scheme focuses on sink's location prediction and its dwelling time adjustment. If there exists the concave obstacle which may result in data reporting failed, obstacle avoidance routing protocols can be adopted to relay the data packets to the mobile sink.

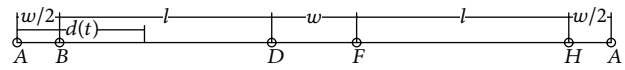


Figure 3: Mapping model of rectangle trajectory.

The time step T_{step} is the time interval for moving between two adjacent sojourn points. It is calculated by the following formula:

$$T_{\text{step}} = \frac{2(w+l)}{nV}. \quad (4)$$

During the loose time synchronization phase, the mobile sink broadcasts a *HELLO* message to achieve loose time synchronization among all the nodes in the network. The parameter $T_s(i, 1)$ in the *HELLO* message is equal to each other; that is, $T_s(1,1) = T_s(2,1) = T_s(3,1) = T_s(4,1) = T_s$. T_s is a constant value and keeps unchanging during a DGP.

To determine the location of the mobile sink at time t, in this paper, we have the moving trajectory of the mobile sink map to a line model. In the model, the starting location of a DGP is chosen as the reference location, for example, point A in Figure 3. Corresponding to the rectangle trajectory illustrated in Figure 1, the line model is shown in Figure 3, in which the $d(t)$ is the distance from the reference location A to the current location of mobile sink in the present DGC.

The moving time of the mobile sink in current DGC is denoted as T_p, which is calculated by the following formula:

$$T_p = \left\lfloor \frac{t - t_0}{n\left(T_{\text{step}} + T_s\right)} \right\rfloor. \quad (5)$$

Based on loose time synchronization, the location of the mobile sink can be calculated by the following formulas:

(1) If

$$0 \le T_p < \frac{nT_s(1,k)}{4} + \frac{w+l}{2V}, \quad (6)$$

then

$$\text{Loc}(t) = \begin{cases} \left(-\dfrac{l}{2}, d(t)\right), & d(t) < \dfrac{w}{2}, \\ \left(-\dfrac{l}{2} + \left(d(t) - \dfrac{w}{2}\right), \dfrac{w}{2}\right), & \text{others} \end{cases} \quad (7)$$

$d(t)$ can be calculated as below:

$$T_1 = \frac{\left(T_p - \left\lfloor T_p / \left(T_{\text{step}} + T_s(1,k)\right)\right\rfloor \left(T_{\text{step}} + T_s(1,k)\right)\right)}{\left(T_s(1,k)\right)} \quad (8)$$

and if $\lfloor T_1 \rfloor = 0$, then

$$d(t) = V \left\lfloor \frac{T_p}{T_{\text{step}} + T_s(1,k)} \right\rfloor T_{\text{step}} \quad (9)$$

else

$$d(t) = V\left(T_p - \left\lceil \frac{T_p}{T_{\text{step}} + T_s(1,k)} \right\rceil T_s(1,k)\right). \quad (10)$$

(2) If

$$\frac{nT_s(1,k)}{4} + \frac{w+l}{2V} \le T_p < \frac{n\left(T_s(1,k) + T_s(2,k)\right)}{4} + \frac{2(w+l)}{2V}, \quad (11)$$

then

$$\text{Loc}(t) = \begin{cases} \left(-\frac{l}{2} + \left(d(t) - \frac{w}{2}\right), \frac{w}{2}\right), & d(t) < \frac{w}{2} + l \\ \left(\frac{l}{2}, \frac{w}{2} - \left(d(t) - l - \frac{w}{2}\right)\right), & \text{others} \end{cases} \quad (12)$$

$d(t)$ can be calculated as below:

$$T_2 = \left(T_p - \frac{nT_s(1,k)}{4} - \frac{w+l}{2V} - \left\lfloor \frac{T_p - nT_s(1,k)/4 - (w+l)/2V}{T_{\text{step}} + T_s(2,k)} \right\rfloor \times \left(T_{\text{step}} + T_s(2,k)\right)\right) \times \left(T_s(2,k)\right)^{-1} \quad (13)$$

and if $\lfloor T_2 \rfloor = 0$, then

$$d(t) = V\left(\frac{w+l}{2V} + \left\lfloor \frac{T_p - nT_s(1,k)/4 - (w+l)/2V}{T_{\text{step}} + T_s(2,k)} \right\rfloor \times T_{\text{step}}\right) \quad (14)$$

else

$$d(t) = V\left(T_p - \frac{nT_s(1,k)}{4} - \left\lceil \frac{T_p - nT_s(1,k)/4 - (w+l)/2V}{T_{\text{step}} + T_s(2,k)} \right\rceil T_s(2,k)\right). \quad (15)$$

(3) If

$$\frac{n\left(T_s(1,k) + T_s(2,k)\right)}{4} + \frac{2(w+l)}{2V} \le T_p < \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} + \frac{3(w+l)}{2V}, \quad (16)$$

then

$$\text{Loc}(t) = \begin{cases} \left(\frac{l}{2}, -d(t) + l + w\right), & d(t) < \frac{3w}{2} + l \\ \left(\frac{l}{2} - \left(d(t) - l - \frac{3w}{2}\right), -\frac{w}{2}\right), & \text{others} \end{cases} \quad (17)$$

$d(t)$ can be calculated as below:

$$T_3 = \left(T_p - \frac{n\left(T_s(1,k) + T_s(2,k)\right)}{4} - \frac{2(w+l)}{2V} - \left\lfloor \frac{T_p - n\left(T_s(1,k) + T_s(2,k)\right)/4 - 2(w+l)/2V}{T_{\text{step}} + T_s(3,k)} \right\rfloor \times \left(T_{\text{step}} + T_s(3,k)\right)\right) \times \left(T_s(3,k)\right)^{-1} \quad (18)$$

and if $\lfloor T_3 \rfloor = 0$, then

$$d(t) = V\left(\frac{2(w+l)}{2V} + \left\lfloor \frac{T_p - n\left(T_s(1,k) + T_s(2,k)\right)/4 - 2(w+l)/2V}{T_{\text{step}} + T_s(3,k)} \right\rfloor \times T_{\text{step}}\right) \quad (19)$$

else

$$d(t) = V\left(T_p - \frac{n\left(T_s(1,k) + T_s(2,k)\right)}{4} - \left\lceil \frac{T_p - n\left(T_s(1,k) + T_s(2,k)\right)/4 - 2(w+l)/2V}{T_{\text{step}} + T_s(3,k)} \right\rceil \times T_s(3,k)\right). \quad (20)$$

(4) If

$$\frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} + \frac{3(w+l)}{2V} \le T_p < \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k) + T_s(4,k)\right)}{4} + \frac{4(w+l)}{2V}, \quad (21)$$

then

$$\mathrm{Loc}(t) = \begin{cases} \left(\frac{l}{2} - \left(d(t) - l - \frac{3w}{2}\right), -\frac{w}{2}\right), \\ \qquad d(t) < \frac{3w}{2} + 2l \\ \left(-\frac{l}{2}, -\frac{w}{2} + d(t) - 2l - \frac{3w}{2}\right), \\ \qquad \text{others} \end{cases} \tag{22}$$

$d(t)$ can be calculated as below:

$$\begin{aligned} T_4 = \Bigg(T_p - \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} - \frac{3(w+l)}{2V} \\ - \Bigg\lfloor \Bigg(T_p - \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} \\ - \frac{3(w+l)}{2V}\Bigg) \times \left(T_{\text{step}} + T_s(4,k)\right)^{-1} \Bigg\rfloor \\ \times \left(T_{\text{step}} + T_s(4,k)\right)\Bigg) \\ \times \left(T_s(4,k)\right)^{-1} \end{aligned} \tag{23}$$

and if $\lfloor T_4 \rfloor = 0$, then

$$\begin{aligned} d(t) = V\Bigg(\frac{3(w+l)}{2V} \\ + \Bigg\lfloor \Bigg(T_p - \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} \\ - \frac{3(w+l)}{2V}\Bigg) \times \left(T_{\text{step}} + T_s(4,k)\right)^{-1} \Bigg\rfloor \\ \times T_{\text{step}}\Bigg) \end{aligned} \tag{24}$$

else

$$\begin{aligned} d(t) = V\Bigg(T_p - \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} \\ - \Bigg\lceil \Bigg(T_p - \frac{n\left(T_s(1,k) + T_s(2,k) + T_s(3,k)\right)}{4} \\ - \frac{3(w+l)}{2V}\Bigg) \times \left(T_{\text{step}} + T_s(4,k)\right)^{-1} \Bigg\rceil \\ \times T_s(4,k)\Bigg). \end{aligned} \tag{25}$$

We define $\lfloor x \rfloor$ as the largest integer no more than x, $\lceil x \rceil$ as the smallest integer no less than x, and $[x1/x2]$ as the reminder of $x1$ divided by $x2$. To keep the formula as simple as possible, we require the starting point of data gathering to be on the intersection of x axis or y axis. Without loss of generality, we choose the location A as shown in Figure 1 as the starting point and deduce a series of formulas above.

As described in Section 3, the moving trajectory can be a rectangle or circle. When the sink's moving trajectory is a circle, the time-location formulas can be acquired in a similar way. The difference is that the moving trajectory of the mobile sink is mapped into a polar coordinate system, because of calculation convenience. Figure 4 is a model of polar system, assuming the mobile sink starts its data gathering process from point A at time t_0 and reaches point B at time t, the arc length of $\overset{\frown}{AB}$ is $R\theta$ $(0 \le \theta < 2\pi)$. When the sink moves at speed V in the network, there is $V(t - t_0) = R\theta$. The corresponding polar coordinate of point B is (R, θ).

Similar to rectangle trajectory, based on loose time synchronization, the location of the mobile sink can be acquired accurately when the trajectory is a circle. For example, when the mobile sink locates at the first quadrant, then T_p meets $0 \le T_p < (n/4)T_s(1,k) + 2\pi R/4V$, and the polar angle of the sink is calculated by the following formulas:

If

$$\left\lfloor \frac{T_p - \left\lceil T_p / \left(T_{\text{step}} + T_s(1,k)\right)\right\rceil \left(T_{\text{step}} + T_s(1,k)\right)}{T_s(1,k)} \right\rfloor = 0, \tag{26}$$

then

$$\theta = \frac{2\pi}{n}\left\lfloor \frac{T_p}{T_{\text{step}} + T_s(1,k)} \right\rfloor, \tag{27}$$

else

$$\theta = \frac{V}{R}\left(T_p - \left\lceil \frac{T_p}{T_{\text{step}} + T_s(1,k)} \right\rceil T_s(1,k)\right). \tag{28}$$

When the sink locates at other quadrants, the corresponding location information can be obtained in a similar manner.

According to the calculated location information, source nodes upload their sensed data to the mobile sink by multihop communication. When events occur in the monitoring area, the sensors outside the communication range of the mobile sink route data packets to their next hop directly. It is unnecessary to judge the current state of the mobile sink, that is, moving between sojourn points or gathering data packets at a sojourn point. Only the neighbor nodes of the mobile sink need to judge the state of the mobile sink. If the mobile sink is moving between sojourn points, the neighbor nodes have to wait for a period of time T_{wl} until the former arrives at its next sojourn point. Otherwise, they transmit the data packet to the mobile sink directly. For instance, as shown in Figure 1, we assume the current location of the mobile sink is point D; if the events occur in quadrant III, then data packets can be routed along the shortest routing path to point D. When the data packets reach the neighbor node of the mobile sink, it will judge the state of mobile sink according to its local time clock. If the time of the node meets $t_0 + k(T_{\text{step}} + T_s(i,k)) < t < t_0 + k(T_{\text{step}} + T_s(i,k)) + T_s(i,k)$, which means the mobile sink is

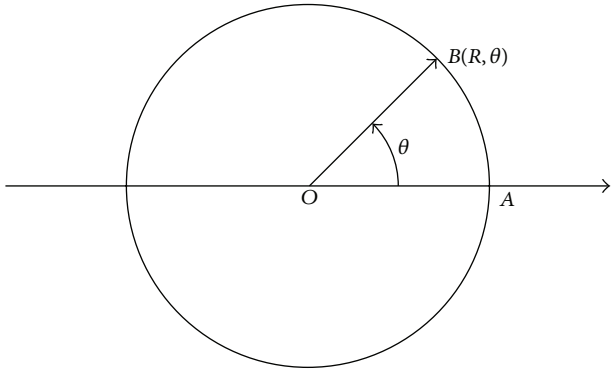


Figure 4: The model of polar system.

still gathering data at the sojourn point D, then this one-hop neighbor node of the sink transmits the data packets directly to the mobile sink; otherwise, for example, the sink is now located at D', it needs to wait time T_{wl} and then transmit the data packet to the mobile sink. The time T_{wl} is calculated by the following formula:

$$T_{wl} = t_0 + (k+1)\left(T_{\text{step}} + T_s(i,k)\right) - t. \tag{29}$$

During routing data packets to the mobile sink, hop-by-hop acknowledgement mechanism is applied to ensure the data transmission rate, that is, if the receiver *Node2* gets the data packets from sender *Node1*, it will reply with an *ACK* message to *Node1*. If *Node1* does not receive the *ACK* message from its next hop node *Node2* within time T, *Node1* considers that the packet transmission is failed and will cache the data packets and wait for a random time and then retransmit the packets to its next hop again. We assume that T is equal to the propagation time of a packet between two farthest nodes of the network.

6. Performance Evaluation

In this section, we evaluate the performance of our scheme through extensive simulations. In addition to the proposed scheme, we implemented the *Adaptive* [16] for comparison. The reason is that *Adaptive* is also a discrete data gathering protocol. Besides, the mobile sink in *Adaptive* dwells at sojourn points on the predefined trajectory for data collection, and the dwelling time is dynamically adjusted. Two performance metrics, energy consumption and data delivery latency, are investigated. Energy consumption is the average energy that is consumed by nodes during one DGP. And data delivery latency is the time interval from a message creation to the mobile sink receiving it.

6.1. Simulation Environment. We implement our proposed scheme in MatLab. In our simulation, the deployment area has a 500 m × 500 m square sensing field and sensor nodes are randomly deployed in the network, that is, area length L equals its width W. The communication range of the nodes and the mobile sink is set to 60 m. The mobile sink moves along the predefined trajectory for 10 circles in every DGP, and in every DGC, 5% of sensor nodes act as source nodes, which send messages toward the mobile sink continually when the sink is dwelling at sojourn points.

Different simulation environments with varying sink moving trajectory, number of nodes N, and mobile sink speed V are studied. We set the trajectory as circle and rectangle, the length of side or radius of trajectory is set as $L/4$, $L/2$, $3L/4$, and L. And we varied N from 800 to 1200, V from 4 m/s to 20 m/s. Several groups of simulation experiments are carried out. The threshold of adjusting dwelling time T_h is set as 0, 0.25, 0.5, 0.75, and 1. $T_h = 0$ means the dwelling time of the mobile sink needs to be changed if the proportion of packets amount quantity generated from every quadrant is not exactly the same as previous DGC, while $T_h = 1$ means the dwelling time keeps unchanging during one DGP.

6.2. Simulation Results with Varying Sink Moving Trajectory. Now we discuss the influence of sink moving trajectory on network performance. We compare two scenarios of the trajectory; they are rectangle and circle, and the mobile sink moves along the predefined trajectory for one DGP. The simulation results are shown in Figures 5, 6, 7, and 8.

Comparing Figures 5 and 7 with Figures 6 and 8, respectively, it is noticed that the energy consumption of nodes is lower when the moving trajectory is rectangle than that of circle when the deployment area is rectangle. The sharp of sink's moving trajectory has influence on energy consumption of nodes even if the path is predefined. In the following simulation, the moving trajectory of the mobile sink is rectangular which is similar to network deployment area.

As shown in Figures 7 and 8, the energy consumption is the lowest while mobile sink moves along the track with $L/4$ and $3L/4$ length of side or radius. It is different from the theory proposed in [15] that peripheral movement is the best strategy, because the ideal load-balanced routing is hard to satisfy. Under the condition of a certain trajectory, there is an outstanding performance when T_h is less than 1; this is because the dynamic adjustment of dwelling time is beneficial to the performance of network. Besides, when $T_h = 0.75$, as shown in Figure 5, the energy consumption of control message is very low. This is because there is an appropriate tradeoff between the control overhead and the balanced energy consumption among different quadrants when $T_h = 0.75$. In contrast, the dwelling time adjustment frequency is too high when $T_h = 0$, resulting in much energy overhead. When $T_h = 1$, the energy consumption of control message equals 0, which means there is no dwelling time adjustment and the energy consumption among nodes is not well balanced.

6.3. Simulation Results with Varying Number of Sensor Nodes. Now we discuss the performance of our scheme by setting the number of sensor nodes N varying from 800 to 1000 when the moving trajectory is rectangle. The simulation results are shown in Figures 9 and 10.

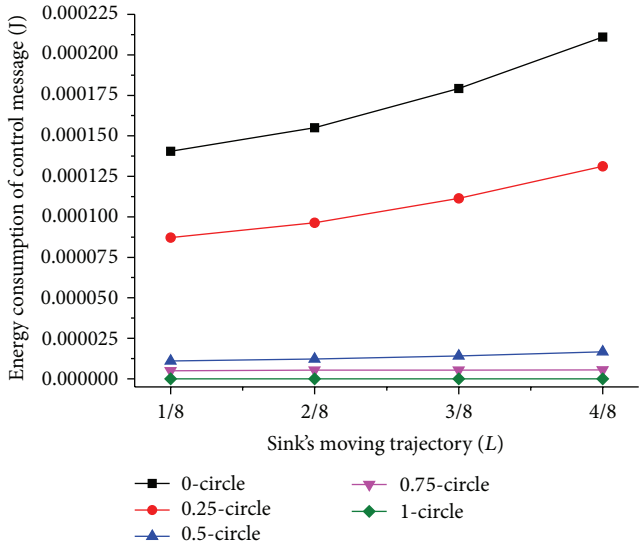


Figure 5: Energy consumption of control message with circle trajectory.

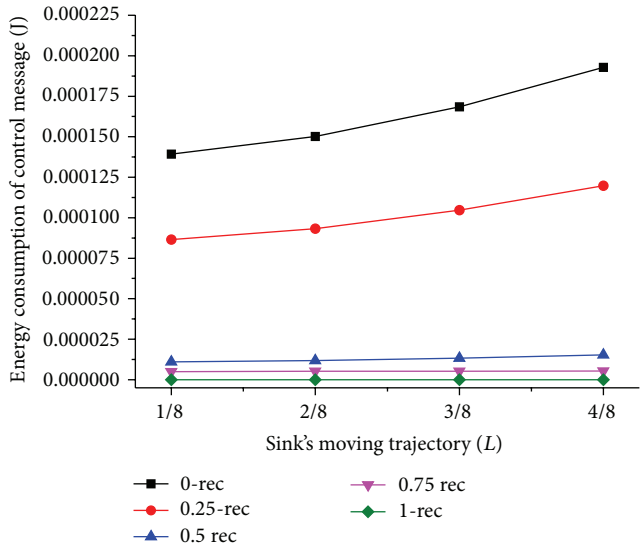


Figure 6: Energy consumption of control message with rectangle trajectory.

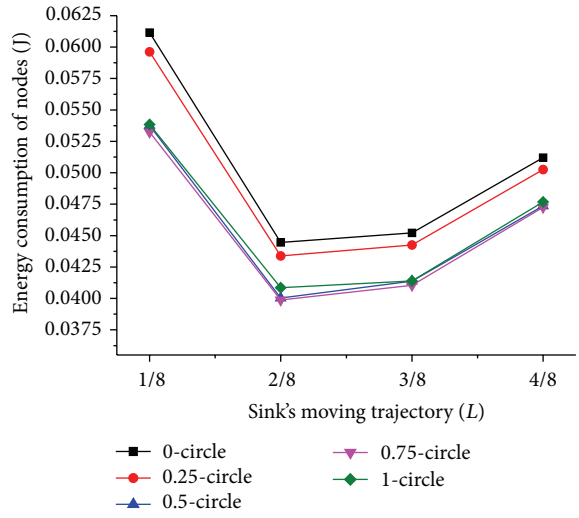


Figure 7: Energy consumption of nodes with circle trajectory.

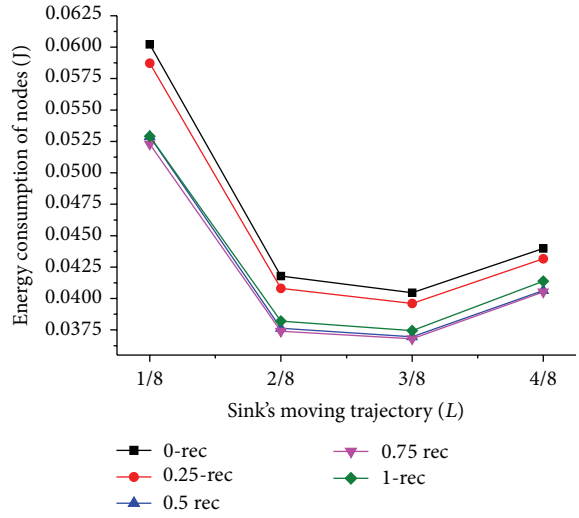


Figure 8: Energy consumption of nodes with rectangle trajectory.

As illustrated in Figure 10, the energy consumption of nodes decreases first with increased number of nodes and then increases when N is more than 1000. It is because, with the increase of N, the number of selectable next hop neighbors increases, as a result, the hop distance and the routing path between the mobile sink and source nodes are improved, which results in less energy consumption for relaying the same quantity of data packets. When N is more than 1000, the increase of the number of source nodes causes more packets generating in the network, which results in more energy consumption of nodes when N increases. The performance is outstanding compared to the others when T_h is 0.75, and the reason is the same as explained in Section 6.2.

6.4. Simulation Results with Varying Sink Speed. Now we evaluate the network performance when sink's moving speed V varies from 4 m/s to 20 m/s. The results are shown in Figures 11 and 12.

It is noticed that as the mobile sink speed goes up, the energy consumption decreases. This is because with the increase of sink speed, the time the mobile sink spending for moving between two adjacent sojourn points decreases, and the mobile sink can receive the data packets timely from sensor nodes. This results in less energy consumed.

6.5. Simulation Results of Data Transmission Delay and Energy Consumption. We simulated our proposed scheme, as well as the adaptive algorithm and constant algorithm described in the adaptive sink mobility scheme proposed by Kinalis et al. [16], to evaluate the performance of data delivery latency and energy consumption by varying sink's moving speed.

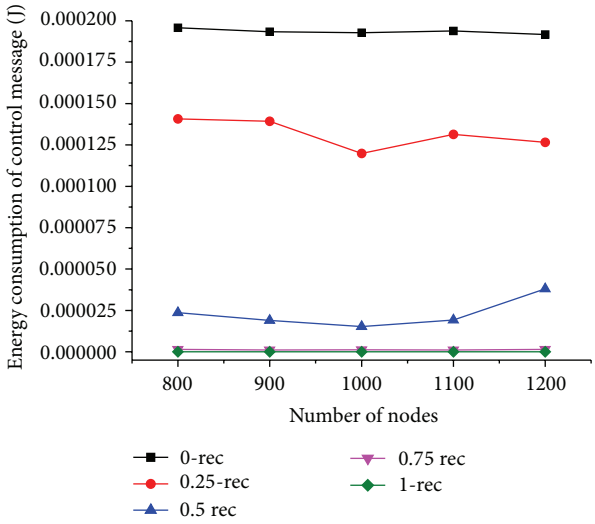


Figure 9: Energy consumption of control message under varying number of nodes.

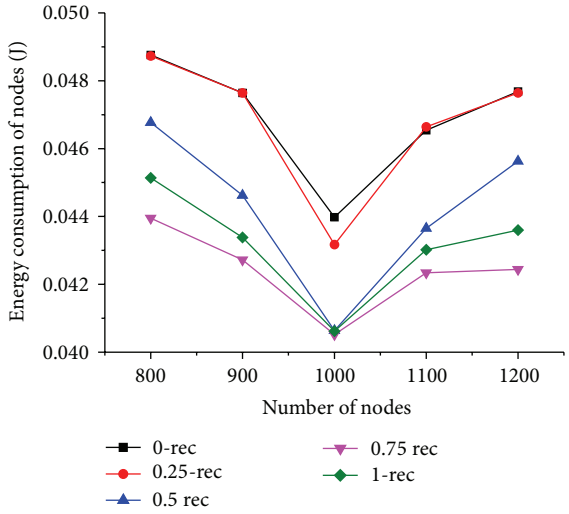


Figure 10: Energy consumption of nodes under varying number of nodes.

The results are shown in Figure 13. Our LPTA scheme outperforms the Adaptive scheme in the attribute of latency. The reason is that in LPTA, the latency is mainly caused by the mobile sink turning off its communication model when moving between two adjacent sojourn points. But in adaptive and constant algorithms, the sink has to traverse all vertexes, which results in large time delay. Besides, the increase of speed is beneficial for our LPTA. It is because the time needed decreases for moving the same distance with higher moving speed. Hence, the data transmission delay significantly reduced with the increase of sink speed.

Figure 14 shows the performance of energy consumption with the change in velocity. When the sink moving speed is relatively small, *adaptive* algorithm performance is almost the same as our LPTA. However, with the increasing of sink's moving speed, the energy consumption of *adaptive*

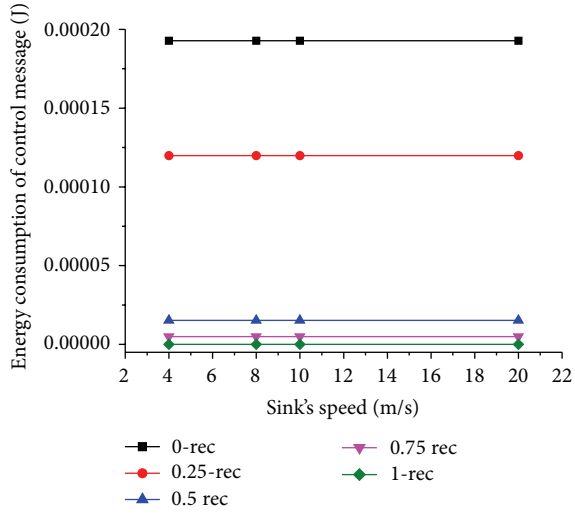


Figure 11: Energy consumption of control message under varying sink speed.

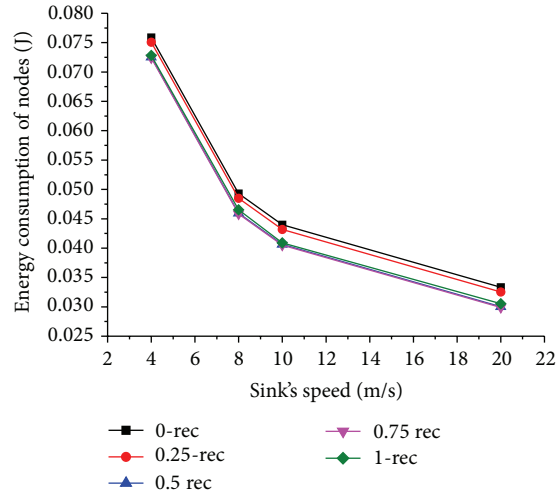


Figure 12: Energy consumption of nodes under varying sink speed.

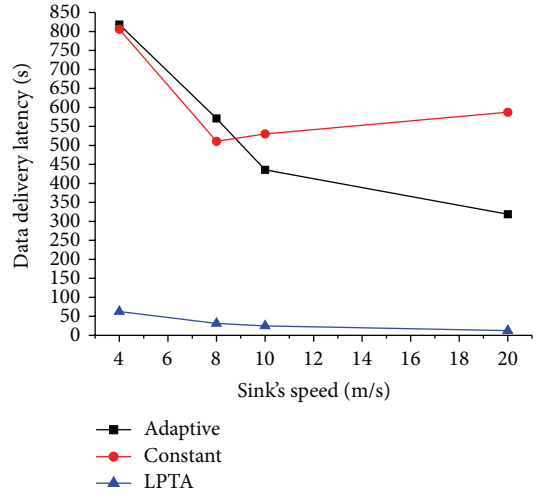


Figure 13: Data transmission latency under varying sink speed.

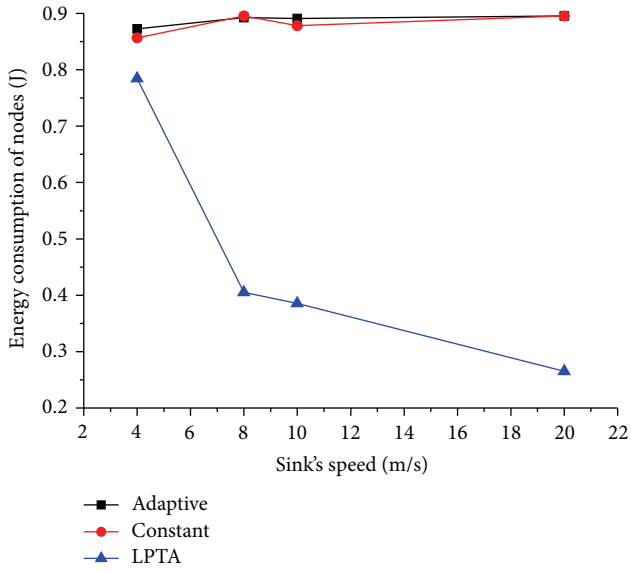


FIGURE 14: Energy consumption under varying sink speed.

and *constant* algorithms are much more than LPTA. It is because with the increasing of sink's speed, the time the mobile sink is spending for moving between two adjacent sojourn points decreases. As a result, the mobile sink can receive the data packets timely from sensor nodes, and the hops and routing path are shortened. Therefore, the energy consumption decreases accordingly.

7. Conclusion

In this paper, we propose a location predictive and time adaptive data gathering scheme with one mobile sink. Based on the loose time synchronization, nodes can calculate the latest location information of the mobile sink. Therefore, source nodes are able to route data packets timely to the mobile sink by multihop relay. As a result, the energy overhead for updating sink's location is largely reduced. Along with the location predictive algorithm, this study also describes a dwelling time adjustment method for the mobile sink to efficiently balance the energy consumption among nodes. Simulation results show that the proposed data collection scheme provides improved performance on time latency and energy consumption compared to *adaptive* algorithm. However, as described in Section I, the environments of the deployment area are often harsh and involve obstacles. When some sojourn points happen to be on huge obstacles, it may have bad effect on data packets uploading to the mobile sink. Under this condition, when the mobile sink dwells on these sojourn points, data packets have to be uploaded for many times and even be lost, which will degrade the energy consumption performance and packet delivery ratio. For future research, we plan to enhance the LPTA scheme by considering huge obstacles avoidance to further improve the performance of data gathering algorithm.

Notations

$T_s(i,k)$: The dwelling time of mobile sink at each sojourn point in quadrant i during the kth DGC in current DGP
T_{syn}: The time needed for the network to achieve time synchronization
$P_{\text{data}}(i,k)$: The proportion of the collected data from quadrant i to the whole network during the kth circle
T_p: The time since the beginning of present DGC
T_{bl}: The time before the mobile sink leaving the network in present DGP
r: The communication radius of nodes and the mobile sink
R: The radius of the circle moving trajectory
w/l: The width/length of the rectangle moving trajectory
n: The number of sojourn points
N: The number of sensor nodes
V: The speed of the mobile sink.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The work is supported by the Science and Technology Pillar Program of Changzhou (Social Development), no. CE20135052. Joel J. P. C. Rodrigues's work has been supported by the Fundamental Research Funds for the Central Universities (Program no. HEUCF140803), by *Instituto de Telecomunicações*, Next Generation Networks and Applications Group (NetGNA), Covilhã Delegation, by Government of Russian Federation, Grant 074-U01, and by National Funding from the FCT—*Fundação para a Ciência e a Tecnologia* through the Pest-OE/EEI/LA0008/2013 Project.

References

[1] G. Han, H. Xu, J. Jiang, L. Shu, T. Hara, and S. Nishio, "Path planning using a mobile anchor node based on trilateration in wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 13, no. 14, pp. 1324–1336, 2011.

[2] C. Zhu, C. Zheng, L. Shu, and G. Han, "A survey on coverage and connectivity issues in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 619–632, 2012.

[3] G. Han, H. Xu, T. Q. Duong, J. Jiang, and T. Hara, "Localization algorithms of wireless sensor networks: a Survey," *Telecommunication Systems*, vol. 52, no. 4, pp. 2419–2436, 2013.

[4] G. Wang, J. Cao, H. Wang, and M. Guo, "Polynomial regression for data gathering in environmental monitoring applications," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 1307–1311, IEEE, Washington, DC, USA, November 2007.

[5] C. Chen, J. Ma, and K. Yu, "Designing energy efficient wireless sensor networks with mobile sinks," in *Proceedings of the ACM Sensys'06 Workshop WSW'06*, pp. 1–9, Boulder, Colo, USA, 2006.

[6] G. Xing, T. Wang, Z. Xie, and W. Jia, "Rendezvous planning in wireless sensor networks with mobile elements," *IEEE Transactions on Mobile Computing*, vol. 7, no. 12, pp. 1430–1443, 2008.

[7] S. Basagni, A. Carosi, E. Melachrinoudis, C. Petrioli, and Z. M. Wang, "Controlled sink mobility for prolonging wireless sensor networks lifetime," *Wireless Networks*, vol. 14, no. 6, pp. 831–858, 2008.

[8] G. Wang, T. Wang, W. Jia, M. Guo, and J. Li, "Adaptive location updates for mobile sinks in wireless sensor networks," *Journal of Supercomputing*, vol. 47, no. 2, pp. 127–145, 2009.

[9] K. Shin and S. Kim, "Predictive routing for mobile sinks in wireless sensor networks: a milestone-based approach," *Journal of Supercomputing*, vol. 62, no. 3, pp. 1519–1536, 2012.

[10] K. Lee, Y. Kim, H. Kim, and S. Han, "A myopic mobile sink migration strategy for maximizing lifetime of wireless sensor networks," *Wireless Networks*, vol. 20, no. 2, pp. 303–318, 2013.

[11] X. Li, A. Nayak, and I. Stojmenovic, "Sink mobility in wireless sensor networks," in *Wireless Sensor and Actuator Networks: Algorithms and Protocols for Scalable Coordination and Data Communication*, pp. 153–184, Wiley, 2010.

[12] J. Sheu, P. K. Sahoo, C. Su, and W. Hu, "Efficient path planning and data gathering protocols for the wireless sensor network," *Computer Communications*, vol. 33, no. 3, pp. 398–408, 2010.

[13] Y. Yang, M. I. Fonoage, and M. Cardei, "Improving network lifetime with mobile wireless sensor networks," *Computer Communications*, vol. 33, no. 4, pp. 409–419, 2010.

[14] W. Liang, J. Luo, and X. Xu, "Network lifetime maximization for time-sensitive data gathering in wireless sensor networks with a mobile sink," *Communications and Mobile Computing*, vol. 13, no. 14, pp. 1263–1280, 2011.

[15] J. Luo and J. Hubaux, "Joint mobility and routing for lifetime elongation in wireless sensor networks," in *Proceedings of the 4th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, vol. 3, pp. 1735–1746, IEEE, March 2005.

[16] A. Kinalis, S. Nikoletseas, D. Patroumpa, and J. Rolim, "Biased sink mobility with adaptive stop times for low latency data collection in sensor networks," *Information Fusion*, vol. 15, pp. 56–63, 2012.

[17] Z. M. Wang, S. Basagni, E. Melachrinoudis, and C. Petrioli, "Exploiting sink mobility for maximizing sensor networks lifetime," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*, p. 287a, 2005.

[18] C. H. Liu, K. F. Ssu, and W. T. Wang, "A moving algorithm for non-uniform deployment in mobile sensor networks," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 4, no. 3, pp. 271–290, 2011.

[19] F. Ye, H. Luo, J. Cheng, S. Lu, and L. Zhang, "A two-tier data dissemination model for large-scale wireless sensor networks," in *Proceedings of the 8th Annual International Conference on Mobile Computing and Networking*, pp. 148–159, Atlanta, Ga, USA, September 2002.

[20] L. Shi, B. Zhang, H. T. Mouftah, and J. Ma, "DDRP: an efficient data-driven routing protocol for wireless sensor networks with mobile sinks," *International Journal of Communication Systems*, vol. 26, no. 10, pp. 1341–1355, 2012.

[21] K. Fodor and A. Vidács, "Efficient routing to mobile sinks in wireless sensor networks," in *Proceedings of the 3rd International Conference on Wireless Internet (WICON '07)*, pp. 1–7, 2007.

[22] X. Liu, H. Zhao, X. Yang, and X. Li, "SinkTrail: a proactive data reporting protocol for wireless sensor networks," *IEEE Transactions on Computers*, vol. 62, no. 1, pp. 151–162, 2013.

[23] W. M. Aioffi, C. A. Valle, G. R. Mateus, and A. S. da Cunha, "Balancing message delivery latency and network lifetime through an integrated model for clustering and routing in wireless sensor networks," *Computer Networks*, vol. 55, no. 13, pp. 2803–2820, 2011.

[24] D. P. Liu, K. Zhang, and J. Ding, "Energy-efficient transmission scheme for mobile data gathering in wireless sensor networks," *China Communications*, vol. 10, no. 3, pp. 114–123, 2013.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 646953, 13 pages
http://dx.doi.org/10.1155/2014/646953

Research Article

IoT-Based Smart Garbage System for Efficient Food Waste Management

Insung Hong, Sunghoi Park, Beomseok Lee, Jaekeun Lee, Daebeom Jeong, and Sehyun Park

School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 151-756, Republic of Korea

Correspondence should be addressed to Sehyun Park; shpark@cau.ac.kr

Received 11 April 2014; Accepted 28 May 2014; Published 28 August 2014

Academic Editor: Zhongmei Zhou

Owing to a paradigm shift toward Internet of Things (IoT), researches into IoT services have been conducted in a wide range of fields. As a major application field of IoT, waste management has become one such issue. The absence of efficient waste management has caused serious environmental problems and cost issues. Therefore, in this paper, an IoT-based smart garbage system (SGS) is proposed to reduce the amount of food waste. In an SGS, battery-based smart garbage bins (SGBs) exchange information with each other using wireless mesh networks, and a router and server collect and analyze the information for service provisioning. Furthermore, the SGS includes various IoT techniques considering user convenience and increases the battery lifetime through two types of energy-efficient operations of the SGBs: stand-alone operation and cooperation-based operation. The proposed SGS had been operated as a pilot project in Gangnam district, Seoul, Republic of Korea, for a one-year period. The experiment showed that the average amount of food waste could be reduced by 33%.

1. Introduction

The Internet of Things (IoT) is a concept in which surrounding objects are connected through wired and wireless networks without user intervention. In the field of IoT, the objects communicate and exchange information to provide advanced intelligent services for users. Owing to the recent advances in mobile devices equipped with various sensors and communication modules, together with communication network technologies such as Wi-Fi and LTE, the IoT has gained considerable academic interests.

The term Internet of Things was introduced by Kevin Ashton, who was the director of the Auto-ID Center of MIT in 1999 [1]. The initial technical realization of IoT was achieved by utilizing RFID technology for the identification and tracking of devices and storing device information. However, IoT utilizing RFID technology was limited to object tracking and extracting information of specific objects. The current IoT performs sensing, actuating, data gathering, storing, and processing by connecting physical or virtual devices to the Internet. For IoT applications performing these functions, a variety of researches on IoT services including environmental monitoring [2, 3], object tracking [4], traffic management [5], health care [6], and smart home technology [7, 8] are being conducted.

Owing to the characteristics and merits of IoT services, waste management has also become a significant issue in academia, industry, and government as major IoT application fields. An indiscriminate and illegal discharge of waste, an absence of waste disposal and management systems, and inefficient waste management policies have caused serious environmental problems and have incurred considerable costs for waste disposal. To handle these problems, various researches into waste management based on IoT technology have been conducted, from studies on RFID technology to studies on waste management platforms and systems [9–13]. However, there remains a lack of research into waste management based on IoT technology or on the application of developed waste management systems in Republic of Korea.

This paper proposes an IoT-based smart garbage system (SGS) composed of a number of smart garbage bins (SGBs), routers, and servers. Each SGB, which plays a role in collecting food waste, is battery operated for mobility and, considering the convenience to residents, performs various techniques through wireless communication. The

Table 1: Comparison of the three types of volume-rate garbage disposal systems.

Type	Pros	Cons
Plastic garbage bags	(i) Convenient discharge (ii) High adaptability in poor environments	(i) Inaccurate measurements (ii) Odor problems (iii) Spoils the beauty of the city
Chips and stickers	(i) Remedies for the shortcomings of plastic garbage bags (ii) Various charge commissioning methods	(i) Inaccurate measurements (ii) Elaborate charge commissioning system required (iii) Inconvenient discharge and bin management
RFID-based garbage collection system	(i) Accurate weight measurement (ii) High impact on food waste reduction	(i) Causes server overload owing to data concentration (ii) Low mobility from a fixed power supply (iii) User inconvenience caused by complex discharge process

server collects and analyzes the status of all SGBs and resident information collected through RFID readers. The router is used for server load distribution. The proposed system had been operated as a pilot project in Gangnam district, which is one of the local districts in Seoul, the capital of Republic of Korea, according to the food waste reduction policy of the Korean government. Through the proposed system, not only food waste is reduced but also residents and the government save costs.

The rest of this paper is organized as follows. Section 2 describes the motivation for creating the IoT-based SGS. Section 3 details the architecture of the SGS and the discharge process. Section 4 presents the main techniques of the SGS. Next, Section 5 shows the implementation of the SGS in Gangnam district for a one-year period and the operation results. Finally, some concluding remarks and directions for future work are given in Section 6.

2. Motivation and Background

In existing food waste management, local governments manage food waste by deploying food waste bins and employing multiple pickup businesses for food waste collection. However, the existing food waste management method is based on a flat rate, that is, a price structure that charges a single fixed fee, which causes environmental problems and increases waste discharge because there are no restrictions on heavy producers of food waste and no incentives for lighter producers. Because food waste producers do not have a direct burden of expense for generating waste, it is difficult for their waste amounts to be efficiently reduced. Moreover, the low reliability of statistics on food waste has caused difficulty in adjusting and managing discharge amounts because a local government hires multiple pickup businesses for waste collection, and each of them uses a different measuring method.

To deal with these problems in existing food waste management, a volume-rate garbage disposal system has been introduced. In particular, in Republic of Korea, three types of volume-rate garbage disposal systems, that is, chips and stickers, standard plastic garbage bags, and RFID-based garbage collection systems, are currently being used. Table 1 describes the pros and cons of these three types of volume-rate garbage disposal systems. The most significant difference among them is that an accurate discharge weight can be obtained for an RFID-based garbage collection system when collecting food waste, which is difficult to measure accurately for chips and stickers and standard plastic garbage bags. For example, for standard plastic garbage bags, the weight of each bag may differ according to the resident's discharge habits and contents. In a chip and sticker method, although a collection box is used, thereby decreasing the allowable tolerance, accurate weight data also cannot be provided. Measuring accurate weight data is important, because of providing disposal convenience, after collecting and imposing the right duty for discharging how much food garbage they throw away.

In an RFID-based garbage collection system, an RFID collection bin includes a communication module for communicating with a central server, an RFID tag module for reading the data from an RFID card, automatic garbage entrance, and a scale function to measure the weight of the food waste. However, the collection bin communicates only with a server and lacks machine-to-machine communication with the other collection bins, which may cause a server overload. Furthermore, owing to the delay incurred from the complex discharge process of an RFID-based garbage collection system, users have a lengthy wait; in addition, an RFID-based garbage collection system lacks mobility because of a fixed power supply, causing further user inconvenience.

To solve these problems in existing RFID-based garbage collection systems, an IoT-based SGS is proposed. The proposed SGS fits into the category of IoT applied to external and public environments and was therefore designed to include the necessary components for such applications.

(i) Reliability. In IoT applied to external and public environments, communication is important for service provisioning. In particular, since this type of IoT has a wide service domain, reliable communication is necessary for devices to communicate with each other. Therefore, the SGBs utilized in the proposed system communicate with each other based on a wireless mesh network (WMN), securing communication reliability.

(ii) Mobility. IoT devices in an external environment may need to move on occasion. For a high level of mobility, the proposed system operates with a battery instead of the fixed power source that an existing RFID card system utilizes. With a battery-based power supply, the mobility of the proposed system is secured.

(iii) Service Continuity. In IoT with a wide service domain, data exchanges and services should be conducted seamlessly at any time and any location. Thus, SGBs, which communicate and exchange information based on a WMN, enable users to discharge their food waste anywhere a bin is available.

(iv) User Convenience. User convenience has been enhanced with the advent of IoT. For user convenience, the proposed SGS reduces the process delay time of the existing RFID-based garbage collection systems, which enables users to discharge their food waste without a lengthy wait.

(v) Energy Efficiency. IoT applied to external and public environments relies on an always-on infrastructure and requires mobility, causing a large amount of energy consumption. To solve this problem, the SGBs operate using energy-efficient techniques, increasing their battery lifetimes.

3. Smart Garbage System Architecture

3.1. Architecture Overview. The architecture of the SGS is shown in Figure 1. The SGBs, which are installed near apartment buildings and individual houses, exchange information with each other and send the information to the server through wireless communication. Structurally, the proposed system is divided into two domains: an administration domain and a service domain. In the administration domain, information transferred from a SGB is analyzed and processed. In the service domain, residents throw away their food waste in a SGB, and resident and SGB information is collected and transferred to the administration domain.

(i) Administration Domain. In this domain, registered resident information, payment information, and status information, such as the battery life, memory, and any malfunctions of the SGBs, are collected. To achieve this, three servers are used: a smart garbage maintenance server, a user management server, and a payment management server. The user management server manages food waste discharge information and the personal information of the registered residents who are registered in the user management server through an administrator. Furthermore, information on the discharge amount of food waste is stored and classified based on region, resident, and bin in the user management server. The charge management server conducts the payment process based on the weight of the food waste with the resident's card company. When a resident uses an RFID card to discharge his food waste, his personal card information registered on the RFID card is transferred to the charge management server, which then requests the card company to process the payment. The smart garbage maintenance server plays a role in managing all information related to the SGBs such as the amount of food waste each SGB has, the amount of food waste a collection company has gathered, and the status information of the SGBs. Thus, if a malfunction is detected in a SGB after analyzing the status information, an administrator is sent to check the problem, and the smart garbage maintenance server induces residents to use a nearby SGB. All information managed in the administration domain is also provided through a Web-based service, through which the administrators can determine the state of the system and residents can check the amount of food waste they have thrown away and for how much they have paid.

(ii) Service Domain. This domain is where the residents throw away their food waste. When a resident's RFID card touches the RFID reader of a SGB, the SGB authenticates the resident and opens the lid. The resident then throws away his food waste, and the SGB measures its weight. After the discharge process, the SGB sends the collected information on the resident and the weight of his food waste to the administration domain. Based on the collected information, a garbage collector collects the food waste from the SGB, an administrator inspects or repairs the bin, and a cleaner cleans the bin as necessary. Figure 2 illustrates the network topology of SGBs located in the service domain. The SGBs exchange information such as their capacity, battery life, and resident information through a WMN. Therefore, service continuity is guaranteed even when the same residents use different garbage bins. A header smart garbage bin (HSGB), located within each region, analyzes and manages the other SGBs within its region after collecting their information. The HSGB also exchanges this information with other HSGBs through the WMN, allowing the service continuity to be secured. Furthermore, for network reliability, if a communication problem occurs in a HSGB, header authority is delegated to the most appropriate SGB within the same region.

3.2. Discharge Process of Smart Garbage System. As mentioned above, the proposed system uses a new discharge process to minimize the delay caused by the payment and data transmission processes. Figure 3 shows a comparison between an existing RFID-based garbage collection system and the proposed system. In the existing RFID-based garbage collection system, a resident touches his RFID card to the garbage bin twice. The first touch is for resident authentication, and the second touch is for his payment. Because a data transmission between a garbage bin and a server is required before payment, the process delay incurred from the moment the food waste is weighed until the fee is paid may be lengthy, and residents may be inconvenienced. In the proposed system, however, food waste disposal and the payment process are conducted by touching an RFID card to the SGB only once, thereby reducing the process delay of existing RFID-based garbage collection systems. After the resident authentication and weighing process, the balance of the RFID card is shown on an LCD screen of the SGB using the payment data previously received from the server and the present weight of the food waste. This marks the end of the discharge process requiring the residents to wait. Then, if no other residents are waiting to use the SGB, the SGB then sends the payment data to the server through a router each time it receives a request message from the router, and the server processes the payment data of all residents and charges their fees through their credit card company. Using this discharge process, an additional RFID card touch for payment is unnecessary, which reduces the process delay.

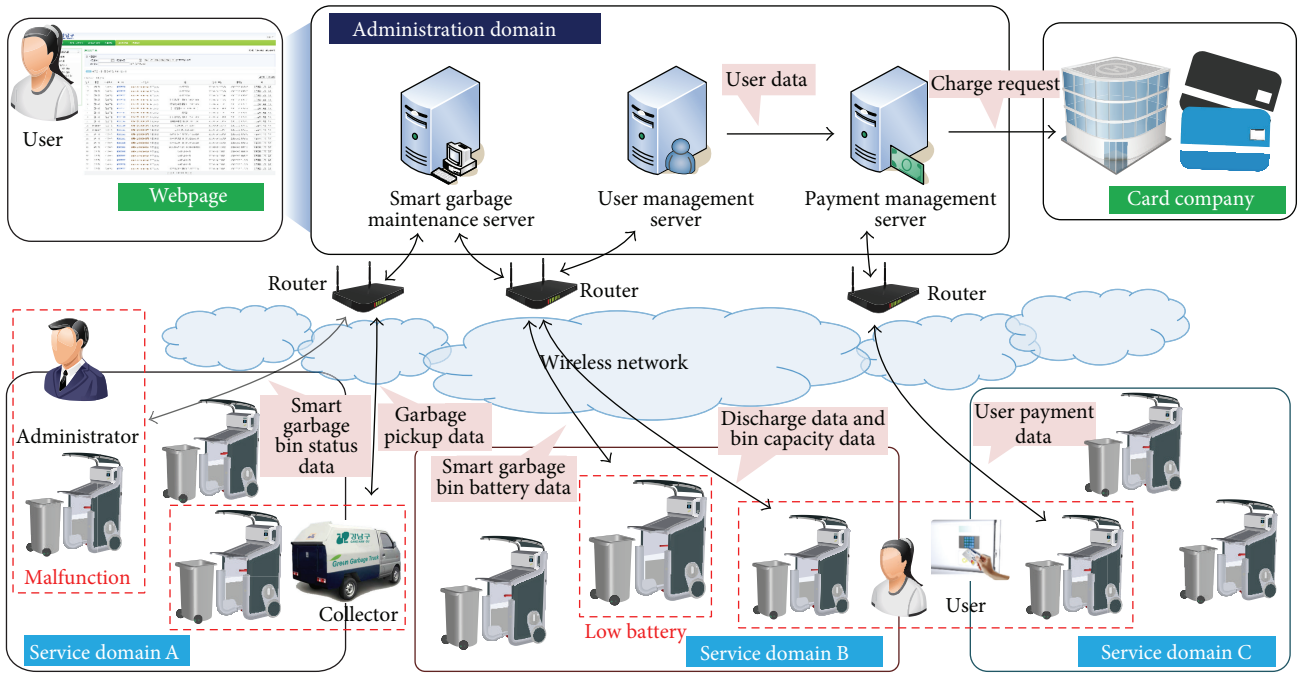


Figure 1: Overview of smart garbage system.

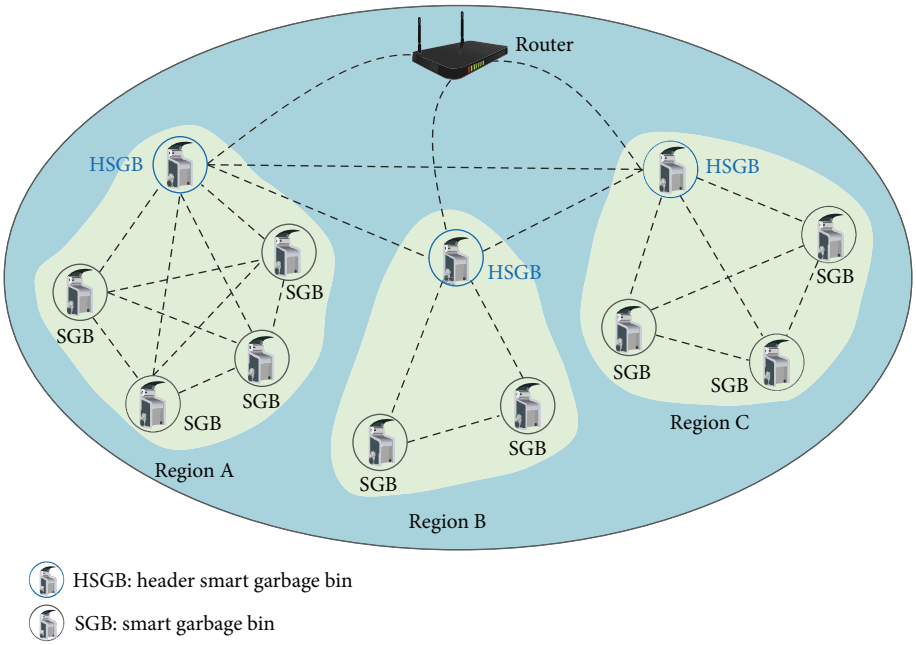


Figure 2: Network topology of smart garbage bins.

3.3. Middleware Architecture of Smart Garbage System. Figure 4 describes the entire middleware architecture of SGS. The service is based on the cooperation between the central server in the administration domain and a SGB in the service domain. The router shown in the figure is included in the administrator domain and acts as a distributed server for supplementing the centralized server's weakness when increasing the number of SGBs. Multiple routers are arranged to minimize the load and manage traffic through the server according to the number of SGBs in service.

The centralized server in the administration domain is composed of the three modules: service management, maintenance management, and charge management modules.

(i) Service Management Module. The service management module is based on the information obtained from a SGB,

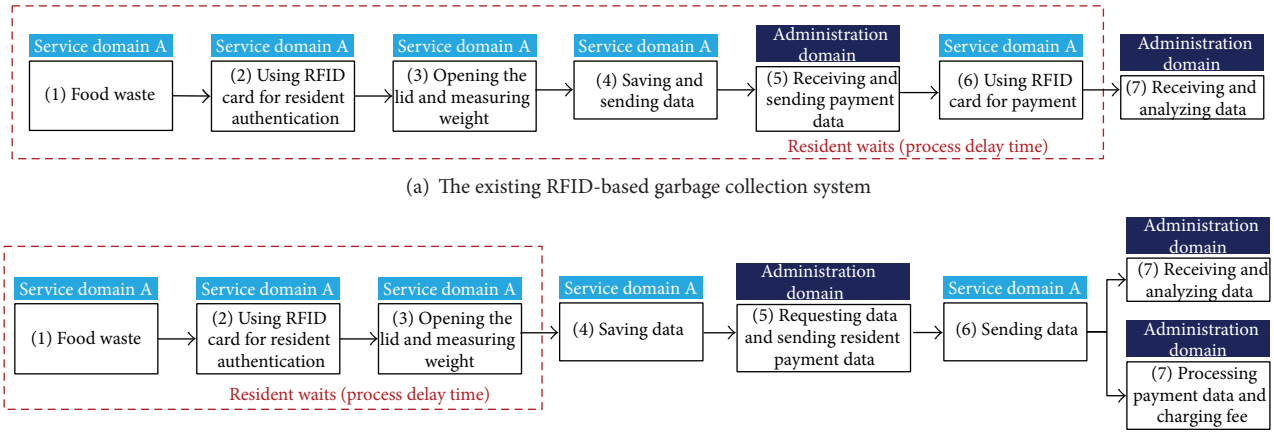


Figure 3: Discharge process of existing RFID-based garbage collection system and the proposed smart garbage system.

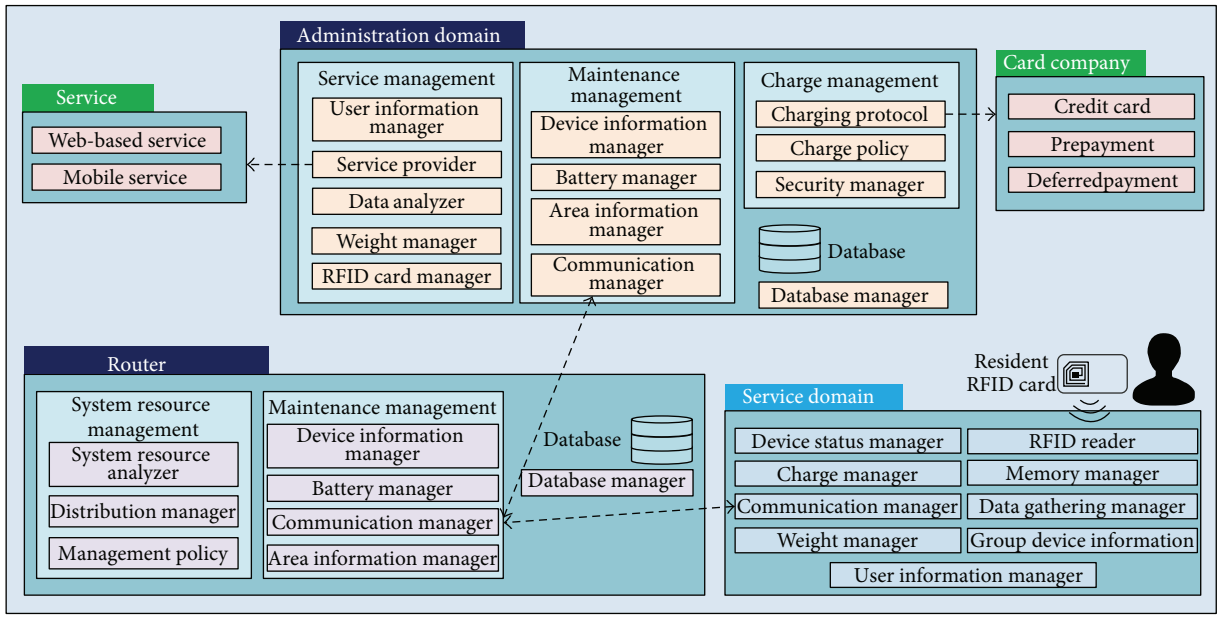


Figure 4: Middleware architecture of the proposed smart garbage system.

and it includes a user information manager for inputting or modifying the user information, a service provider to provide Web-based and mobile services, a data analyzer for analyzing information for compiling statistics, a weight manager for determining the unit price of the food waste, and an RFID card manager for managing the RFID card information.

(ii) Maintenance Management Module. This management module is composed of a device information manager to deal with information on each SGB, a battery manager to check the battery status of the SGBs, a communication manager to manage the communication status, and an area information manager for management of the area information.

(iii) Charge Management Module. This module deals with information regarding the charge process by the SGB and includes three components: a charge protocol to cooperate with an external charge interface, a charge policy component to determine the charge policy according to prepayment and deferred payment policies, and security management for encryption of the charge information.

In addition to these three modules, a database and database manager are used, the latter of which was designed for providing information required by the server or router.

Although the router normally performs maintenance management, it only takes allocated SGBs even though the server deals with all SGBs. A description of the system resource management can be given as follows.

System Resource Management Module. This module monitors the resource status of each SGB and the other routers and includes a distributed manager, which gives a specific role

to each SGB based on analyzed information regarding the status of the battery and memory, and the management policy for system resource distribution. For example, if one SGB is unserviceable, the system resource management sends the necessary information to the SGB to guide residents to neighboring SGBs.

In addition to the system resource management and maintenance management, the database manager in the router cooperates with the database manager on the server and receives the required data on the allocated SGBs.

The middleware of a SGB, located in the bottom layer of the system architecture, is composed of a device status manager module to check the status of the SGB, a weight manager to measure the weight of the inserted food waste, and a data gathering manager to process the data received from other SGBs, the router, or the server.

4. IoT Techniques of Smart Garbage System

4.1. Energy-Efficient Stand-Alone Operation of a Smart Garbage Bin. Owing to the battery-based power supply for a SGB, both basic and low-power operations of a SGB are required to improve the battery efficiency. Existing RFID-based garbage collection systems powered from electric wires are consistently in always-on mode for users. Moreover, whenever a discharge process is conducted, the bins communicate with the server for a data update. However, in the case of a battery-based SGS, there is a problem of inefficient energy use if the proposed system is used in exactly the same manner as an existing electricity-based system. Therefore, the proposed system uses an energy-efficient communication technique for battery saving.

Figure 5 shows a flowchart of an energy-efficient stand-alone operation of a SGB.

(i) Process 1. The SGBs remain in sleep mode for low-power operation. However, because a SGB should be ready to receive a request message from a router or device data from SGBs in the same region, the communication module is always turned on.

(ii) Process 2. There are three different cases of this type of process.

Case 1: a router sends a request message requiring the status information of the SGBs and resident information to the HSGB 12 times a day for a data update of the SGS. Thus, if the HSGB receives a request message from the router, the HSGB enters wake-up mode and sends all information on the residents and SGBs within the same region to the router.

Case 2: if a SGB receives a request message from another SGB in the same region, the SGB enters wake-up mode and sends the requested information to the SGB that sent the message.

Case 3: the HSGB can detect events such as communication problems and a lack of capacity or battery life. Therefore, if the HSGB detects events occurring in another SGB within its region, the HSGB enters wake-up mode and sends the event information to a router without a request message from the router.

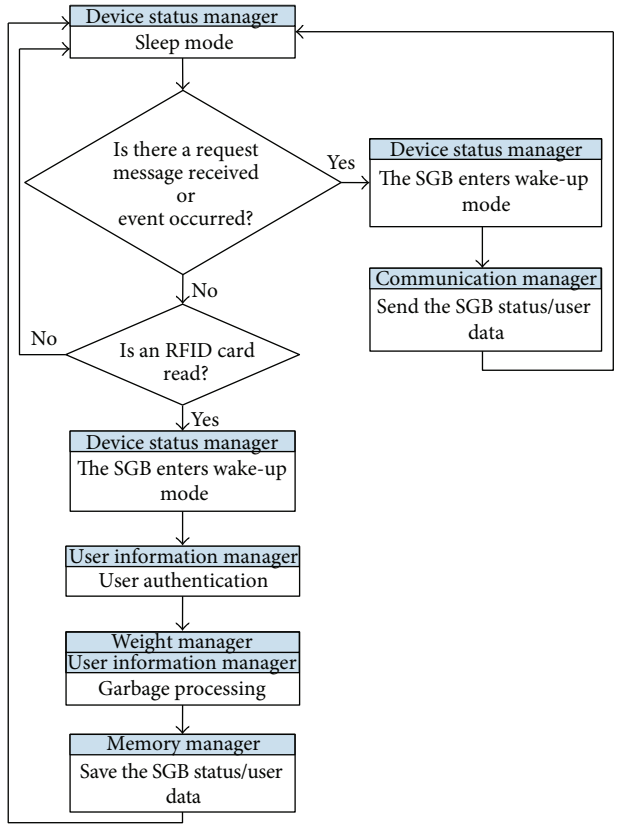


Figure 5: Flowchart of energy-efficient stand-alone operation of a smart garbage bin.

In each of these cases, the SGB or the HSGB enters sleep mode for low-power operation.

(iii) Process 3. In addition to the communication module, the RFID reader of a SGB is also always in an on state, allowing it to read a resident's RFID card at any time. Because the RFID system is event-driven, if the RFID reader of a SGB reads a resident's RFID card, the SGB enters wake-up mode and conducts user authentication and the garbage discharge process. Then, without sending any information, the SGB stores the information and reenters sleep mode.

4.2. Energy-Efficient Cooperation-Based Operation of a Smart Garbage Bin. In addition to their energy-efficient stand-alone operation, the SGBs operate in an energy-efficient manner by cooperating with each other. A router chooses a HSGB according to the battery and memory status of each SGB in the region, and the HSGB then collects information from the other SGBs. However, because the SGBs operate using a battery, a problem may occur if there is only one SGB operating as a header bin. To address this problem, the system resource management in the router checks the status of the

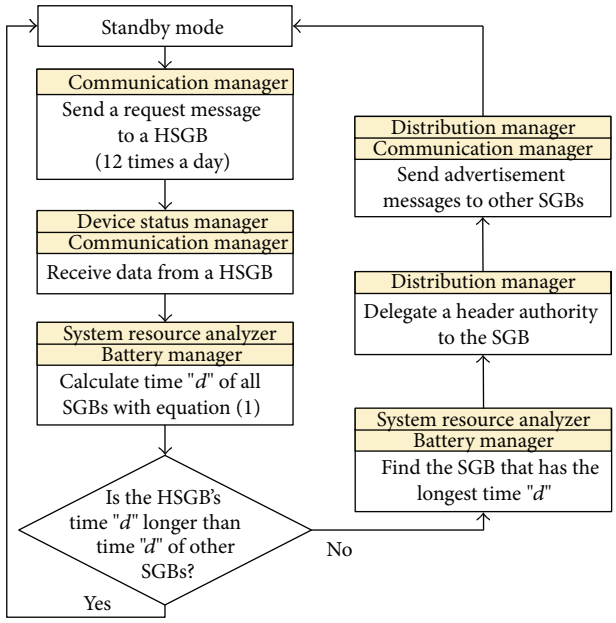


Figure 6: Flowchart of energy-efficient cooperation-based operation of a smart garbage bin.

SGBs and delegates the header authority to the SGB that has the largest amount of battery life and is the least used.

For example, take SGBs A, B, and C, with SGB A being the HSGB. SGB A is chosen to be the HSGB because it has more battery life than bins B and C and, as the least used bin, is expected to consume less energy. However, during operation, if the SGB A becomes frequently used and is thereby expected to consume a significant amount of energy, the router compares its expected power consumption and battery status with that of SGBs B and C. The router then delegates the header authority to either bin B or C accordingly.

To accomplish this process, the system resource management calculates the expected battery time of a SGB through (1). Consider

$$\frac{P_B}{\left(\sum_{i=1}^{7} P_c(i) + P_w(i)\right)/7} = d. \quad (1)$$

P_B represents the current state of charge of a battery, $P_c(i)$ is the power consumption required for communication per day, and $P_w(i)$ is the power consumption for device operation per day. Based on the power consumption for seven days and the P_B, the expected battery use time, d, of a SGB can be calculated. Therefore, based on time d calculated by the router, one of the SGBs, A, B, or C, becomes the HSGB. Figure 6 shows a flowchart of an energy-efficient cooperation-based operation of a SGB.

4.3. Adaptive User-Oriented Charge Policy. The objective of the SGS is a reduction in food waste and efficient garbage management. Even if the residents are motivated to reduce their food waste after seeing the discharge process, expecting all residents to do so may be unrealistic because the cost reduction is low.

To motivate residents to reduce their food waste, the proposed SGS applies an adaptive user-oriented charge policy in place of charging fees per kg of food waste. The basic idea of the adaptive user-oriented charge policy is that the unit cost of food waste per kg is decreased if the food waste amount of a particular month is reduced compared to the amount of the previous month.

For example, take users A and B, where user A's food waste amount for the last month was 20 kg. Therefore, if user A is charged 20,000 Korean won, the unit cost for food waste per kg is 1,000 won. However, if A's food waste amount for the current month is 18 kg, which is a 10% reduction from last month, next month's unit cost for food waste per kg will be 900 won, which is also a 10% reduction from the basic unit cost. In the case of user B, his food waste amount for last month was also 20 kg. However, if his food waste amount for the current month is 22 kg, which is a 10% increase from the previous month, their next month's unit cost for food waste per kg will be 1,100 won, which is a 10% increase from the basic unit cost. The charge policy applied to the proposed SGS can be defined through the following equation:

$$\text{Base Rate} + \left(\frac{\text{Waste Emission}_C}{\text{Waste Emission}_P}\right) \times \text{Past Changable Rate} = \text{Next Month Rate} \quad (2)$$

$$\text{Next Month Rate} - \text{Base Rate} = \text{Past Changable Rate}. \quad (3)$$

Base Rate is the basic monthly charge, Waste Emission_C is the food waste amount for this month, Waste Emission_P is the food waste amount for last month, and Past Changeable Rate is the monthly changing charge. Based on this equation, next month's unit cost of food waste per kg is applied to the residents.

Furthermore, if the capacity of SGBs in a resident region is full, the SGBs show the available SGB list on their LCD screen. In this case, since the residents have to use a SGB in another region, an additional incentive, that is, a 10% reduction in unit cost, is given to the residents to compensate for their inconvenience.

4.4. Food Waste Collection Path and Number Optimization. In existing food waste management, multiple pickup businesses are employed to collect food waste, and these pickup businesses do so from midnight to early morning using several collection vehicles. However, since these vehicles move along a fixed route and the collectors are unable to know the amount of food waste that needs to be collected, unnecessary collections may occur.

The proposed SGS proposes an efficient food waste collection system by monitoring the capacity of the SGBs. When the collectors request the status information of the SGBs along their collection route from the server through a smartphone, the server provides the information on the location and number of SGBs that need to be collected by utilizing the area information in the server middleware to the

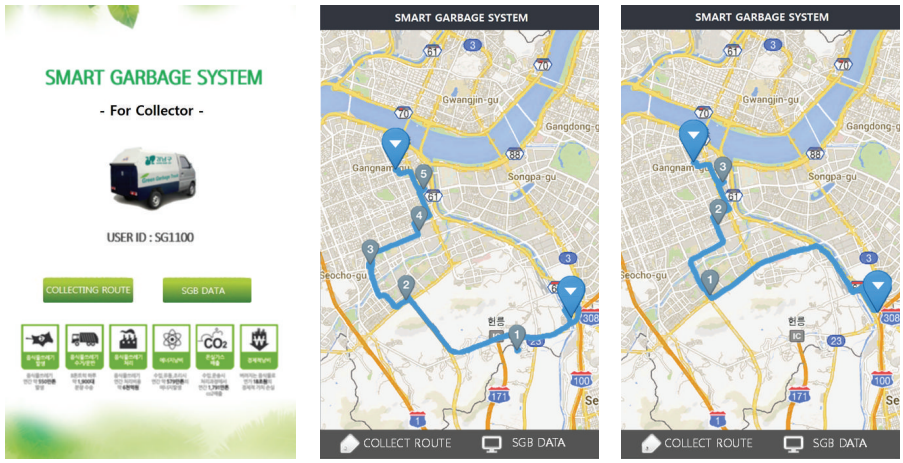


Figure 7: Mobile application for the collectors.

collector's smartphone. Figure 7 shows a mobile application that uses an open-map API. The mobile application shows the location of the SGBs that need to be collected, as well as the optimized collection path generated based on the status information of the SGBs.

Food waste collection is commonly conducted once per day. However, in a commercial area where there is more food waste than in other locations, the food waste collection should be performed more than once per day. The server therefore adjusts the food waste collection time based on the total amount of food waste accumulated over the past seven days:

$$\sum_{t=1}^{24} \frac{E_t}{k} = S, \qquad \left(\sum_{t=1}^{24} \frac{E_a}{7} = E_t \right), \tag{4}$$

where E_t is average discharge amount of food waste per hour, k is capacity of a smart garbage bin, S is number of food waste collection, and $\sum E_a$ is total amount of food waste at a certain time.

The above equation is for the food waste collection time interval. Based on the average discharge amount of food waste per hour, the number of food waste collections is calculated. Using the value of S, the SGS adjusts the collection time and establishes efficient collection plans.

4.5. Event-Based IoT Techniques for the Smart Garbage System. User convenience should be considered as the first priority in the operation of SGBs. Therefore, for service continuity and user convenience, the SGBs should induce the residents to use them by cooperating with each other when events such as a lack of capacity or battery occur. Furthermore, when a communication problem occurs in a HSGB, the header authority is delegated to another SGB for communication reliability. A sequence diagram for the operation of the SGS for two different events is illustrated in Figure 8.

(i) Event 1: Lack of Capacity or Battery. In an existing RFID-based garbage collection system, residents may be unable to discharge their food waste owing to a lack of capacity or when the garbage bins are turned off during the discharge process because of a lack of battery power. The proposed system, however, can prevent such events before they occur. After the discharge process, a SGB stores its status information. At this time, if the capacity of the SGB exceeds 90% or if the battery life drops below 5%, the SGB sends its status information to the HSGB and enters sleep mode. The HSGB, which has received the status information, then checks the other SGBs within the same region. The HSGB then sends a control message and the status information of the other SGBs to the SGB in which an event occurred to show a list of available SGBs on the LCD screen. In addition, the HSGB sends all information on its group of SGBs to the server through a router. The server then updates the Web page, sends SMS messages to the residents located in the area where the event has occurred, and sends an administrator to take action.

(ii) Event 2: A Communication Problem Occurs. If a communication problem occurs in a certain SGB, the problem can be detected when SGBs communicate with each other. The communication problem is then reported to the server through the HSGB. However, in the case of a communication problem in the HSGB, the header authority should be delegated to another SGB. Thus, if a SGB does not receive an acknowledgement message from the HSGB within 5 seconds after sending data, the SGB detects that a communication problem has occurred in the HSGB. The first SGB that detects the problem reports the situation to the router. The router then calculates the available battery life of the other SGBs in the same region and delegates the header authority to the most appropriate bin.

5. System Implementation and Experimental Results

5.1. Hardware Structure of a Smart Garbage Bin. Table 2 shows the specifications of a SGB. A load cell for measuring the weight of the food waste is located at the bottom of each SGB. The full size of a SGB was determined by considering

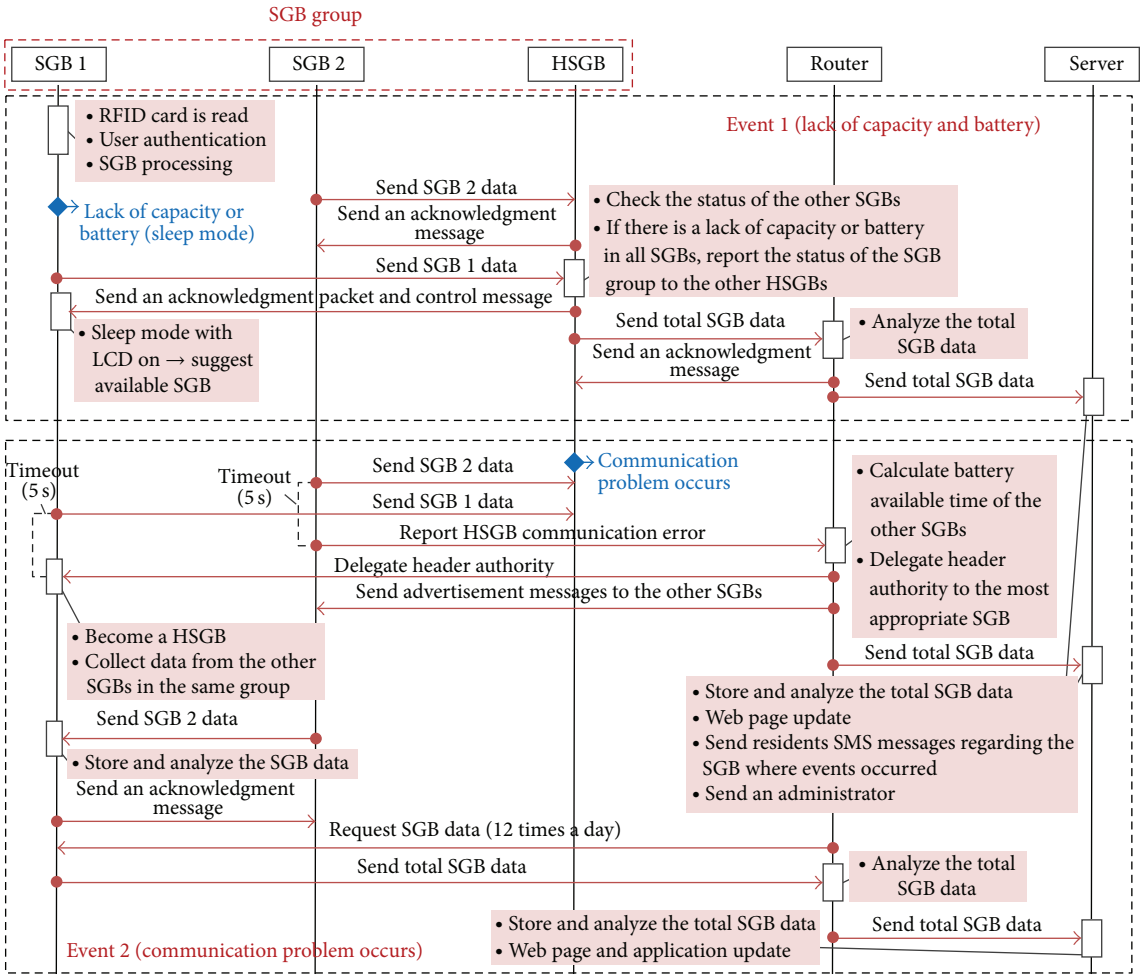


Figure 8: Sequence diagram for operation of the smart garbage system for two different events.

Table 2: Specifications of a smart garbage bin.

	Specification
Type of scale	Substructure
Size	590 × 680 × 1170 mm
Battery capacity	Lithium-ion battery, 7.4 V, 92.4 Ah
Communication	CDMA2000 EV-DO
Type of RFID	ISO 14443, frequency: 13.56 MHz
Weight	Maximum weight: 105 kg, weight unit: 50 g

the height of the users. Furthermore, for mobility, a lithium-ion battery is utilized as a power supply. However, depending on the circumstances, a SGB can use a fixed power source.

The hardware structure of a SGB is composed of eight modules: load cell, main system, interface, modem, motor, LCD display, AD converter, and RFID reader. First, the load cell [14] measures the initial analog value and sends it to the main system through the AD converter module attached to the main system. The AD converter module converts the analog value into a digital value. The value processed by the AD converter module is converted into a weight result in the main system. During this process, the characteristic of the load cell is not linear according to the weight change, and thus it should be corrected in the main system. The weight result is transferred to the interface. The interface sends the result to the modem or LCD display as demanded. Moreover, the interface also manages all the operations in the SGB, such as analyzing the input data from the RFID reader and driving the motor to open or close the lid of the bin. The actual SGB is shown in Figure 9.

5.2. System Implementation. The proposed SGS had been operated as a pilot project in Gangnam district. In total, 136 SGBs were deployed in Gangnam's six subdistricts. The bins were applied to apartment housing areas in five of the subdistricts and to detached housing areas in the other district. Figure 10 shows the locations where the SGBs were deployed, their number, and the system implementation.

As shown in Figure 10, a SGB is structured with a conventional food waste bin placed inside. System implementation was performed by simply placing the SGB at the location where a conventional food waste bin was previously located and fixing the conventional food waste bin inside the smart bin. In addition, since the SGB operates on battery

(a) Smart garbage bin (b) LCD display, keypad, and RFID reader

FIGURE 9: (a) Smart garbage bin and (b) LCD display, keypad, and RFID reader.

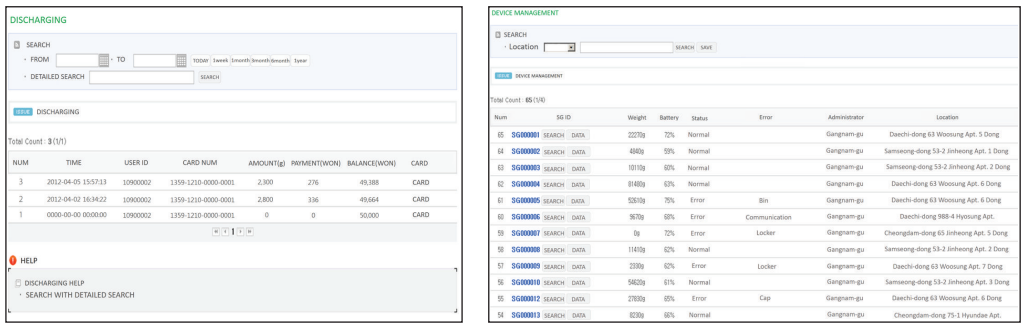
FIGURE 10: Implementation of a smart garbage system.

power, additional construction connecting it to a neighboring commercial electricity line was unnecessary.

The Web-based service structure is presented in Figure 11. The SGS provided an ID and password to each user for their RFID card and Web-based service. The users were divided into three classes: an administrator, collector, and the residents. The administrator can check the present and accumulated amount of food waste of each SGB, the status of all SGBs, and the time log. The administrator is then able to classify the information based on region, resident, and SGB. Moreover, a service enabling new users and RFID cards to be registered was provided to the administrator. While the administrator is given complete authority, a resident can only check the discharge amount of their food waste and payment information, and the collector can check the status of the collected food waste and receive a notification whenever the capacity of a SGB exceeds 90%.

5.3. Experimental Results

(i) Energy Efficiency. To verify the energy efficiency of stand-alone and cooperation-based operations, two comparison



(a) User interface Web page (b) Administrator interface Web page

Figure 11: Web-based food waste management service.

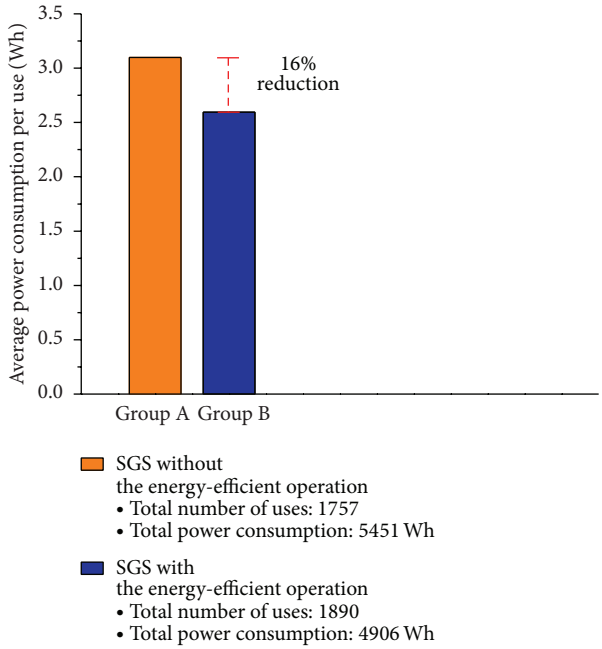


Figure 12: Comparison of average power consumption per use in Groups A and B.

Table 3: Experimental results of Group A.

SGB	Number of uses	Battery remains (%)	Power consumption (Wh)
1	161	29.93	479
2	151	28.99	485
3	197	10.89	609
4	204	8.26	627
5	196	6.59	638
6	144	38.02	423
7	202	6.096	642
8	156	27.51	495
9	177	22.59	529
10	169	23.77	512

groups, A and B, were set up. Group A had ten SGBs operating in the same manner as an existing RFID-based garbage collection system. Therefore, group A was normally kept in sleep mode, and it entered wake-up mode whenever a user utilized a bin. Group A conducted the discharge process and communicated with the server every time this process was finished. Furthermore, the SGBs in group A communicated with the router, and no header bin was used. For group B, which also had ten SGBs, a HSGB was delegated by the router and was changed to another bin depending on the battery status of all SGBs in the area. The two groups were used for a two-week period. To generate identical experimental conditions, two locations with a similar number of users and households were selected. The experimental results of Groups A and B are detailed in Tables 3 and 4, respectively. And Figure 12 shows the comparison of average power consumption per use in Groups A and B.

Table 4: Experimental results of Group B.

SGB	Number of uses	Battery remains (%)	Power consumption (Wh)
1	211	20.07	546
2	199	26.75	500
3	172	35.96	437
4	223	12.50	598
5	192	28.94	485
6	154	39.97	410
7	180	31.35	469
8	207	18.28	558
9	171	37.08	430
10	181	31.47	468

From the results, when the same number of service provisions and the same quality are assumed, the energy efficiency improved by 16% for the SGB group that applied an energy-efficient operation. Although there was little effect on the energy efficiency of device operation when including the opening of the lid and the use of an LCD screen, energy efficiency improvement was achieved by controlling the amount of communication with the server or router and applying an energy-efficient operation.

(ii) Food Waste Reduction. To evaluate the performance of the proposed SGS, the amount of food waste discharged by

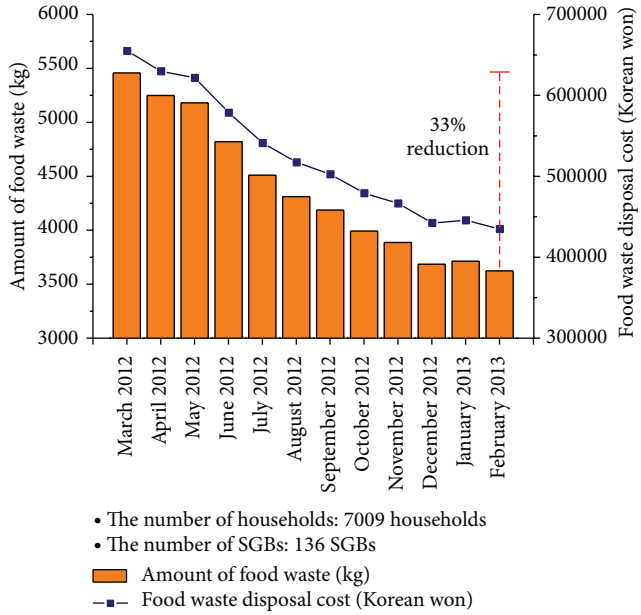


Figure 13: Amount of food waste discharged by the inhabitants of Gangnam district and disposal cost according to the amount of food waste for a one-year period.

the inhabitants of Gangnam district was analyzed. Because the proposed SGS provides a Web-based service, the amount of food waste can be easily analyzed through statistical data collected by the Web-based service.

As mentioned before, 136 SGBs were deployed in the Gangnam district, which consists of six subdistricts. The bins were applied to an apartment housing area in five subdistricts and to detached housing areas in the other district. The adaptive user-oriented charge policy was also applied to the SGS.

Figure 13 shows the amount of food waste discharged by the inhabitants of Gangnam district and disposal cost according to the amount of food waste for a one-year period. There was not a notable result for the initial three months but significant reduction results occurred from June 2012, when an effect of the adaptive user-oriented charge policy appeared. In the last month, the amount of food waste generated per month was decreased by about 33%.

Of course, it is somewhat difficult to consider that the decreased amount accurately shows the performance of the SGS owing to the reliability of a conventional collection system. In addition, the reduction in food waste may be a temporary phenomenon caused from an aversion to the new system. If the reduction occurred constantly, the expectation effectiveness because of the reduction caused by the Pay as You Throw (PAYT) system based on RFID would be expected to be expanded.

6. Conclusions and Future Works

In this paper, we proposed an IoT-based SGS for replacing existing RFID-based garbage collection systems. To provide differentiation from passive collection bins and other types of RFID-based food garbage collection systems, we also proposed components required in external and public environments and designed the SGS based on these components. The basic system structure of a SGB is a centralized structure in which information gathered in each bin is transferred to the server; we also designed a HSGB for improving the battery efficiency of each SGB.

An adaptive user-oriented charge policy is used to motivate residents to reduce their food waste, and Web-based services are provided to achieve more efficiency in the disposal and collection processes. Furthermore, based on the proposed system using SGBs, we implemented the proposed system in Gangnam district for a one-year period as a pilot project and verified the results. The energy efficiency of the proposed SGBs shows 16% energy saving result, which shows that SGBs can contribute to not only a reduction of food waste but also energy saving. The proposed system along with the adaptive user-oriented charge policy resulted in a reduction of food waste of about 33%, and it is expected that the proposed system will thereby improve the efficiency of food waste management.

Nevertheless, the proposed SGS requires more maintenance cost than the existing system, and there is a tradeoff owing to the proposed system's battery-based power structure. The most important issue is how to improve the battery life of a SGB. To solve this problem, photovoltaic power generation [15] is being considered. Moreover, high-intensity plastic materials are also being considered for durability against external impact and corrosion from humidity.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Human Resources Development (no. 20124030200060) of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant funded by the Korean government Ministry of Trade, Industry and Energy, by the MSIP (Ministry of Science, ICT and Future Planning), Republic of Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1015) supervised by the NIPA (National ICT Industry Promotion Agency), and by the Chung-Ang University Excellent Student Scholarship.

References

[1] K. Ashton, "That "internet of things" thing," *RFiD Journal*, vol. 22, pp. 97–114, 2009.

[2] M. T. Lazarescu, "Design of a WSN platform for long-term environmental monitoring for IoT applications," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 3, no. 1, pp. 45–54, 2013.

[3] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, "Towards the implementation of IoT for environmental condition monitoring in homes," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3846–3853, 2013.

[4] K. Gama, L. Touseau, and D. Donsez, "Combining heterogeneous service technologies for building an internet of things middleware," *Computer Communications*, vol. 35, no. 4, pp. 405–417, 2012.

[5] L. Foschini, T. Taleb, A. Corradi, and D. Bottazzi, "M2M-based metropolitan platform for IMS-enabled road traffic management in IoT," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 50–57, 2011.

[6] A. J. Jara, M. A. Zamora, and A. F. G. Skarmeta, "An internet of things-based personal device for diabetes therapy management in ambient assisted living (AAL)," *Personal and Ubiquitous Computing*, vol. 15, no. 4, pp. 431–440, 2011.

[7] S. Tozlu, M. Senel, W. Mao, and A. Keshavarzian, "Wi-Fi enabled sensors for internet of things: a practical approach," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 134–143, 2012.

[8] X. Li, R. Lu, X. Liang, X. Shen, J. Chen, and X. Lin, "Smart community: an internet of things application," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 68–75, 2011.

[9] I. Nielsen, M. Lim, and P. Nielsen, "Optimizing supply chain waste management through the use of RFID technology," in *Proceedings of the IEEE International Conference on RFID-Technology and Applications*, pp. 296–301, Guangzhou, China, June 2010.

[10] Z. Lizong, A. Anthony, and H. Yu, "Knowledge management application of internet of things in construction waste logistics with RFID technology," *International Journal of Computing Science and Communication Technologies*, vol. 5, no. 1, pp. 760–767, 2012.

[11] B. Chowdhury and M. U. Chowdhury, "RFID-based real-time smart waste management system," in *Proceedings of the Telecommunication Networks and Applications Conference*, pp. 175–180, December 2007.

[12] M. A. Hannan, M. Arebey, R. A. Begum, and H. Basri, "Radio Frequency Identification (RFID) and communication technologies for solid waste bin and truck monitoring system," *Waste Management*, vol. 31, no. 12, pp. 2406–2413, 2011.

[13] P. Pratheep and M. A. Hannan, "Solid waste bins monitoring system using RFID technologies," *Journal of Applied Sciences Research*, vol. 7, no. 7, pp. 1093–1101, 2011.

[14] I. Muller, R. de Brito, C. E. Pereira, and V. Brusamarello, "Load cells in force sensing analysis—theory and a novel application," *IEEE Instrumentation & Measurement Magazine*, vol. 13, no. 1, pp. 15–19, 2010.

[15] K. Touafek, M. Haddadi, and A. Malek, "Modeling and experimental validation of a new hybrid photovoltaic thermal collector," *IEEE Transactions on Energy Conversion*, vol. 26, no. 1, pp. 176–183, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 918536, 8 pages
http://dx.doi.org/10.1155/2014/918536

Research Article

The Deployment of Routing Protocols in Distributed Control Plane of SDN

Zhou Jingjing, Cheng Di, Wang Weiming, Jin Rong, and Wu Xiaochun

College of Information & Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

Correspondence should be addressed to Zhou Jingjing; zhoujingjing@zjgsu.edu.cn

Received 3 May 2014; Accepted 13 July 2014; Published 28 August 2014

Academic Editor: Xiaoxuan Meng

Software defined network (SDN) provides a programmable network through decoupling the data plane, control plane, and application plane from the original closed system, thus revolutionizing the existing network architecture to improve the performance and scalability. In this paper, we learned about the distributed characteristics of Kandoo architecture and, meanwhile, improved and optimized Kandoo's two levels of controllers based on ideological inspiration of RCP (routing control platform). Finally, we analyzed the deployment strategies of BGP and OSPF protocol in a distributed control plane of SDN. The simulation results show that our deployment strategies are superior to the traditional routing strategies.

1. Introduction

With the rapid development of science and technology, the old-fashioned network architectures can not meet the various needs for modern life. People want to change the existing network architecture urgently and set out to redesign new network architecture. The future network should have these properties, such as the underlying data plane which is dumb, simple, and minimal, the separation of control plane and data plane, the control plane which can completely control the entire network, and the upper layer which provides a common application programming interface of external parts. In this way, researchers can program via calling API (application programming interface) on the control plane to achieve innovation of network.

Software defined network is the powerful enabler of owing innovative network ideas. SDN revolutionizes the existing network architecture to provide the methods of programmable networks and decouples network architecture into the data plane, control plane, and control plane applications. So, SDN separates the data plane and control plane and, at the same time, improves the performance and scalability of network. Through the functions of centralized control plane, the open capabilities of network programming, SDN can reduce or even get rid of the limitations of the network infrastructure and architecture to improve the network efficiency.

In the architecture of SDN, routing control and topology control are still the core functions of the control plane. Centralized control of the routing causes some shortcomings, such as the bottlenecks of performance, a failure of single point, and poor scalability. We will research the deployment of routing protocols in distributed control plane of SDN in this paper. Currently, Kandoo is a hot spot of distributed architecture of SDN; we will research the crucial problem of routing protocol based on distributed architecture of Kandoo, such as the internal communication in distributed control plane and the deployment of distributed routing protocol.

The paper is organized as follows. Firstly, we research the distributed architecture of Kandoo which changes the traditional architecture of single control unit to form a distributed control plane architecture of multicontrol unit, thereby to achieve interconnection of multiple controller units. Then, we will analyze the distributed routing protocol of SDN and research the efficient implementation and deployment of distributed control plane. Then, we improved and optimized the two levels of controller of Kandoo based on the idea of RCP and analyzed the BGP and OSPF routing protocol using the embodiment of SDN distributed control plane.

Finally, we give the simulation results, which show that our deployment strategies are superior to the traditional routing strategies.

2. Related Work

Forwarding and Control Element Separation (ForCES) Working Group (WG) in IETF Routing Area is one of the most influential research organizations in open programmable network research area [1]. The WG specializes in the architecture and protocol standards of open programmable IP network element (NE, such as router, firewall [2], or load balancer).

Open programmable networks are considered as the most prospective architectural approach to meet the above demands. In open programmable networks, a NE (e.g., a router/switch) is systematically separated into a control plane and a forwarding plane. Forwarding plane receives packets from outer networks, processes the packets according to functional requirements of the NE, and then outputs the packets back to outer networks. Forwarding plane usually needs the ability to process packets at line speed. Control plane controls forwarding plane for the whole forwarding process and provides adequate parameters for the process. More importantly, the interface between the control plane and the forwarding plane is standardized. Moreover, resources at the forwarding plane, which are used to process packets, are also described in a standardized way [3]. As a result, control plane can access and control the forwarding plane resources in a standard way. This makes it feasible for control plane and forwarding plane to be separated at their product level; that is, control plane and forwarding plane as separate products from different vendors can work together to form one NE with full interoperability [4]. On this basis, ForCES achieves the separation of the control software of the device and the underlying hardware physically and the virtualization of a variety of basic network functions module. The ForCES Working Group has completed the formulation of the ForCES Requirements, the ForCES Framework. The formulation of ForCES protocol and ForCES FE Model has been basically completed [5].

Before the concept of SDN is proposed, ForCES has already become the crucial technology of Forwarding and Control Element Separation, and, on the basis of ForCES, SDN will get more adequate theoretical support.

3. SDN Distributed Control Plane Architecture

Architecture of Kandoo has changed the traditional structure of a single control unit and formed a distributed multicontrol plane architecture by interconnecting to multiple controller units [6]. Control plane of Kandoo can distinguish local controller applications from nonlocal applications substantially. Kandoo establishes a two-level controller: (a) local controller which performs local application as close as possible to the switch and (b) running logic centralized root controller of nonlocal application. As shown in Figure 1, on the one hand, a number of local controllers are deployed throughout the network, and each controller controls one or a small number of switches. On the other hand, the root controller controls all local controllers.

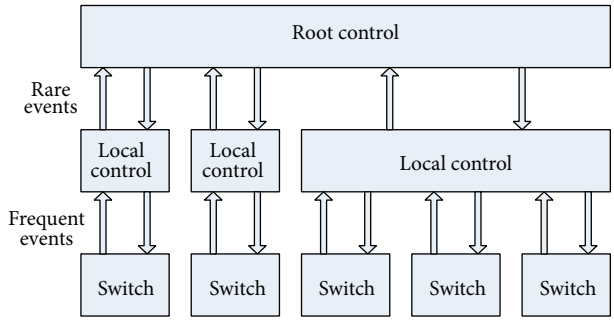


Figure 1: Kandoo two levels of controllers [4].

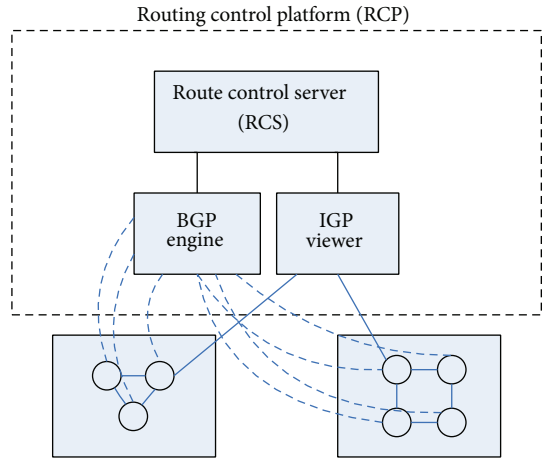


Figure 2: Schematic structure of the routing control platform.

4. Distributed Routing Protocol for SDN Architectures

Before the advent of Kandoo architecture, MIT and AT & T have proposed the idea of separating routing from router [7]. According to these thoughts, they proposed the routing control platform (RCP). This architecture is based on the circumstances in which the network topology and the corresponding management strategies deal with routing and exchange the reachable message between different autonomous domains. As shown in Figure 2, the control platform consists of three modules: IGP indicator, BGP routing engine, and router control server. Network distributes the functions of measurement and management through a number of different routers. But it is difficult to quickly perform the strategy of wide area network or deploy new services. As for such questions, RCP provides the direct capabilities of network control for the network operator, rather than indirectly affecting the network through a router. RCP can reduce the router's configuration status remarkably, thereby decreases the configuration errors and diminishes the complexity of

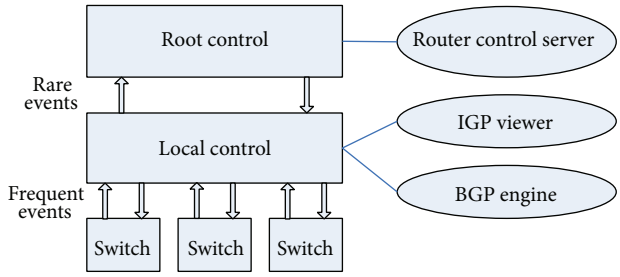


Figure 3: Improved Kandoo architecture based on RCP.

(0) Ignore if egress router unreachable
(1) Highest local preference
(2) Lowest AS path length
(3) Lowest origin type
(4) Lowest MED
(5) eBGP-learned over iBGP-learned
(6) Lowest IGP path cost to egress router
(7) Lowest router ID of BGP speaker

Algorithm 1: Steps of the process of BGP route selection.

network software management, and can deploy application services quickly.

Inspired by the thought of the controller function modules of the RCP architecture, we improved and optimized the two levels of controllers of Kandoo and analyzed the more efficient implementation of the routing process of BGP and OSPF protocols in the distributed control plane of AS autonomous system; the specific structures are shown in Figure 3.

In the following, we will describe the implementation details of local controller and the root controller on distributed BGP protocol.

4.1. The Function Expansion of Local Controller. The local controller completes the function of IGP indicator and BGP engine.

The IGP indicator can monitor the IGP topology and provide the information to the root controller; the IGP indicator also can create IGP adjacency to accept the link state advertisements (LSAs) of IGP. In order to ensure the IGP indicator does not route packets, we set a big IGP weight between the IGP indicator and the routers. The IGP indicator can keep the newest topology state of IGP.

The BGP engine can maintain the iBGP sessions of each router in the AS system. These iBGP sessions allow the root controller to understand the candidates of routing. The BGP engine also can communicate the routing decisions with other routers. The iBGP runs over TCP, so the BGP engine does not need to be adjacent to each router physically.

We make a reasonable assumption as the connecting of the two IGP endpoints is sufficient to establish a BGP session. In fact, the continuing obstruction and incorrect configuration will affect that assumption, but these situations are unusual cases. Usually, the router will configure the BGP packets to forwarding path with high priority in order to ensure the transmission of these packets.

In order to accept the BGP updating, the root controller will send the BGP routing to the router using iBGP session. Because BGP updating has the property of next hop, the BGP engine can advertise BGP routing to other routers with the next hop. This feature does not allow BGP engine to forward packets. The BGP routing usually carries the attribute of next hop based on the egress router. Therefore, the root controller can send the routing to the router, the router's next hop is not changed, and the router can forward the packets to the exit routing.

The interaction of BGP engine and a router is the same as the interaction of the BGP spokesman and the router. But the BGP engine can send a different routing to a router. After choosing a new optimal routing from the neighbor AS, a router will send the BGP updating to the BGP engine. Similarly, BGP engine sends the update message only when the routing of a router needs to change.

4.2. The Function Expansion of Root Controller. Root controller accepts the message of IGP topology and BGP routing from the local controller, then calculates the best routing for a group of routers, and assigns the results to the appropriate router through the BGP engine. According to the selection process of BGP routing, in the first step of the route table, the best routing has been selected from a number of candidates of routers. And the root controller no longer assigns the routing to the router. In order to make the right routing decisions for a group of routers in the same partition, it must meet the condition in which the root controller must be able to receive the topology message of IGP and the routing message of BGP in this partition.

Although the root controller has considerable excellent flexibility in assigning routes, a more reasonable approach is to choose routing in condition of iBGP configuration of the whole network. For the purpose of simulating an iBGP configuration of the whole network, according to Algorithm 1, the root controller needs to perform BGP routing process.

The reasons that root controller can perform calculations are as follows.

(a) It knows the IGP topology; the root controller can select the reachable egress routers from their visible routers in partition.

(b) From step 1 to 4, the corresponding property is compared through BGP message.

(c) Step 5: the root controller learns about message of iBGP through other routers and considers learning eBGP message.

(d) Step 6: the root controller compares the path cost of IGP via the message which IGP engine publishes.

(e) Step 7: because iBGP message maintains each router and the BGP engine, the root controller knows the routing ID of each router. By calculation, the root controller sends the proper routing to each router.

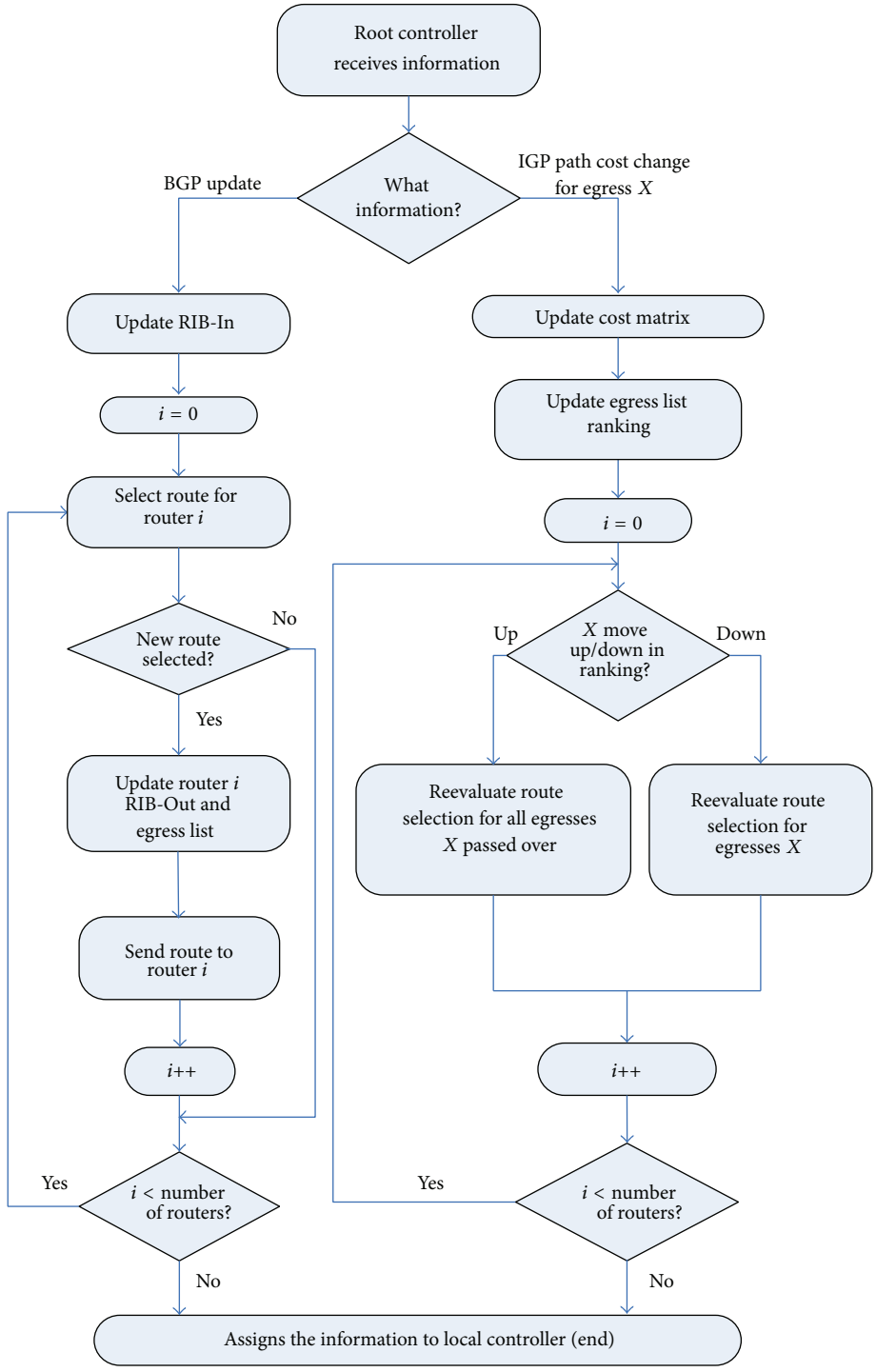


Figure 4: The implementation process of BGP protocol on Kandoo.

5. The Deployment Strategy of Distributed Routing Protocol of SDN

5.1. The Deployment of Distributed BGP on the Root Controller of SDN. Root controller acts the functional roles as BGP speaker of AS. Root controller receives the information from the local controller. Figure 4 shows the processes of the implementation of root controller. Root controller receives the updating message from the local controller, and the learned routing message is stored in the routing table. Root controller performs the routing of each router and stores the selected routing in the RIB-Out table. The RIB-In table maintains the routing cluster of each prefix, and each BGP has the property of next hop to uniquely identify the egress router.

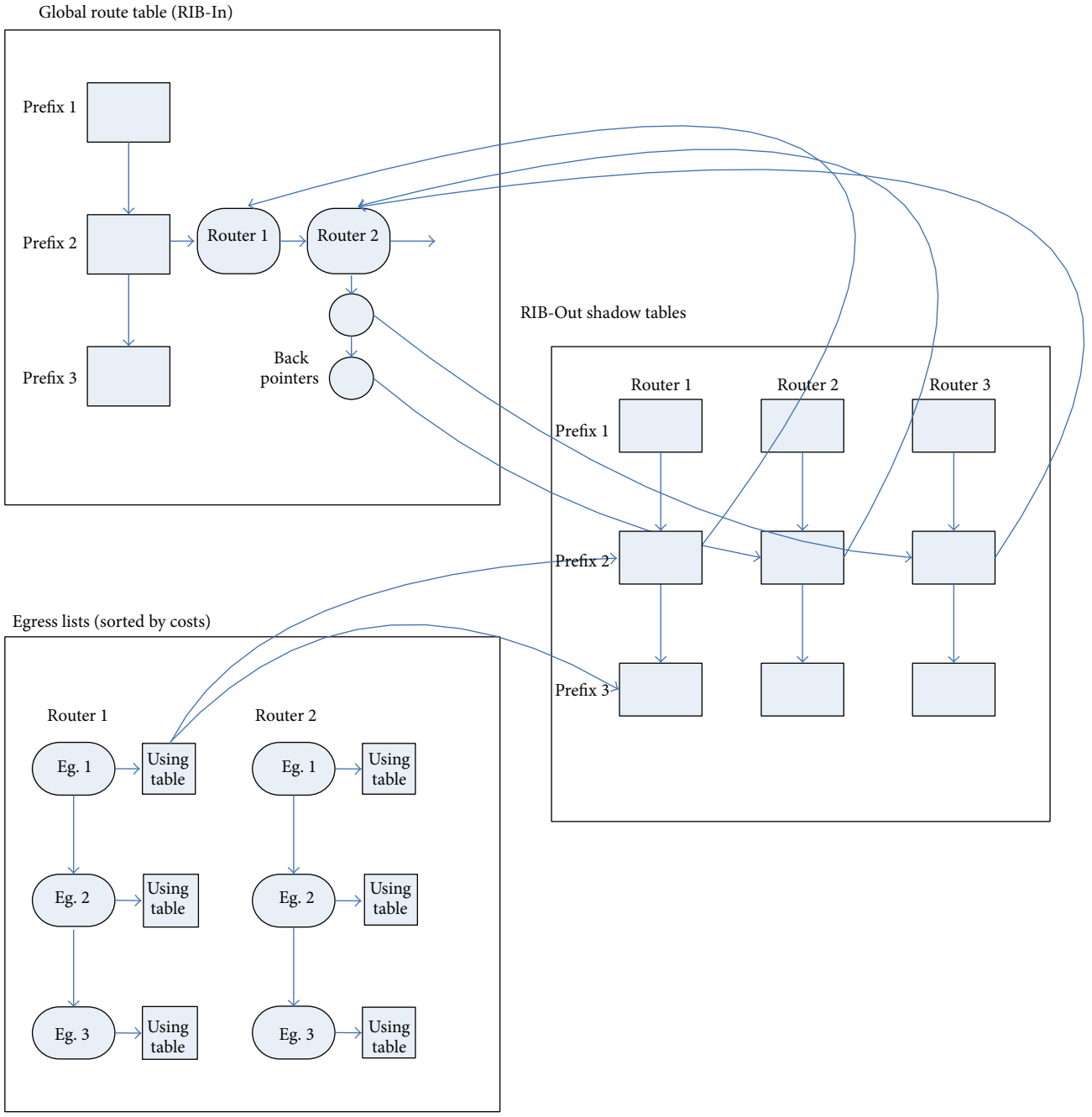


Figure 5: Data structures of RIB-Out and RIB-In route table and egress lists of root controller.

Root controller also accepts the cost of IGP path of each router from the local controller. Root controller calculates the optimum message of BGP routing using the RIB-In table, then assigns the information to local controller, and sends updating message to the router. When you receive the change of cost information of the path from the local controller, if the root controller makes the decisions of selecting the best routing using step 6, namely, considering the cost of IGP path, the root controller should recalculate the best routing.

To improve the ability of calculating the routing message of the root controller, the method is shown as follows.

When root controller calculates the best routing, finding the affected routing is actually a process of high price. We present a method for efficient execution based on level arrangement of the egress router. The details of the method are shown in Figure 5.

5.1.1. Only Storing One Copy of the BGP Routing. It needs a large number of additional stored overheads to store the detached copy of BGP routing of each destination prefix. In order to reduce the storage overhead, root controller routing message is stored only in RIB-In table. The property of next

hop of BGP routing can uniquely identify the egress router. According to the updating packets, root controller searches the RIB-In table based on the prefix and adds, updates, or deletes the corresponding routing based on the property of next hop. To implement the RIB-Out table, root controller table should regard the router images containing RIB-In pointer as a search tree of prefix. Figure 5 shows two examples of how to implement the RIB-Out from the RIB-In pointer.

5.1.2. Keeping in Touch with the Routers Which Have Been Assigned to Each Routing. When one routing path is withdrawn, root controller needs recalculation of a new routing for the router which is using the routing path. In order to identify the affected routers quickly, the route which is stored in the RIB-In table should contain a reverse pointer list pointing to the router. For example, in the RIB-In table of Figure 5, routing path 2 with the prefix 2 has reverse pointer indicating that router 2 and router 3 have been assigned to that path.

5.1.3. Maintaining One Array of Egress Based on the Cost of IGP Path for Each Router. The change of single IGP path cost may affect the BGP decision of root controller for the egress routing of the destination address prefix. In order to avoid the reselection of routing for every prefix and every router, root controller needs to maintain a level arrangement of egress router, as shown in Figure 5, the egress lists. For each egress, root controller stores the pointer to the prefix and route link. For example, router 1 arrives at prefixes 2 and 3 through egress 1. If the IGP path cost increases from router 1 to egress 1, the BGP speaker puts down the level of egress 1, until it comes up with the higher one. Finally, the root controller recalculates the prefix of egress 1 to get the BGP routing.

5.1.4. Assigning Routing to a Group of Associated Routers. Each router does not need to calculate the BGP routing, and the root controller can assign the BGP routing to a group of routers with the same destination prefixes.

5.2. The Deployment of Distributed OSPF on Local Controller. Local controller connects one or more routers to accept the link state advertisement. While local controller accepts the message of BGP routing, it sends the BGP routing messages to the root controller and also sends BGP routing to single router. Local controller maintains a newest network topology and calculates the path cost of each pair of the routers. Figure 6 shows the principle of the running condition of OSPF protocol on the local controller of SDN.

Through analyzing the architecture of Kandoo, we know that one of the important works of the local controller is to download the task of the root controller. We will study the methods of how to change local controller to reduce the load of root controller. To improve the ability of downloading the load of root controller to the local controller, the methods are studied as follows.

5.2.1. Only Sending the Change Information of Path Cost. Even if the network is in a stable condition, local controller not only creates LSA according to the changes of network, but also regularly updates the LSAs message. Local controller maintains the steady status of network according to the topology model and determines whether to change or not the network topology of the updating LSA. For the changed LSA, local controller calculates the shortest path to determine the newest cost of path from the perspective of each router. Local controller does not send the information of all the paths to root controller and only sends part of the information which has been changed through the calculation of path cost.

For the purpose of scalability, OSPF domain is divided into several regions to form the topology with a radiation center. Area 0, as the backbone of the region in the center, provides connections to other nonbackbone areas to form radiation. Each link belongs to a determined area. A router connecting multiple regions is known as the backbone router, which is also known as local controller. The local controller will learn the whole topology of the area. The local controller does not learn the topology of other remote areas; but it will learn all the information of path cost to a remote routing node.

Local controller performs the shortest path first (SPF) algorithm according to the whole topology; it looks like ignoring of the domain boundary, but OSPF will assign the path of the routers belonging to the domain to be inside the domain, despite the existence of the shortest path of cross domain boundaries. So, it does not ignore the problem of domain boundary in the process of calculation, therefore using two stages to calculate. The first stage, called the inside-domain phase, as shown in Figure 6, calculates the path cost for each domain using the LSAs message within the domain. In the second stage, called outside-domain stage, local controller calculates the path cost for the router in different domains through the integration of paths.

5.2.2. Reducing the Load of Root Controller through Aggregation Router. Local controller can use the structure of domain to reduce the number of routers of root controller. To achieve this purpose, local controller requires (1) provision of the path cost information of nonbackbone domain routers and backbone domain routers and (2) the formation of each of the nonzero domains as a router group and its provision of information of the group. Furthermore, local controller need not be physically connected to a nonzero domain, because the total LSAs message obtained from the backbone domain allows the local controller to calculate the path cost from the router of the zero domain to the other routers. Meanwhile, local controller will determine the group relationship of routers from the total message of LSAs.

5.2.3. Caching BGP Routing Message. Local controller stores the message of RIB-In and RIB-Out table locally. When a failure of BGP speaker occurs, the caching of RIB-In will bring in a new duplication or recover rapidly according to the latest duplication in the condition of no influence to the routers. When the connection of IGP is interrupted temporarily, the caching of RIB-Out will resend the routing to this router.

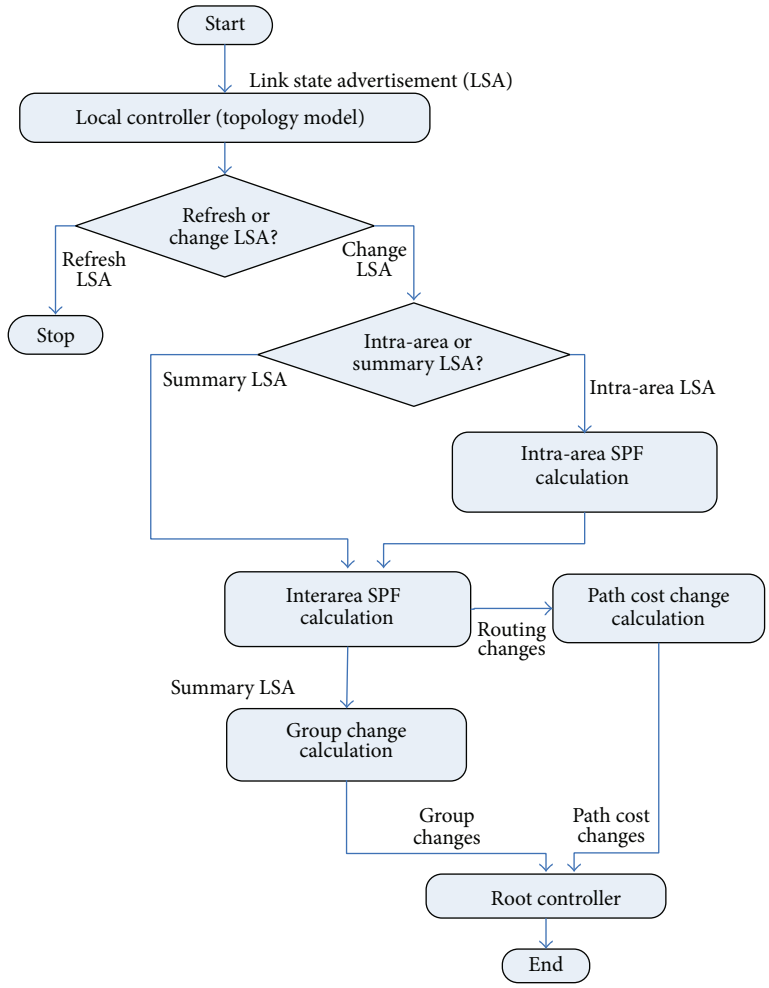


Figure 6: The implementation process of OSPF protocol on local controller.

5.2.4. The Low-Level Exchanges of Managing the Router. Local controller creates simple and stable communication with other routers through a number of flow tables to maintain the BGP sessions and multiplexes the updating message to form a single flow to send to the root controller.

6. Simulation

In order to evaluate the performance of BGP under the Kandoo architecture, we simulate the Kandoo topology using Cisco GNS3. We, respectively, simulate 5 root controllers (or AS), 30 root controllers, 45 root controllers, and 60 root controllers based on the Kandoo architecture. And each root controller connects 6 local controllers. We compare the convergence performance and the numbers of updating message of the Kandoo-BGP (K-BGP) with the traditional BGP (T-BGP). The results are shown in Figures 7 and 8.

From the simulation results of Figure 7, we can see that the numbers of updating message of Kandoo-BGP are less than traditional BGP based on difference topology scale; in particular, when the topology scale is larger and larger, the numbers of updating message significantly decrease.

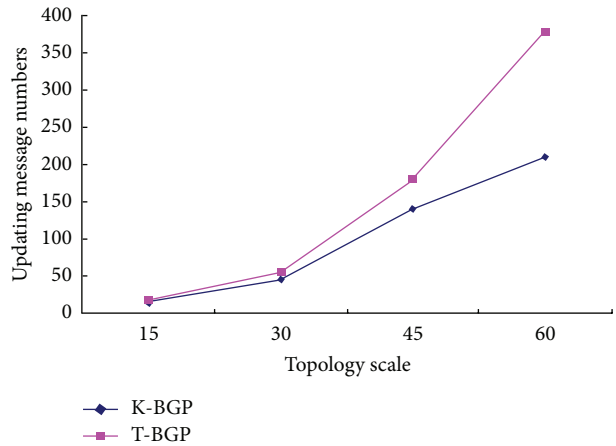


Figure 7: Comparison of numbers of updating message based on difference topology scale.

From the simulation results of Figure 8, we can see that the convergence performance of Kandoo-BGP is superior to traditional BGP based on difference topology scale; in

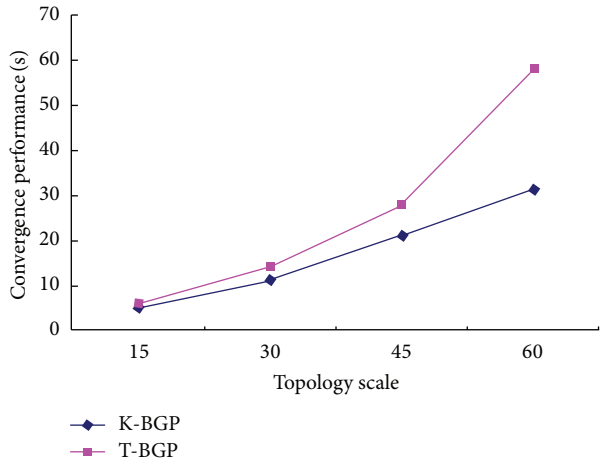


Figure 8: Convergence performance comparison based on difference topology scale.

particular, when the topology scale becomes larger, the convergence performance of K-BGP significantly increases.

7. Summary

This paper mainly researches the deployment of routing protocols in distributed control plane under SDN. The distributed characteristics of Kandoo are studied deeply, which is achieved by multiple controller units interconnected to form a distributed control plane architecture of multicontrol unit. We improved and optimized Kandoo's two levels of controllers based on ideological inspiration of RCP and analyzed the implementation and deployment of BGP and OSPF protocol in a distributed control plane of SDN. We give the simulation results, which show that our deployment strategies are superior to the traditional routing strategies.

Although the deployment strategies have achieved the desired goal, because of the constraints of time and the objective conditions, there are still some deficiencies in this paper. The future works carried out are (1) continuing to deepen the study of a variety of control architectures of SDN and extending the function of Kandoo so that it can deploy and implement efficiently and (2) researching the interdomain deployment strategy of BGP protocol combined with the distributed control plane architecture of SDN and making BGP to fully implement distributed characteristics.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by a Grant from the National Basic Research Program of China (973 Program) (no. 2012CB315902), the National Natural Science Foundation of China (nos. 61379120, 61170215, and 61102074), Zhejiang Leading Team of Science and Technology Innovation (nos. 2011R50010-16, 2011R50010-20), Zhejiang Provincial Key Laboratory of New Network Standards and Technologies (NNST) (no. 2013E10012), and the Natural Science Foundation of Zhejiang (Q12F020074).

References

[1] C. Li, W. Wang, and S. Zhang, "Forwarding performance model in ForCES system under FBM-based traffic arrivals," *Journal of Computer*, vol. 24, no. 1, pp. 46–55, 2013.

[2] M. M. Rashidi, E. Erfani, and B. Rostami, "Optimal homotopy asymptotic method for solving viscous flow through expanding or contracting gaps with permeable walls," *Transaction on IoT and Cloud Computing*, vol. 2, no. 1, pp. 76–100, 2014.

[3] H. Huei-Chen, "A knowledge sharing simulation of team learning," *Transaction on IoT and Cloud Computing*, vol. 1, no. 1, pp. 26–38, 2013.

[4] W. M. Wang, L. G. Dong, and B. Zhuge, "ForTER—an open programmable router based on forwarding and control element separation," *DCABES*, vol. 2, pp. 1069–1077, 2006.

[5] Q. Fang, W. Wang, and X. Wu, "How to improve the independent ability of ForCES routers," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 3, 2013.

[6] S. Hassas Yeganeh and Y. Ganjali, "Kandoo: a framework for efficient and scalable offloading of control applications," in *Proceeding of the 1st ACM International Workshop on Hot Topics in Software Defined Networks (HotSDN '12)*, pp. 19–24, New York, NY, USA, August 2012.

[7] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe, "Design and implementation of a routing control platform," in *Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation (NSDI '05)*, vol. 2, pp. 15–28, 2005.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 429041, 9 pages
http://dx.doi.org/10.1155/2014/429041

Research Article

Research on the Trajectory Model for ZY-3

Yifu Chen and Zhong Xie

National Engineering Research Center for Geographic Information System, China University of Geosciences, 388 Lumo Road, Wuhan 0086-430074, China

Correspondence should be addressed to Yifu Chen; yifuchenyf@gmail.com

Received 11 March 2014; Accepted 18 June 2014; Published 27 August 2014

Academic Editor: Zhongmei Zhou

The new generation Chinese high-resolution three-line stereo-mapping satellite Ziyuan 3 (ZY-3) is equipped with three sensors (nadir, backward, and forward views). Its objective is to manufacture the 1 : 50000 topographic map and revise and update the 1 : 25000 topographic map. For the push-broom satellite, the interpolation accuracy of orbit and attitude determines directly the satellite's stereo-mapping accuracy and the position accuracy without ground control point. In this study, a new trajectory model is proposed for ZY-3 in this paper, according to researching and analyzing the orbit and attitude of ZY-3. Using the trajectory data set, the correction and accuracy of the new proposed trajectory are validated and compared with the other models, polynomial model (LPM), piecewise polynomial model (PPM), and Lagrange cubic polynomial model (LCPM). Meanwhile, the differential equation is derivate for the bundle block adjustment. Finally, the correction and practicability of piece-point with weight polynomial model for ZY-3 satellite are validated according to the experiment of geometric correction using the ZY-3 image and orbit and attitude data.

1. Introduction

Most high-resolution remote-sensing satellites are the near polar satellite; these satellites generally run on their trajectory below 1000 km in order to acquire the higher resolution for earth observation [1]. Ziyuan 3 (ZY-3) is the first Chinese civilian high-resolution stereo-mapping satellite that is equipped with three-line sensors (nadir, backward, and forward views) which have the separated optic system, respectively, and an additional multispectral sensor. The resolutions of nadir, backward, and forward views, are 2.5 m, 4.0 m, and 4.0 m, respectively, and the resolution of the multispectral sensor is 8 m [2]. The trajectory height of ZY-3 is relatively lower, 505 km. The satellite therefore is easily impacted by various disturbing forces from space and various flutters and jitters from the internal mechanical motion of the satellite such as high-frequency flutter from Gyro-Star and flywheel and the low-frequency jitter from solar panels. All of these factors result in the high-frequency flutter and low-frequency jitter of satellite when it is running on its trajectory [3].

For the linear push-broom satellite, every acquired image line has different data of orbit and attitude, and the instrument just records the data at regular intervals, but not all. The unrecorded data at a certain time therefore needs to be interpolated using exterior orientation model [4, 5]. For the traditional bundle block adjustment, it is almost impossible to acquire the solution using the orbit and attitude data of every image line. Thus, the high-precision exterior orientation model is needed to be researched and proposed, which is crucial for the geometry-data processing of linear push-broom satellite. In the geometry-data processing with block bundle adjustment, the difficult and key problem is how to reduce and eliminate the correlation between the interior and exterior orientation model parameters and improve the interpolation's accuracy of orbit and attitude, which avoid the transmission of the interpolation error to the interior orientation model in order to improve the solved accuracy of parameters [6, 7]. The interior and exterior model parameters will affect each other and make the cross-correlations in the calculation with bundle block adjustment. When the correlation among the parameters is strong, the systematic error cannot be described completely and accurately with these parameters, which result in the unstable oscillation of solved parameters and the decreased solution accuracy. In this process, the interpolation error of exterior elements

(orbit and attitude) with trajectory model will also be transmitted to interior orientation as a part of systematic error, increase the systematic error of interior orientation, and decrease the solved accuracy and stability of geometric data processing. An ideal trajectory model not only can ensure a high interpolation accuracy for the attitude and orbit of every image line but also can decrease and eliminate the strong correlations among the solved parameters in bundle block adjustment and reduce the transmission of systematic error between the interior and exterior orientation models.

Generally, the satellite running trajectory is relatively stable in a short period so that the orbit and attitude in a short interval trajectory can be modeled with the polynomial, therefore avoiding the complex stress analysis of the satellite [8], therefore avoiding the complex stress analysis of the satellite. Currently, the trajectory models used for high-resolution remote-sensing satellite have polynomial model (LPM), piecewise polynomial model (PPM), and Lagrange cubic polynomial model proposed by Hofmann (LCPM) [9–11]. By research and comparison with the trajectory models, however, these models have some limitations. The LPM and PPM can acquire the smooth fitted curves, and the interpolation accuracies are very low especially for the unstably oscillated trajectory data. The LCPM can acquire a higher interpolation accuracy comparing with the LPM and PPM; however the LCPM easily causes the data oscillation, when the higher-order model is utilized in the block adjustment.

For satellite ZY-3, a new trajectory model (piece-point polynomial with weight model) is proposed to acquire the higher interpolation's accuracy in this paper, based on analyzing the orbit and attitude data of ZY-3. In addition, the data set of ZY-3 is used to validate the correction of piece-point with weight polynomial model and compare it with the other trajectory models, LPM, PPM, and LCPM, used for other satellites. Meanwhile, the differentiation equation of the proposed trajectory model is derivate and it is validated with the block bundle adjustment. According to geometric correction experiment with the different ground control points, the correctness and applicability of piece-point with weight polynomial model are validated and assessed to ensure and improve the high accuracy of geometric correction for ZY-3 satellite.

2. ZY-3 Trajectory Model

2.1. Piece-Point with Weight Polynomial Model. Trajectory model is a mathematic relationship elucidating the orbit and attitude of satellite vary with the different time in its track. For the push-broom satellite, every acquired image line has a different data of orbit and attitude, and the instrument just records the data at regular intervals, but not all the data. The unrecorded data at a certain time thereby needs to be interpolated using exterior orientation model.

Piece-point with weight polynomial model (PWPM) is proposed according to the researching and analyzing of the data of ZY-3's orbit and attitude in the long and short period. In comparison with the other models, LPM, PPM, and LCPM, the weight value is used in the new model to perform interpolation's calculation. The new model therefore can acquire a higher accuracy, have a better flexibility, and can reduce the correlation among the exterior orientation elements. The PWPM is an interpolation model with weight value. Using the model to interpolate satellite's orbit and attitude, the weight is calculated by the difference from any interpolation's time to the time assumed as known. Through the least square method, the polynomial parameters are solved, and then the exterior orientation at any time on the orbit can be acquired with the polynomial parameters. The PWPM is represented by (1), and the weight value is described with P:

$$
\begin{aligned}
X_{St} &= X_0^i + X_1^i \cdot t + X_2^i \cdot t^2, \\
Y_{St} &= Y_0^i + Y_1^i \cdot t + Y_2^i \cdot t^2, \\
Z_{St} &= Z_0^i + Z_1^i \cdot t + Z_2^i \cdot t^2, \\
\phi_t &= \phi_0^i + \phi_1^i \cdot t + \phi_2^i \cdot t^2, \\
\omega_t &= \omega_0^i + \omega_1^i \cdot t + \omega_2^i \cdot t^2, \\
\kappa_t &= \kappa_0^i + \kappa_1^i \cdot t + \kappa_2^i \cdot t^2,
\end{aligned}
\qquad P = \frac{1}{\|t - t_i\|} \quad P = \frac{1}{(t - t_i)^2} \tag{1}
$$

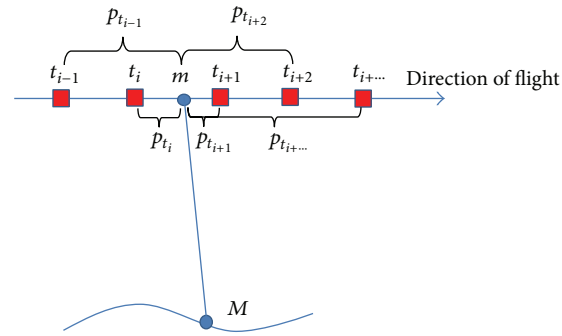


Figure 1: Diagram of piece-point with weight polynomial model.

Figure 1 is the diagram of PWPM. M is a ground point. Its imaging point is shown by m. t is the time that the point m is imaged. On the orbit, t_{i-1}, t_i, t_{i+1}, t_{i+2}, and $t_{i+\cdots}$ are the known times that the exterior orientations are acquired with GPS and star sensor, which are represented by red square. p_{i-1}, p_i, p_{i+1}, p_{i+2}, and $p_{i+\cdots}$, respectively, represent the weight of the time$t$ to the known time t_{i-1}, t_i, t_{i+1}, t_{i+2}, and $t_{i+\cdots}$. The unrecorded orbit and attitude at every time like time t can be interpolated with the exterior orientations at the time t_{i-1}, t_i, t_{i+1}, t_{i+2}, and $t_{i+\cdots}$ using PWPM.

Assuming that the four known times t_1, t_2, t_3 and t_4 for orbit and attitude are selected as known points, the value of orbit and attitude at time t can be calculated and acquired with PWPM. The same polynomial coefficients are solved in the piece-wise polynomial model for the orbit and attitude respectively. For the convenient expression, the error

equation is represented only with the angle Kappa as (3) according to (1) and the general form of error

$$V = A \cdot X - L \tag{2}$$

$$v_i = \kappa_0^i + \kappa_1^i \cdot t + \kappa_2^i \cdot t^2 - \kappa_t \quad (i = 1, 2, 3, 4). \tag{3}$$

Due to the four known times, the four error equations can be established and the coefficient matrix A, observed value matrix L, and unknown vector matrix X are constructed by error equations. Matrices A, L, and X are represented by

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \end{bmatrix}; \qquad L = \begin{bmatrix} \kappa_1 \\ \kappa_2 \\ \kappa_3 \\ \kappa_4 \end{bmatrix}; \qquad X = \begin{bmatrix} k_0 \\ k_1 \\ k_2 \end{bmatrix}. \tag{4}$$

According to the equation of weight, the weights p_1, p_2, p_3, and p_4 for times t_1, t_2, t_3, and t_4, respectively, are calculated and shown in (5), and then the weight matrix P is established. The value of weight reflects the influence degree of the orbit and attitude at times t_1, t_2, t_3, and t_4 for the unknown time t:

$$\begin{aligned} p_1 &= \frac{1}{(t - t_1)^2}; \qquad p_2 = \frac{1}{(t - t_2)^2}; \\ p_3 &= \frac{1}{(t - t_3)^2}; \qquad p_4 = \frac{1}{(t - t_4)^2} \end{aligned} \tag{5}$$

$$P = \begin{bmatrix} p_1 & & & \\ & p_2 & & \\ & & p_3 & \\ & & & p_4 \end{bmatrix}. \tag{6}$$

According to the least square method, normal equation coefficient matrix N and free vector U are constructed and shown by

$$\begin{aligned} N &= A^T P A \\ &= \begin{bmatrix} p_1 + \cdots + p_4 & p_1 t_1 + \cdots + p_4 t_4 & p_1 t_1^2 + \cdots + p_4 t_4^2 \\ p_1 t_1 + \cdots + p_4 t_4 & p_1 t_1^2 + \cdots + p_4 t_4^2 & p_1 t_1^3 + \cdots + p_4 t_4^3 \\ p_1 t_1^2 + \cdots + p_4 t_4^2 & p_1 t_1^3 + \cdots + p_4 t_4^3 & p_1 t_1^4 + \cdots + p_4 t_4^4 \end{bmatrix} \\ U &= \begin{bmatrix} \kappa_1 \cdot p_1 + \cdots + \kappa_4 \cdot p_4 \\ \kappa_1 \cdot t_1 \cdot p_1 + \cdots + \kappa_4 \cdot t_4 \cdot p_4 \\ \kappa_1 \cdot t_1^2 \cdot p_1 + \cdots + \kappa_4 \cdot t_4^2 \cdot p_4 \end{bmatrix}. \end{aligned} \tag{7}$$

For convenience, the matrices N^{-1} and U are represented by the below formation, shown in

$$N^{-1} = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix} \qquad U = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}. \tag{8}$$

Through least square adjustment, the polynomial parameters k_0, k_1, and k_2 are solved and shown by

$$X = \begin{bmatrix} k_0 \\ k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}. \tag{9}$$

Afterwards, the Kappa value at t time can be calculated using (1), and the general formula of the Kappa value is represented by

$$\begin{aligned} \kappa(t) &= (c_1 u_1 + c_2 u_2 + c_3 u_3) \\ &\quad + (c_4 u_1 + c_5 u_2 + c_6 u_3)\, t \\ &\quad + (c_7 u_1 + c_8 u_2 + c_9 u_3)\, t^2. \end{aligned} \tag{10}$$

In the process of interpolation, the PWPM can solve the different parameters corresponding to the different attitude and orbit at any time with the different weight value. In this paper, the two weight equations are given out, the reciprocal of the absolute value of time difference and the reciprocal of the square of the time difference. The selection of weight equation has a large impact for the interpolation's accuracy of orbit and attitude. According to analysis and research, the reciprocal of the square of the time difference is adopted when the trajectory is relatively unstable. On the contrary, the reciprocal of the absolute value of time difference is utilized.

2.2. Two Interpolation Methods of Trajectory Model. For the PWPM, the new trajectory has two kinds of interpolation methods to acquire the data set of orbit and attitude at any time. One is using the several known times selected round the unknown time to interpolate the orbit and attitude at the unknown time. The other is using all the selected known times to interpolate the orbit and attitude at the unknown time. The impaction from the selected known time for the orbit and attitude at the unknown time is measured and assessed according to the weight value. In other words, the time difference between the unknown time and the known time is more far, and the impaction from the unknown time is more great. The interpolation accuracies with the two methods are different, which is determined by the stability of the satellite trajectory, the number of selected known times, and the location of the selected known time. In the practical application, the two methods of PWPM are utilized together or respectively, which is determined by the stability of orbit and attitude.

2.3. The Differential Expression of PWPM. Sensor's imaging model describes the mathematic transformation relationship between the coordinate of image point (x, y) and the coordinate of ground point (X, Y, Z). It includes two kinds of models, rigorous imaging model and general imaging model [12, 13]. Based on the structure of CCD-array equipped on the ZY-3 and the sight vector of every CCD, scanning the ground, the rigorous imaging model for ZY-3 satellite is established, represented by (11), and the one to one correspondence relationship between image point and ground point is built

up with sensors coordinate systems, satellite's trajectory coordinate system, and ground reference system.

For ZY-3 satellite, the data received from dual-frequency GPS represents the location of the phase center of GPS and the attitude data from star sensor is measured in the J2000 coordinate [14, 15]. In the process, the displacement matrix from the phase center of GPS (GPS antenna) to the coordinate of satellite body, $[Dx\ \ Dy\ \ Dz]^T$; the displacement matrix from CCD-array center to the coordinate of satellite body, $[dx\ \ dy\ \ dz]^T$; the rotation matrix from the coordinate of star sensor to the coordinate of satellite body, $R_{\text{star}}^{\text{body}}$; and the rotation matrix form imaging space coordinate to the coordinate satellite body, $R_{\text{camera}}^{\text{body}}$, are needed:

$$\begin{aligned}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{WGS}_{84}} &= m \cdot R \cdot \left[\begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} + R_{\text{camera}}^{\text{body}} \cdot \begin{pmatrix} x \\ y \\ -f \end{pmatrix} \right] \\ &\quad + \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{GPS}} \\ R &= R_{\text{J2000}}^{\text{WGS84}} \cdot R_{\text{orbit}}^{\text{J2000}} \cdot R_{\text{body}}^{\text{orbit}},\end{aligned} \tag{11}$$

where $[x\ \ y\ \ -f]^T$ are point coordinates in the image system; $[X\ \ Y\ \ Z]_{\text{GPS}}^T$ and $[X\ \ Y\ \ Z]_{\text{WGS}_{84}}^T$ are perspective center position and position coordinates in WGS84 coordinate system; $R_{\text{body}}^{\text{orbit}}$, $R_{\text{orbit}}^{\text{J2000}}$, and $R_{\text{orbit}}^{\text{J2000}}$ are rotation matrices, respectively, from satellite body system to satellite orbit system, from satellite orbit system to J2000 coordinate system, and from J2000 coordinate system to WGS84 coordinate system; m represents the scale. According to satellite's structure design and the result of laboratory calibration, the three-line imaging model, nadir, forward, and backward, can be acquired with the different value of displacement matrix $[dx\ \ dy\ \ dz]^T$ and rotation matrix $R_{\text{camera}}^{\text{body}}$.

For the push-broom high-resolution satellite, the objective of the high-accuracy trajectory model is to acquire the accurate elements of exterior orientation, $[X_{\text{GPS}}, Y_{\text{GPS}}, Z_{\text{GPS}}]^T$ and $R_{\text{body}}^{\text{orbit}}$, at any time.

Assuming that the n known times are selected in a scene image of satellite and the orbit and attitude at time t are needed to be interpolated, the four known times t_1, t_2, t_3, and t_4 round the unknown time t are selected and the their weight values p_1, p_2, p_3, and p_4 are correspondingly calculated with (1). For the convenient expression, the Pitch angle is picked up as a sample; therefore, the value of Pitch angle at t time can be calculated and represented by (12), which has the same form expression as (10):

$$\begin{aligned}\text{Pitch}(t) = {} & (c_1u_1 + c_2u_2 + c_3u_3) \\ & + (c_4u_1 + c_5u_2 + c_6u_3)\, t \\ & + (c_7u_1 + c_8u_2 + c_9u_3)\, t^2.\end{aligned} \tag{12}$$

In the calculation of bundle block adjustment, the differential expression of t time for the four unknown times t_1, t_2, t_3, and t_4 is derived and represented by

$$\begin{aligned}\frac{\partial \text{Pitch}(t)}{\partial \text{Pitch}(t_1)} = {} & \left(c_1 + c_4t + c_7t^2\right) p_1 \\ & + \left(c_2 + c_5t + c_8t^2\right) p_1t_1 \\ & + \left(c_3 + c_6t + c_9t^2\right) p_1t_1^2 \\ \frac{\partial \text{Pitch}(t)}{\partial \text{Pitch}(t_2)} = {} & \left(c_1 + c_4t + c_7t^2\right) p_2 \\ & + \left(c_2 + c_5t + c_8t^2\right) p_2t_2 \\ & + \left(c_3 + c_6t + c_9t^2\right) p_2t_2^2 \\ \frac{\partial \text{Pitch}(t)}{\partial \text{Pitch}(t_3)} = {} & \left(c_1 + c_4t + c_7t^2\right) p_3 \\ & + \left(c_2 + c_5t + c_8t^2\right) p_3t_3 \\ & + \left(c_3 + c_6t + c_9t^2\right) p_3t_3^2 \\ \frac{\partial \text{Pitch}(t)}{\partial \text{Pitch}(t_4)} = {} & \left(c_1 + c_4t + c_7t^2\right) p_4 \\ & + \left(c_2 + c_5t + c_8t^2\right) p_4t_4 \\ & + \left(c_3 + c_6t + c_9t^2\right) p_4t_4^2.\end{aligned} \tag{13}$$

Similarly, the differential expression of the other elements of exterior orientation, (Roll, Yaw) and (X_s, Y_s, Z_s), can be derived.

3. Systematic Error Model

The systematic error model for the interior orientation is to describe the various distortions from satellite's sensor such as the CCD-array distortions, the distortions of optic lenses, and principal point's distortion. In order to realize the high-precision geometric correction for ZY-3 image, it is very necessary to establish the various error models based on the analysis of satellite's structural parameters; thus the system error coming from the interior orientation, radial direction, and tangential direction distortion of optics lens and CCD-line's distortion and rotation will be modeled [16, 17]. According to the analysis of satellite's structure and the imaging characteristics, the systematic error model of the interior orientation is established and represented by

$$\begin{aligned}\Delta x = {} & -\frac{\Delta f}{f}\overline{x} + \left(k_1r^2 + k_2r^4\right)\overline{x} \\ & + p_1\left(r^2 + 2\overline{x}^2\right) + 2p_2\overline{xy} + \overline{y}\sin\theta\end{aligned}$$

$$\Delta y = -\frac{\Delta f}{f}\overline{y} + \left(k_1 r^2 + k_2 r^4\right)\overline{y} + p_2\left(r^2 + 2\overline{y}^2\right) + 2p_1\overline{x}\overline{y} + s_y\overline{y}$$
$$\overline{x} = (x - x_0), \quad \overline{y} = (y - y_0), \quad r = \sqrt{\overline{x}^2 + \overline{y}^2}, \tag{14}$$

where $(-\Delta f/f)\overline{x}$ and $(-\Delta f/f)\overline{y}$ represent the errors generated by the image principal point and focal length; Δf and f mean the difference of focal length and the optic focal length, respectively; $(k_1 r^2 + k_2 r^4)\overline{x}$ and $(k_1 r^2 + k_2 r^4)\overline{y}$ describe the optics lens distortion of radial direction in along-track and cross-track directions, respectively; k_1, k_2 mean the distortion's coefficient of radial direction; r means the distance from one point on the optic lens to the lens's center; $p_1(r^2 + 2\overline{x}^2) + 2p_2\overline{x}\overline{y}$ and $p_2(r^2 + 2\overline{y}^2)$ represent the distortions of tangential direction in along-track and cross-track directions, respectively; p_1, p_2 mean the distortion's coefficient of tangential direction; $\overline{y}\sin\theta$ represents the error of CC-array rotation, and θ is the rotation angle; $s_y\overline{y}$ represents the distortion of CCD-array in the cross-track direction generated by the temperature variation. The CCD-array distortion in along-track direction is particle, owing to only one CCD arranged in this direction; the distortion therefore can be ignored.

According to the analysis of the correlations among the model's parameters in the block bundle adjustment, the correlation between the principal point and focal length is very strong, so that the parameters are combined in order to reduce the parameters correlation and improve the stability and accuracy of the block bundle adjustment. Equation (14) is represented as (15) after the parameters combination:

$$\Delta x = x_0 + \left(k_1 r^2 + k_2 r^4\right)\overline{x} + p_1\left(r^2 + 2\overline{x}^2\right) + 2p_2\overline{x}\overline{y} + \overline{y}\sin\theta$$
$$\Delta y = y_0 + \left(k_1 r^2 + k_2 r^4\right)\overline{y} + p_2\left(r^2 + 2\overline{y}^2\right) + 2p_1\overline{x}\overline{y} + s_y\overline{y} \tag{15}$$
$$\overline{x} = (x - x_0), \quad \overline{y} = (y - y_0), \quad r = \sqrt{\overline{x}^2 + \overline{y}^2}.$$

4. Data and Method of Experiment

In this paper, data set of orbit and attitude used to validate the correction and accuracy of piece-point polynomial model is acquired from 609th track of ZY-3. In order to validate the high accuracy of the new proposed trajectory model, the LPM, PPM, LCPM, piece-point with weight polynomial with four known times model (PWP4M) and piece-point with weight polynomial model with all known times (PWPM) are utilized to interpolate and compare the interpolation accuracy. In the process, the different numbers of the known times, 10, 15 and 20, are selected from trajectory data set, and are used to interpolate the other unknown times' orbit and attitude data with the different models, respectively. Finally, the result of interpolation is represented by the table and curve, and the advantage of PWPM is illuminated according to researching and analyzing the result.

In order to validate the correction and accuracy of PWPM, ZY-3 orbit, and attitude data, ground control point (GCP) and systematic error model of interior orientation are used in the bundle block adjustment of geometric correction. Based on the nadir image of ZY-3, the 74 GCPs are picked up from the image, and 27 GCPs are selected as check points (CPs) that do not take part in the block bundle adjustment. For validating the correction and stability of the proposed models, the 16, 26, 36, and 46 GCPs are performed, respectively, in the geometric correction experiment. Figure 2 shows the distribution of GCPs and the error's distribution of image points, corresponding to GCPs, before geometric correction.

5. Result and Validation of Experiment

5.1. Result and Analysis of Trajectory Model. According to analyzing the stability of the orbit and attitude of ZY-3, it can be seen obviously that the orbit and attitude angles of Yaw of ZY-3 are very stable, but the attitude angles of Pitch and Roll are unstable relatively. The curves of attitude angles in 10 seconds are shown in Figures 3, 4, and 5. From the diagram of curves, it is obvious that the attitude angles of Pitch and Roll are unstable. Hence, the interpolation experiment is performed using the angle Pitch and Roll.

In Figure 6, the result of interpolation is represented using the different trajectory models with 15 selected known times in a scene image. The curves on the left of Figure 6 show the interpolation's result with angle Pitch, and the curves on the right of Figure 6 show the result with angle Yaw. The red curve and red circle, respectively, mean the fitting curve and the selected known times, and the green curve means the original curves. From top to bottom Figures 6(a), 6(b), 6(c), 6(d), and 6(e) represented, respectively, the models LPM, PPM, LCPM, PWP4M, and PWPM.

From Figure 6, it can be seen clearly that the interpolation's accuracy with LCPM, PWP4M, and PWPM is higher than LPM and PPM. The curves of LCPM, PWP4M, and PWPM are relatively similar. Furthermore, Table 1 shows the interpolation accuracy results with the different orbit and attitude models using the different numbers of known times selected from 609th track orbit in a scene image. The accuracy of the interpolation for the angles Roll, Pitch, and Yaw with PWPM is the highest with 10 and 15 selected known times. Using the 20 known times, the interpolation accuracy of Yaw is the highest with PWP4M, 4.539. The accuracy of Pitch with PWP4M is 5.727 lower than the accuracies with PWPM 5.593. In this case, PWP4M and PWPM can be used together in order to acquire the highest accuracy of interpolation for any angle.

5.2. Result and Analysis of Geometric Correction Based on PWPM. The new proposed trajectory model (PWPM) has higher interpolation's accuracy and more flexibility than the other models according to the upper experiment and analysis.

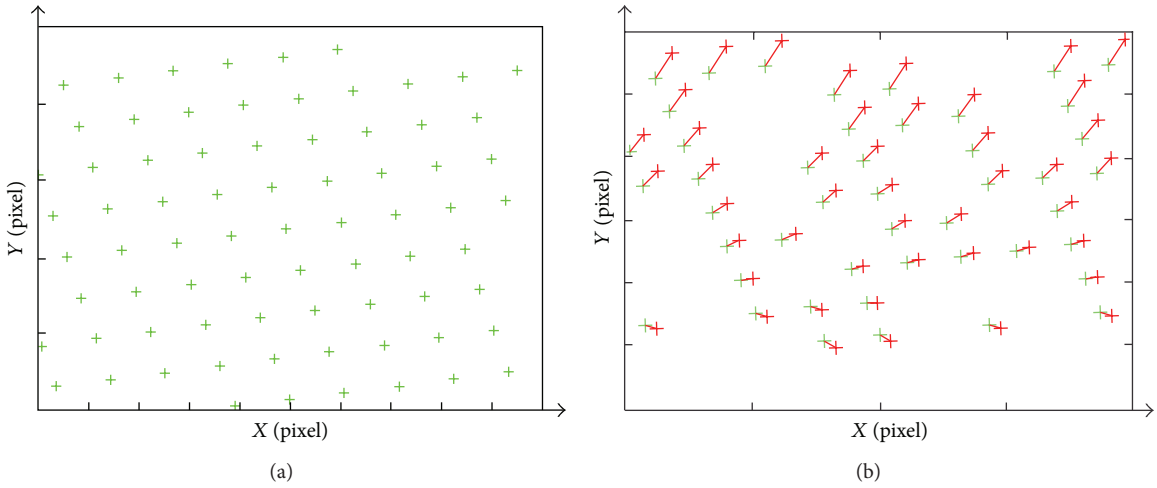


Figure 2: (a) Diagram of the distribution of GCPs; (b) the error's distribution of image points corresponding to GCPs.

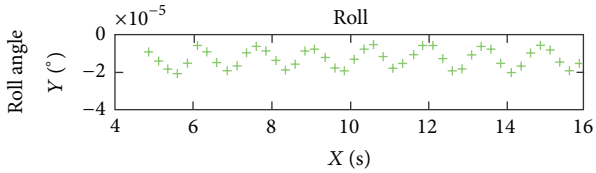


Figure 3: The attitude angle (Roll) curve in 10 seconds.

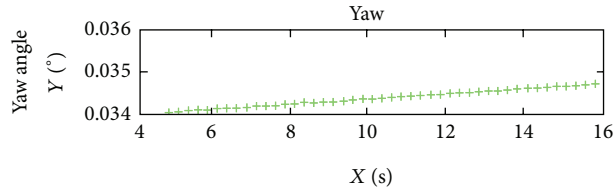


Figure 5: The attitude angle (Yaw) curve in 10 seconds.

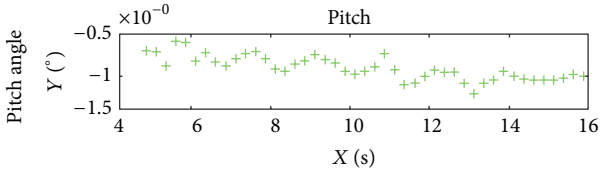


Figure 4: The attitude angle (Pitch) curve in 10 seconds.

In order to validate the correction and accuracy of PWPM in the bundle block adjustment, the geometric correction experiment is performed using the data set of ZY-3. Before the process, which one interpolation's method of PWPM is utilized according to the analysis of the orbit and attitude corresponding to the used image range? Thus, geometric correction is performed and the result of correction is represented by Figure 7.

In Figure 7(a), the residuals distribution of 46 GCPs after the geometric correction is represented and the residuals distribution after checking with 27 CPs is shown in Figure 7(b) and the assessed accuracy is 0.0793 pixels. From Figure 7, it is obvious that the accuracy of geometric correction is very high based on the PWPM. Furthermore, Table 2 shows the assessment of accuracy for geometric correction using the 10, 16, 26, 36, and 46 GCPs, respectively. The accuracies with the different number of GCPs are all high; the highest accuracy reaches 0.5293 pixels with 26 GCPs and the lowest accuracy is 0.0841 pixels. With the increasing number of GCPs, the accuracy of geometric correction increased gradually until the 26 GCPs.

5.3. Validation and Analysis. Analyzing and comparing Table 1 and Figure 6, it is obvious that the interpolation accuracy of PWPM is the highest. When the orbit and attitude are unstable, the weight value is acquired with the reciprocal value of the absolute value of time difference. On the contrary, the weight value is calculated with the reciprocal value of the square of time difference. In comparison with the LPM, PPM, and LCPM, PWPM can solve the different parameters corresponding to the different attitude at any time with the different weight value. The proposed new trajectory model can therefore reach higher accuracy of interpolation than others, especially when the orbit and attitude are unstable. In addition, the PWPM has two interpolation methods, PWP4M and PWPM, and the two methods can be used together in order to acquire the higher interpolation accuracy. Owing to the higher accuracy, the new trajectory model has the ability that can improve the interpolation accuracy of orbit and attitude and avoid the interpolation error is transmitted into interior orientation as a part of systematic error, which will increase the systematic error of interior orientation. Thus, the solved accuracy of parameters in the systematic model is improved, and the accuracy of geometric correction is also increased correspondingly. According to the analysis and research, the different form of PWPM in the bundle block adjustment only relates to weight, selected

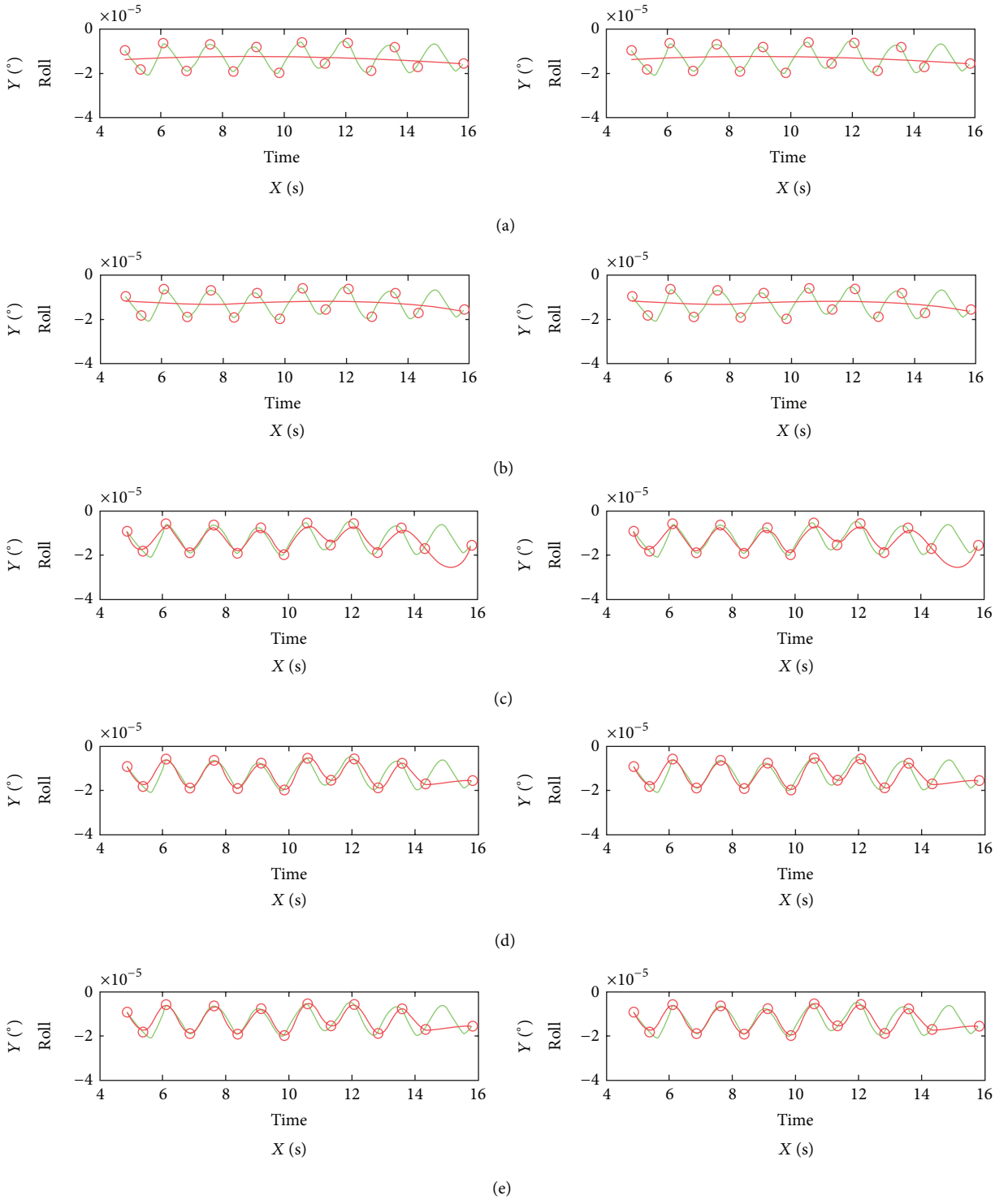


Figure 6: The fitting curves with the different trajectory models.

known time, and needed interpolation time; thus the correlation among the orbit and attitude is decreased, and the calculation's accuracy and stability of the block adjustment can be improved.

In the geometric correction experiment based on the PWPM, the accuracies with the different number of GCPs also reach a high level totally, which is represented by Table 2 and Figure 7, and validate the correctness and applicability of the PWPM. For the different number of GCPs, the accuracy varies in the geometric correction mostly owing to two reasons, the distribution of the different GCPs and the correlations among the parameters of the systematic

Table 1: The fitting accuracy comparison of the different attitude and orbit modes selecting the different known times on the orbit (unit: degree).

σ ($e-007$)	10			15			20		
Angle	Roll	Pitch	Yaw	Roll	Pitch	Yaw	Roll	Pitch	Yaw
1-LPM	64.610	12.772	15.100	47.859	9.103	15.297	47.362	8.569	15.122
2-PPM	63.136	12.237	16.210	47.768	8.737	15.027	47.281	8.532	14.948
3-LCPM	63.523	11.955	13.919	51.758	8.275	10.257	45.901	5.946	9.554
4-PWP4M	63.668	11.732	13.956	34.760	8.034	7.7186	15.628	5.727	4.539
5-PWPM	60.287	10.387	12.534	32.837	7.648	7.5958	22.261	5.593	5.140

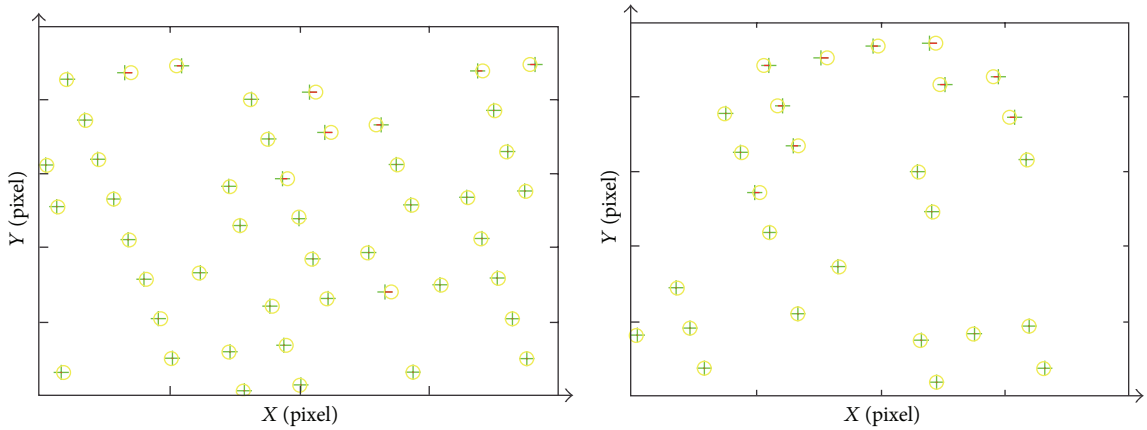


Figure 7: Diagram of geometric correction: (a) the residuals of GCP; (b) assessment with CP.

Table 2: The assessment of accuracy for geometric correction (unit: pixel).

Number of GCPs	σ_x	σ_y	σ_{sum}
46	0.0767	0.0215	0.0797
36	0.0706	0.0217	0.0739
26	0.0508	0.0146	0.0529
16	0.0725	0.0161	0.0743
10	0.0795	0.0276	0.0841

error model. On the one hand, the various distributions of the different GCPs will cause the accuracy to oscillate in a very small range; on the other hand, the bundle block adjustment will generate correlations among the parameters. The correlations result in the following: the solved results of parameters vibrate in a range unstably, and the systematic error cannot be described completely and accurately with these parameters. Thus, a better systematic error model is needed to be proposed according to further researching and analyzing of the satellite sensor's overall structure design and the imaging geometric characteristics.

6. Conclusion

In this study, the new trajectory model, PWPM, is proposed according to the researching and analyzing of the data of ZY-3's orbit and attitude in the long and short period. By comparison with the other trajectory models, the PWPM can acquire a higher interpolation's accuracy and has more flexibility. Meanwhile, the differentiation equation of the proposed trajectory model is derivate and it is validated through the bundle block adjustment. In the geometric correction experiment based on the PWPM, the accuracies of geometric correction with the different number of GCPs also reach a high level totally. According to the analyzing and researching of the assessment results with GCPs and CPs, the correctness and applicability of the PWPM are validated and assessed to ensure and improve the high accuracy of geometric correction for ZY-3 satellite. The further study will be performed to experiment with the real image data of ZY-3 and GCP to research better systematic error model for interior orientation, in order to explore the potentials of using ZY-3 data for stereo mapping.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] F. Long, W. Zhang, and J. Liu, "Effect of satellite attitude control accuracy on TDI CCD cameras," *Journal of Harbin Institute of Technology*, vol. 34, no. 3, pp. 382–384, 2002.

[2] C. Sun, X. Tang, Z. Qiu, and X. Wu, "Introducing ZY-3: China's first civilian high-res stereo mapping satellite," in *Proceedings of the Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS '12)*, Melbourne, Australia, September 2012.

[3] G. Zhang, *Rectification for high resolution remote sensing image under lack of ground control points [Ph.D. thesis]*, Wuhan University, Wuhan, China, 2005.

[4] A. Bouillon, E. Breton, F. De Lussy, and R. Gachet, "SPOT5 geometric image quality," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '03)*, vol. 1, pp. 303–305, Toulouse, France, 2003.

[5] T. Toutin, "Geometric processing of remote sensing images: models, algorithms and methods," *International Journal of Remote Sensing*, vol. 25, no. 10, pp. 1893–1924, 2004.

[6] K. Jacobsen, "Calibration of optical satellite sensors," in *Proceedings of the International Calibration and Orientation Workshop EuroCOW*, Casteldefels, Spain, 2006.

[7] D. Mulawa, "On-orbit geometric calibration of the orb-view3 high-resolution imaging satellite," in *Proceedings of the ISPRS 20th Congress*, Commission 1, Remote Sensing and Spatial Information Sciences, Istanbul, Turkey, July 2004.

[8] X. Li, L. Zhang, and W. Xu, "Precise acquisition of ZY-3 orbit and attitude parameters based on metadata file," *Journal of Atmospheric and Environmental Optics*, vol. 3, no. 8, pp. 166–173, 2013.

[9] F. J. Ponzoni, J. Zullo Jr., R. A. C. Lamparelli, G. Q. Pellegrino, and Y. Arnaud, "In-flight absolute calibration of the Landsat-5 TM on the test site salar de uyuni," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 12, pp. 2761–2766, 2004.

[10] C. Valorge, "40 years of experience with SPOT in-flight calibration," in *Proceedings of the ISPRS Workshop on Radiometric and Geometric Calibration*, Gulfport, Miss, USA, December 2003.

[11] S. Kocaman and A. Gruen, "Orientation and self-calibration of ALOS PRISM imagery," *Photogrammetric Record*, vol. 23, no. 123, pp. 323–340, 2008.

[12] S. Riazanoff, *SPOT 123-4-5 Geometry Handbook*, GAEL Consultant, 2004, http://www-igm.univ-mlv.fr/~riazano/publications/GAEL-P135-DOC-001-01-04.pdf.

[13] X. Zhu, G. Zhang, X. Tang, and L. Zhai, "Research and application of CBRS02B image geometric exterior calibration," *Geography and Geo-Information Science*, vol. 25, no. 3, pp. 16–18, 2009.

[14] G. Zhang, Z. Li, H. Pan, Q. Qiang, and L. Zhai, "Orientation of spaceborne SAR stereo pairs employing the RPC adjustment model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 7, pp. 2782–2792, 2011.

[15] X. Tang, G. Zhang, X. Zhu et al., "Triple linear-array imaging geometry model of Ziyuan-3 surveying satelite and its validation," *Acta Geodaetica et Cartographica Sinica*, vol. 41, no. 2, pp. 191–198, 2012.

[16] D. Poli, *Modelling of spaceborne linear array sensors [Ph.D. thesis]*, Swiss Federal Institute of Technology, Zurich, Switzerland, 2005.

[17] D. Poli, "Indirect georeferencing of airborne multiline array sensors: a simulated case study," in *Proceedings of the ISPRS Commission Symposium, International Archives of Photogrammetry and Remote Sensing*, vol. 34, part B3, pp. 246–251, Graz, Austria, September 2002, part B3.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 432123, 11 pages
http://dx.doi.org/10.1155/2014/432123

Research Article

An Opportunistic Routing Mechanism Combined with Long-Term and Short-Term Metrics for WMN

Weifeng Sun, Haotian Wang, Xianglan Piao, and Tie Qiu

Software School, Dalian University of Technology, Dalian 116620, China

Correspondence should be addressed to Tie Qiu; qiutie@dlut.edu.cn

Received 30 April 2014; Accepted 16 July 2014; Published 26 August 2014

Academic Editor: Xiaoxuan Meng

WMN (wireless mesh network) is a useful wireless multihop network with tremendous research value. The routing strategy decides the performance of network and the quality of transmission. A good routing algorithm will use the whole bandwidth of network and assure the quality of service of traffic. Since the routing metric ETX (expected transmission count) does not assure good quality of wireless links, to improve the routing performance, an opportunistic routing mechanism combined with long-term and short-term metrics for WMN based on OLSR (optimized link state routing) and ETX is proposed in this paper. This mechanism always chooses the highest throughput links to improve the performance of routing over WMN and then reduces the energy consumption of mesh routers. The simulations and analyses show that the opportunistic routing mechanism is better than the mechanism with the metric of ETX.

1. Introduction

Wireless mesh network based on IEEE 802.11 is a wireless multihop network which is easy to deploy, is of low cost, and has wide coverage. WMN is an efficient extension of wired network. It overcomes the limitation of the harsh geographical conditions and provides high-speed transmission. WMN is also the suitable access solution of last mile. However, compared with wired network, the channel of wireless networks cannot provide a guaranteed link, and there also exists asymmetric, immediate loss and transmission error (bit error) in wireless networks [1]. From the research in [2], data packet loss is not caused only by distance and signal attenuation; SNR (Signal-to-Noise Ratio, S/N) and multipath fading with interference are also the important reasons which lead to packet loss. In wireless networks, SNR and BER (Bit Error Ratio) should be taken account of the information of neighbor nodes; thus, the information is relatively unreliable. Traditional link-state and distance vector based routing protocols such as OSPF (Open Shortest Path First Interior Gateway Protocol) [3] and RIP (Routing Information Protocol) [4] do not work very well. According to traditional routing protocols, packet loss which is caused by physical layer will react to the transport layer and lead to the unavailability of users' services.

In [5], Ahmeda and Esseid claimed that each individual routing metric considers some features and it is difficult to satisfy all the requirements of WMNs by using a single metric. So we consider combining two different metrics to avoid the disadvantage of single metric.

Figure 1 is a simple example of the problem of ETX. The data flow starts from node A to node E, and the ETX of links B → C and B → D are 3 and 5, respectively. According to traditional mechanism with the routing metric of ETX, node B will choose node C as the next hop to transmit. If the link of B → C has a lower probability of successful transmission, while the probability of successful delivery over BD is higher at the same time, the link of B → D should be chosen to achieve better performance of transmission.

In this paper, we study the problem of original ETX used in WMN. We consider a wireless mesh network deployed in a simple topology which has three source destination pairs of data which start simultaneously. Two important parameters that are related to the problem are data transmission rate and probe interval of ETX. Particularly, we are interested in the following three specific questions.

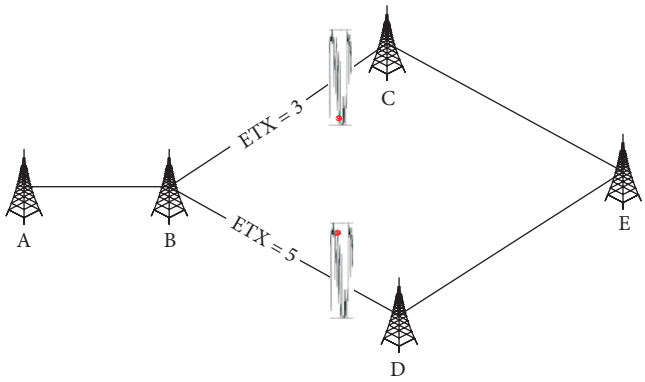


Figure 1: The state of links in data transmission.

(i) What is the optimized probe interval of ETX that leads to maximum throughput of WMN?

(ii) Does the data rate of transmission of WMN effect the optimization of probe interval of ETX?

(iii) Could another metric be added in the original ETX to overcome the shortcoming of original ETX?

Through the cross-layer process between MAC, routing, and application layers, a quality aware routing with opportunity based on ETX is proposed in this paper. This mechanism which is under the assurance of QoS with long-term (longer probe interval for calculating the metric) routing metric (ETX, expected transmission count) concludes a candidate forwarding link set. Then, according to short-term (shorter probe interval for calculating the metric) routing metric (RTT, Round-Trip Time), the link set will be checked and the candidate link with the best wireless link condition will be chosen by this mechanism to transmit packets. Through the combination of long-term and short-term routing metrics, this opportunistic routing mechanism not only chooses the link with smaller ETX to assure the high probability of transmission success, but also guarantees that the transmission over each hop has higher throughput and improves the efficiency of data transmission.

The contribution of our work includes the following aspects:

(i) analyses of the disadvantage of ETX over WMN with different data transmission rates and probe intervals of ETX through experiments of simulation;

(ii) we propose an opportunistic routing mechanism combined with long-term and short-term metrics based on OLSR and ETX to overcome the disadvantage and improve the routing performance;

(iii) analyses of theory and implementation to prove the advantage of the opportunistic routing mechanism;

(iv) simulation results evaluate the performance of the opportunistic routing mechanism, and it performs better than ETX.

The rest of paper is organized as follows: Section 2 introduces the background of ETX, formulates the network problems under study, and reviews related work. The opportunistic routing mechanism is described in Section 3. Section 4 gives the analyses and performance evaluation of the mechanism. Finally, Section 5 summarizes the paper and our future work.

2. Related Work

In this section, we introduce the background of ETX and formulate the network model under study and review related work.

2.1. Background of ETX. The routing metric of WMN includes hop count [6], RTT [7], Per-Hop Packet Pair Delay (PktPair) [8], ETX, expected transmission time (ETT) [9], and Weighted Cumulative Expected Transmission Time (WCETT) [10]. The selected path by routing algorithm may be the one of the farthest, weakest signal and the least energy with the routing metric of hop count. In particular, in dense scenario, the routing algorithm with minimum hop count does not work efficiently and the chosen path has weak signal and low bandwidth and it brings more interference, so a new routing metric in wireless multihop network is needed. According to the research in [11], the results in the test-bed of Roofnet show that the probability of one packet's successful transmission cannot be predicted accurately. So ETX with statistical function according to the Allen variance was designed. ETX calculates the expected transmission count of a packet over a link. ETT, WCETT, and some others are all the routing metrics based on ETX. From the results of tests with variety of routing metrics in [12], the metric of hop count reflects the actual topology state better and achieves better performance ad hoc, while, for WMN which has relatively fixed nodes, ETX performs better.

In ETX, a mesh router delivers the probe packets to its neighbor nodes every 1 second. Through the statistic of successful forwarding count d_f and successful receiving of ACK (acknowledgment) count d_r, EXT calculates the expected count of failure forwarding and receiving P_f and P_r. So the probability of failure transmission is $P = 1 - (1 - P_f) * (1 - P_r)$ and the probability of successful transmission after k times' retransmission is $s(k) = P^{k-1} * (1 - P)$. By summing, the expected successful transmission count is

$$\mathrm{ETX} = \sum_{k=1}^{\infty} k * s(k) = \frac{1}{1-P} = \frac{1}{\left(1-P_f\right) * \left(1-P_r\right)}. \quad (1)$$

ETX calculates the expected transmission count through the statistical results in the latest 10 seconds and updates every 1 second. If the quality and transmission rate of link fluctuate wildly in a certain period of time, it will not work well for this method which predicts current transmission rate according to the historical data in wireless multihop networks.

2.2. Related Work. In this section, we briefly review existing work on the improvement of ETX in WMN.

It will encounter some problem when the metric of ETX is used over WMN in the practical scenario. Some new metrics based on ETX were proposed in several literatures. Through

this kind of method, most of these new metrics could reflect network conditions more accurately. In [13], Koksal and Balakrishnan describe two new metrics, called modified expected number of transmissions (mETX) and effective number of transmissions (ENT) that work well under a wide variety of channel conditions. Empirical observations of a real-world wireless mesh network suggest that mETX and ENT could achieve a 50% reduction in the average packet loss rate compared with ETX.

In order to solve the problem that ETX performs poorly in the case of multiradio and multichannels, WCETT was proposed in [14] to extend ETX and make it support multichannels. Instead of the probe mechanism with single minimum rate, WCETT adopts the probe mechanism with multirates to predict the packet loss rate accurately.

In the implementation of ETX, it acquires the global information firstly and selects the path from the source point to the end point with minimum expected transmission count. This kind of method according to the expected transmission count (including retransmission count) considers the influence of packet loss rate and asymmetry of link in wireless network. Mogaibel et al. experimentally verified ETX's high-speed transmission rate, and the performance was improved most especially on single-channel WMN in [12].

Also, the metric of ETX was implemented over different multihop routing protocols like optimized link state routing (OSLR) protocol [15]. OLSR adopts a proactive, optimized link state scheme to spread topology information while keeping the message overhead low. The key idea is that link-state information is generated and flooded in the network only by selected nodes, called Multipoint Relays (MPRs). Any source-destination route is bidirectional and includes only MPRs as relay nodes. OLSR has been extensively used around the world for building low-cost community owned mesh networks and the metric of ETX has been added in OLSR protocol in several researches.

Based on ETX over OLSR, some modification of this protocol was proposed. This kind of method usually modifies the details of metric and makes the metric overcome the disadvantage of ETX over OLSR. In [16], Johnson and Hancke presented an experimental comparison of OLSR using the standard routing metric and ETX metric in a 7 by 7 grid of closely spaced Wi-Fi nodes. The results show that the ETX metric which has been extensively used in mesh networks around the world is fundamentally flawed when estimating optimal routes in real mesh networks. Houaidia et al. pointed out the shortcoming of the ETX metric for eventual optimizations towards a more efficient routing through using several real experiments [17]. And then they presented improvements of the ETX metric based on link availability for accurately finding high throughput paths in multihop wireless mesh networks. In [18], Pinheiro et al. proposed the OLSR-Fuzzy ETX Queue (OLSRFEQ) protocol to overcome the limitations of OLSR-ETX regarding queue availability and QoS and QoE assurance. OLSR-FEQ optimizes network and user-based parameters by coordinating queue availability, QoS, and fuzzy issues in the routing decision process as a way of allocating the best paths for multimedia applications.

On the other hand, ETX was also implemented over AODV [19], and modifications of ETX based on AODV are proposed in current research. In [20], Ni et al. proposed a modified solution in which they repeatedly broadcast RREQ (Route Request) packets. Simulation results show that their modified solution improves ETX in the initial route selection in both single flows and multiple flows cases.

In another work of ours, we try to implement a mechanism which changes the probe interval of ETXs probe packet to adapt different transmission rates of wireless mesh network. Different from previous methods, this mechanism does not modify the protocol itself but attaches a cross-layer module to dynamically change the value of factors in protocol according to network conditions.

In order to adjust to the variety of wireless networks and reflect link state more accurately, referencing the idea of ExOR in [21] an opportunistic routing mechanism combining long-term and short-term metrics is designed: every mesh router maintains a set of forwarding links and each candidate link in this set has the opportunity to be chosen. In this scheme, the metric of ETX plays the role of main factor in the set and every mesh router will change the set according to the values of ETX. This mechanism not only will choose the link with minimum ETX, but also avoids data transmission over the link with poor quality. Also, this mechanism will enlarge the scale of transmission and improve the performance of transmission under the assurance of users' QoS.

3. Network Model

To understand our proposed mechanism better, we provide an overview of the network model assumed in this section. We consider a wireless mesh network deployed on a simple topology which has three competing data flows. One of the flows needs relay nodes to transmit packets and there are two different ways to transmit. The other two flows are used to influence the routing of ETX. This wireless mesh network adopts the protocol of IEEE 802.11 and antenna type is omniantenna. The nodes in the network are fixed and density of network is also fixed; however, the transmission rate of every node is variable. And hop count of this network is from 1 to 3.

In the network model, we assume that the size f data packet transmitted over the wireless link is the same as the size of probe packet which is used to acquire the information of RTT. In our opportunistic outing mechanism, $\mathrm{ETX}_{\mathrm{threshold}}$ is a threshold to filter candidate links and it should be optimized to achieve better outing performance. The notations used in this paper can be found in Notations.

4. Opportunistic Routing Mechanism

4.1. Disadvantage of ETX. We did simulation with a simple topology based on NS-2 [22] as follows: simple topology and random topology. Through the evaluation metrics of throughput, end-to-end delay, and packet delivery rate, we find the problem that the original ETX will not always perform well when the total data rate of WMN changes

rapidly. That is because its fixed probe interval will not adapt to the change of network environment.

To verify the weakness of the routing metric of ETX, we process the simulation as follows: Figure 2 is a simple topology that is used in our simulation studies. It consists of 7 nodes which are static in the 2000 m ∗ 2000 m ground. The maximum transmission range of nodes is 250 m and the distance between any two neighboring nodes makes one node only communicate with its neighboring nodes directly. Node $n0$ transmits data packets to the destination node $n4$ and there are other two data flows: $n6 \rightarrow n2$ and $n5 \rightarrow n3$.

Our simulator is NS-2.34 and the routing protocol in the simulation is UM-OLSR [23] which is the extension of OLSR with ETX [24]. And the other settings of parameters in our simulation are shown in Table 1.

We vary the probe interval from 0.5 second to 4.0 (0.5, 1.0, 1.5,. . ., 4.0) seconds with different data transmission rates of the data flow $n0 \rightarrow n4$ which is from 0.5 Mbps to 2.0 Mbps, and the data transmission rates of the other two data flows $n6 \rightarrow n2$ and $n5 \rightarrow n3$ are 0.3 Mbps and 1.0 Mbps, respectively. This simulation lasts 50 seconds and these data flows all start at 10 s.

We get the statistical results of throughput, delay, and packet delivery rate of node $n4$ during the simulation.

Figure 3(a) shows that the throughput will decrease with the increment of probe interval when the data rate is not very low (1.0, 1.5, and 2.0 Mbps). However, the throughput will not change a lot with the increment of probe interval when the data flow is low (0.5 Mbps). That is because the collision of packets on WMN is not very severe and the bandwidth of network is wide enough for data packets and probe packets. While the collision between data packets and probe packets will increase when the data rate is high, we could set the probe interval as a smaller value to achieve higher throughput. That is because the more frequently the probe packets are transmitted, the more quickly ETX mechanism could select the better path to transmit data packets. But we can see that there exist inflection points of probe interval which makes the throughput increase (3.5 seconds of probe interval with 2.0 Mbps of data rate). The reason of this phenomenon is that the advantage of reducing packet collisions recovers the disadvantages of switching to the better path slowly. However, to the higher rate of data flow, the longer probe interval will achieve lower end-to-end delay which could be seen in Figure 3(b). So the longer probe interval could be a better choice when the data rate is higher. And in Figure 3(c) we can see almost the same circumstance which is shown in Figure 3(a).

From the results of simulation, we can find that too long probe interval will make ETX perform poorly, while too short interval will cause the waste of bandwidth and lead to longer end-to-end delay. And there exists an optimum interval which will achieve the best performance of ETX, while this interval will change with BER and it is difficult to be predicted. Since ETX does not reflect the change of BER in time with its long-term routing metric, we propose an opportunistic routing mechanism which combines with long-term and short-term routing metrics to solve the problem of ETX when it is applied in practice.

Table 1: Parameter setting.

Parameter	Value
CBR/UPD ∗ 3	Variable
Packet size	1 Kbytes
Routing algorithm	OLSR_ETX
ETX probe interval	Minimum RTT in candidate set
Queue length	50
Mac type	802.11b
Bandwidth	11 Mbps
Propagation	Model two-ray ground
Antenna	Type Omniantenna
Frequency band	2.4 GHz
Simulation time	50 s
RTS/CTS	Off

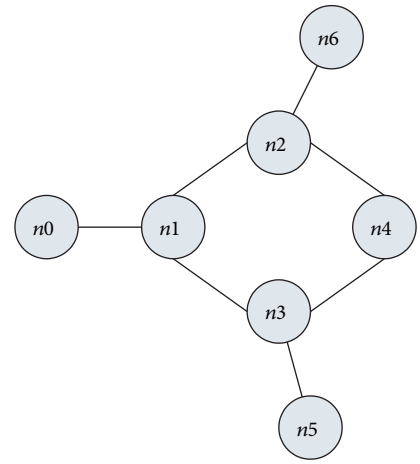


Figure 2: Topology of simulation.

4.2. Opportunistic Routing Mechanism Combined with Short-Term Metric and ETX. The value of ETX of one link is the expected value of statistical link's quality during the latest 10 times. But if the quality of link fluctuates wildly during the period of 10 seconds, the predicted value will be in much warp and make the efficiency of routing lower. From the viewpoint of MAC layer, the node will choose the next hop which is able to reach the destination node and meet the routing algorithm. In the practical process of routing, choosing the path fixedly with minimum cost is not necessarily the best routing and the research in [23] also verified this view. In WMN, which adopts omnidirection antenna, the packets transmitted by source node could cover multineighbor nodes. This kind of routing mechanism with fixed path will not make full use of the sharing medium in wireless networks. So, based on the long-term routing metric such as ETX, combining with the state of links in short-term routing metric, we need to adopt a short-term routing metric to further restrict the routing. In the mechanism with short-term routing metric, the probability of link quality's change is considered as lower in a short period of time (millisecond scale). This mechanism probes the RTT of each candidate link every short period of time and determines the quality of links according to

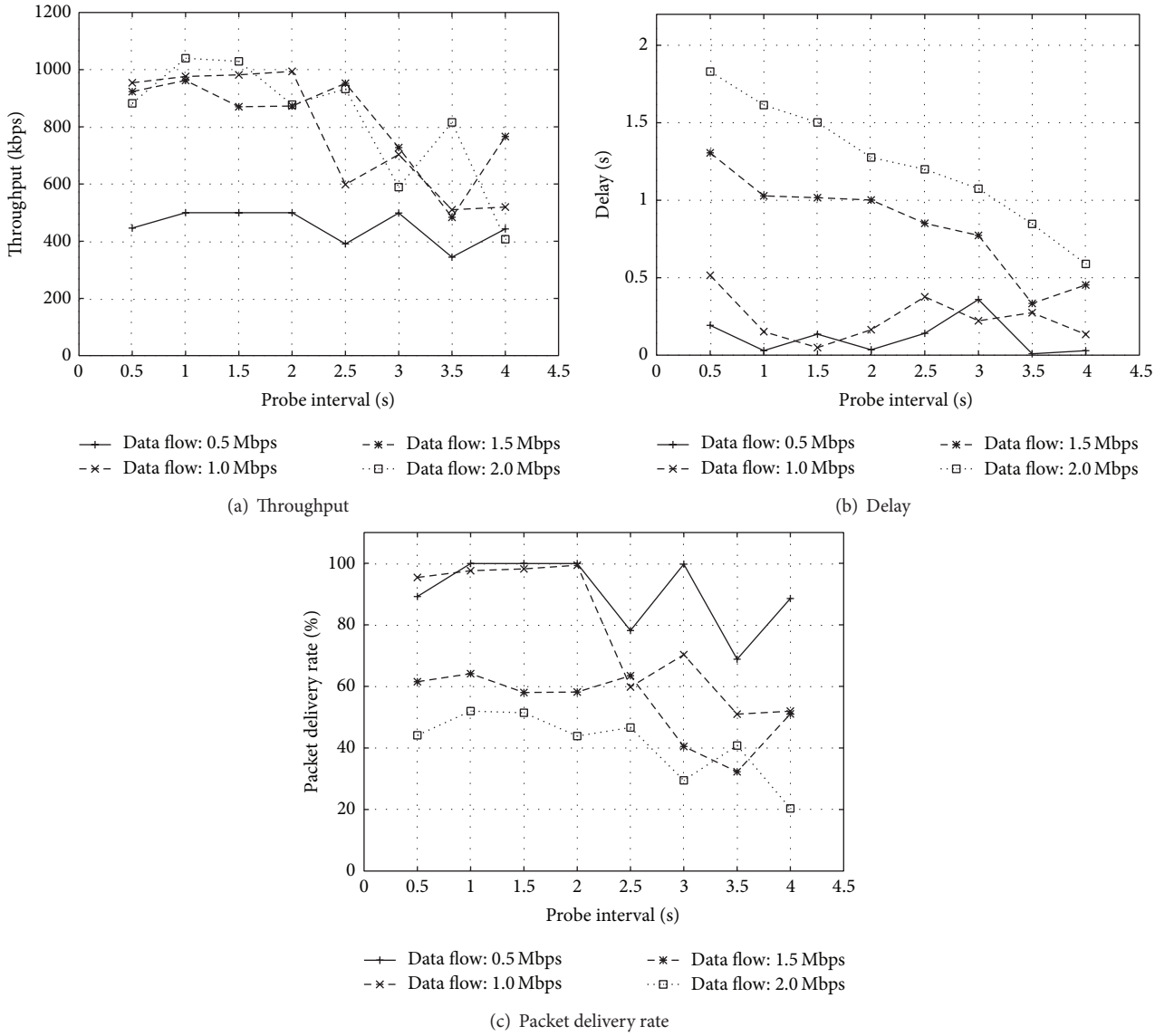


Figure 3: Performance with different probe intervals and data flow rate.

RTT. In this mechanism, some candidate links are selected according to ETX and the next hop will be chosen from these candidate links according to RTT. The short-term routing metric RTT is measured among the candidate links and nodes only maintain the latest one or several values of links' RTTs. Through enlarging the scale of routing in ETX, this mechanism increases the opportunity of being a choice of the links which has smaller ETX and better quality. At the same time, this mechanism increases the opportunity of data transmission over the links with higher quality under the assurance of QoS. And this routing mechanism is a kind of opportunistic method.

This opportunistic mechanism which takes long-term and short-term routing metrics into consideration uses heuristic method, and it is thought that the change of one link's state is continuous in a short period of time (tens of milliseconds). The current state of link is related to the previous two measurements and it will not change suddenly in most cases. This mechanism limits the additional short-term routing metric based on the global routing metric of long term and it improves the performance of routing under the assurance of QoS. Also, this mechanism is a method which takes both limitation of successful transmission count and performance of transmission into account through the additional storage, detection, and comparison in mesh routers.

4.3. Implementation of Opportunistic Routing Mechanism. The opportunistic routing mechanism based on ETX is a method which selects the path with suboptimum value of ETX. There exist multipaths which satisfy the QoS requirement and they will improve the performance of network through the routing of these candidate links by the constraint conditions of network contributes (available bandwidth, etc.).

Combined with short-term routing metric and ETX, the opportunistic routing mechanism assures the choice of link with higher performance every time, while it does not guarantee the same links which are chosen. So the estimation of RTT and constraints of routing are needed in our routing mechanism.

It is thought that the change of one link's state is continuous in a short period of time and the threshold of ETX, $ETX_{threshold}$, is set by network administrator or users. So the implementation of routing with constraints by the short-term routing metric of RTT shows as follows:

(i) the mechanism of short-term routing metric is set by mesh routers and they detect the links whose value of ETX is smaller than $ETX_{threshold}$;

(ii) the mesh routers send probe packets over these candidate links and acquire the current state of network according to RTTs;

(iii) if there is an echo of ACK, the router records the RTT; otherwise, RTT is set by maximum value;

(iv) for one of the candidate links, the router only maintains several historical values and estimates current state of network according to previous two records of RTT;

(v) the router chooses the link with the minimum RTT as the next hop.

If the change of one link's state is continuous, the RTT of the next time can be estimated according to the empirical value of RTT: the RTTs of the latest two moments T_{n-1} and T_{n-2} are $RTT_{n-1}$and RTT_{n-2}, respectively, so RTT_n at T_n is considered RTT_{n-1} approximately. If RTT_{n-1} of candidate links are the same, the mechanism will choose the link with the minimum RTT_{n-2}; otherwise, one of these links will be chosen randomly. The algorithm of opportunistic routing mechanism is shown in Algorithm 1 and the process of opportunistic routing is shown in Figure 4.

In the scenario shown in Figure 5, the data source S records ETX and RTT of every link (Table 2) and $ETX_{threshold}$ takes 4 hops. According to our opportunistic routing mechanism, node S will put links N2 and NM in the candidate set because their values of ETX are smaller than those of $ETX_{threshold}$. Then node S probes the RTT of links N2 and NM and records the latest two results in the storage of S (the unit of RTT in Table 2 is millisecond). Since RTT_{n-1} of N2 and NM is the same, node S will choose NM as the next hop because of its smaller RTT_{n-2}.

This mechanism which combines with long-term and short-term routing metric makes sure that not only ETX of selected link is under the assurance of QoS, but also the bandwidth of the link is the maximum among all of the candidate links. The measurement of long-term routing metric (ETX) is used as that in [2] which sends probe packets every 1 second, while mesh routers just send RTT probe broadcast packets periodically to the neighbor nodes which have the links in the candidate set. This mechanism avoids broadcast storm and will reduce the heavy burden of storage and processing in mesh routers.

Table 2: Parameter setting records in nodes.

Next hop	ETX	RTT_{n-2}	RTT_{n-1}
N1	5	0.4	0.4
N2	3	0.4	0.3
⋮	⋮		
NM	3	0.2	0.3

5. Analyses of Opportunistic Routing Mechanism

The estimation of delay ($Value_{path}$) from source to destination is expressed as the minimum one of the summation of RTT multiplied by EXT of each hop over every path:

$$Value_{path} = \min \left\{ \sum_{\text{hop } i} \left(ETX_i * RTT_i\right) \mid \text{every path} \right\}. \quad (2)$$

For the states of each mesh router over the path, we could transform the model of change of links' states to a model of Poisson distribution and it can be analyzed through Markov process. However, ETX and RTT of wireless links are ever-changing in practical applications, so it cannot be analyzed and implemented by the method of global modeling. On the other hand, there is no clear regulation of wireless link model and no preventative data of wireless condition could become the input for simulation. So we analyze the opportunistic routing mechanism from theory and implementation.

5.1. Analyses of Theory. The transmission rate of ith data packets of a link l at some moment is $R_i = \alpha * (1/RTT_i)$ and α is a factor of transmission time which is related to the size of probe packet. Suppose ETX is measured as E_i and R_i is independent of E_i. Over the link layer, data packets are transmitted with the size of b and probe packets' size is also b. When the transmission rates of data packets and probe packets are the same, the expected transmission time of a packet with the size of b is $t_i = (b/R_i) * E_i = RTT_i \cdot E_i$. For the packet with the size of B, the time of transmitting this packet is shown in (3) ignoring the delay of router's scheduling:

$$T_1 = \sum_{j=i}^{i+\lceil B/b \rceil} t_j = \sum_{j=i}^{i+\lceil B/b \rceil} RTT_j * E_j. \quad (3)$$

For every j, RTT_j and E_j may be different. E_j is the long-term routing metric which will be modified every 1 second and RTT_j reflects the value of short-term routing metric. Suppose that the interval of RTT_j's change is τ, so RTT_j will change $n = \lfloor 1/\tau \rfloor$ times during every E_j's update. If τ takes $15 * 10^{-3}$ s and probe interval of ETX is 1 second, E_j will be updated after RTT_j changes 66 times. In this case, our opportunistic mechanism is able to detect the change of wireless links more accurately and make sure that every packet will be transmitted over the links with better performance.

It will spend $\lceil B/b \rceil$ times for a packet with the size of B over the link with minimum value of ETX. Suppose the

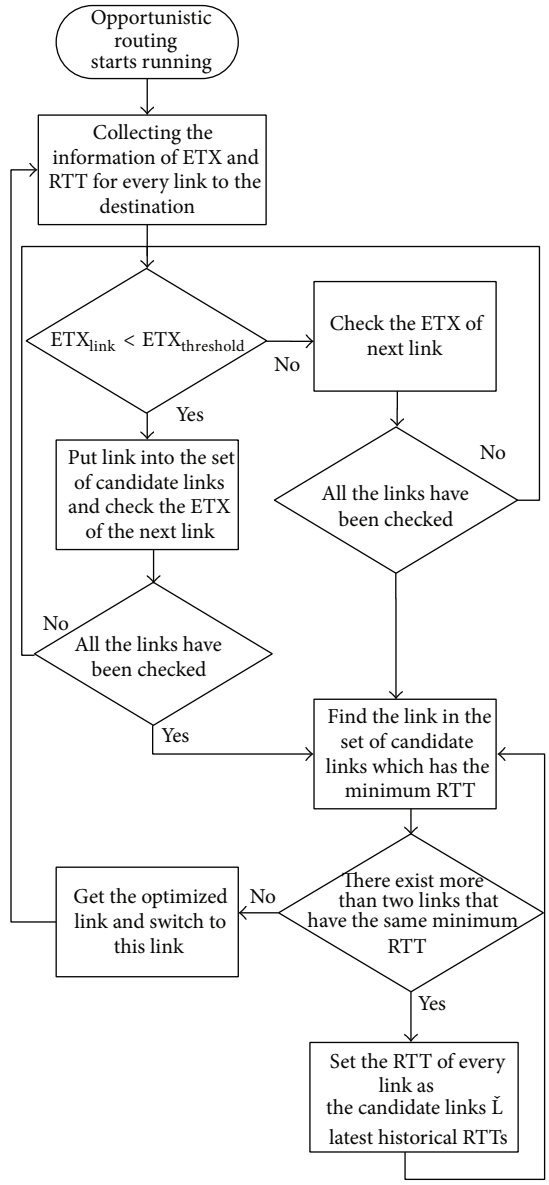


Figure 4: The process of opportunistic routing mechanism.

(1) **for** each *link* from source to destination do
(2) **if** $(ETX_{link} < ETX_{threshold})$
(3) put *link* into the set of candidate links
(4) **end for**
(5) **if** ($Timer_{RTT}$ is timeout)
(6) send RTT probe packet
(7) **if** (ACK of RTT probe is received)
(8) calculate RTT and update it for candiate links
(9) **for** each *link* in the set of candidate links do
(10) if $(RTT_{link} <= RTT_{minimum})$
(11) $link_{chosen} = link$
(12) **end for**

Algorithm 1: The algorithm of opportunistic routing mechanism.

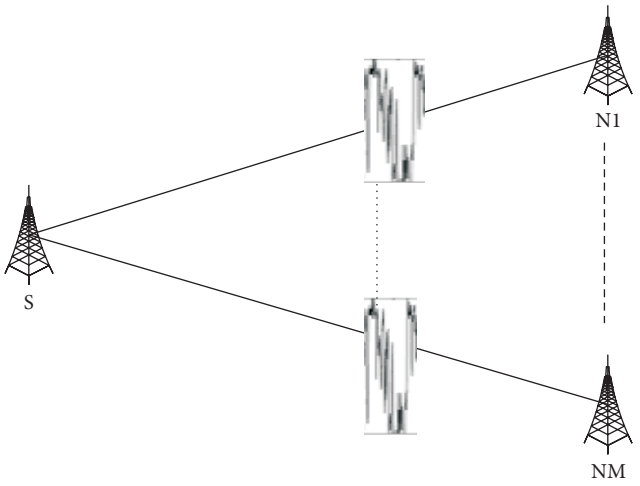


Figure 5: The routing of opportunistic mechanism.

change of links is random and the maximum transmission rate of links is R, so the transmission rate of each packet is $\varepsilon_i R$ ($\varepsilon_i \in [0, 1]$) and the minimum ETX of this link is $E_{\min}$. The time of transmission is shown as follows:

$$T_2 = \left(\sum_{i=1}^{\lceil B/b \rceil} \frac{b}{\varepsilon_i R} \right) * E_{\min} = \left(\sum_{i=1}^{\lceil B/b \rceil} \frac{1}{\varepsilon_i} \right) \mathrm{RTT}_{\min} * E_{\min}. \quad (4)$$

Compared with (3) and (4), E_j is close to $E_{\min}$ because E_j is limited by $\mathrm{ETX}_{\mathrm{threshold}}$. So the margin of deviation between E_j and $E_{\min}$ is very limited ($\xi_j = E_j/E_{\min}$). The value of RTT_j is the minimum one of all the candidate links and it is most close to $\mathrm{RTT}_{\min}$ ($\kappa_j = \mathrm{RTT}_j/\mathrm{RTT}_{\min}$). From (3) and (4), we can get the derivation of

$$\frac{T_1}{T_2} = \frac{\sum_{j=1}^{\lceil B/b \rceil} \kappa_j \xi_j}{\sum_{i=1}^{\lceil B/b \rceil} 1/\varepsilon_i}. \quad (5)$$

In special case, the critical value of T_1/T_2 is 1 when B is equal to b and $\kappa_j \cdot \xi_j = 1/\varepsilon_i$ which can also be expressed by

$$\varepsilon_i = \frac{\mathrm{RTT}_{\min}}{\mathrm{RTT}_j} * \frac{E_{\min}}{E_j}. \quad (6)$$

In common conditions of wireless networks, the state of network accords with the uniform distribution from 0 to 1. So the expectation of the left of (6) is 0.5. The bigger the value of $\mathrm{ETX}_{\mathrm{threshold}}$ is, the more links will be taken into the candidate set and RTT_j will be closer to $\mathrm{RTT}_{\min}$; however, $E_{\min}/E_i$ will be smaller at the same time. So there will also be more links taken in the candidate set with $\mathrm{ETX}_{\mathrm{threshold}}$ of $(E_{\min} + 1)$ when there are more links to the next hop.

At the same time, the actual and maximum throughput of l are Th_1 and Th_2, respectively, and their ratio is shown in (7) under the assumption that B is equal to b.

If the value of $\mathrm{ETX}_{\mathrm{threshold}}$ decreases, the number of candidate links in set will also decrease. And then E_j will be closer to $E_{\min}$; however, it will be unable to predict the changing of RTT_j because of the decrement of candidate links. At the same time, although sending RTT probe packet more frequently will make RTT_j more accurate, overmuch probe packets will lead to more collisions between packets and then E_j may be larger. So the settings of RTT probe packets and $\mathrm{ETX}_{\mathrm{threshold}}$ are very significant to increase throughput of network. Under the assumption that $\mathrm{ETX}_{\mathrm{threshold}}$ and the interval of RTT probe packets are optimized, since the value of ε_i is relatively constant, the ratio between Th_1 and Th_2 will be closed to 1. And the throughput of the chosen link will achieve the maximum. In particular, in relative dense network, this mechanism of routing which takes short-term and long-term routing metric into account will perform better:

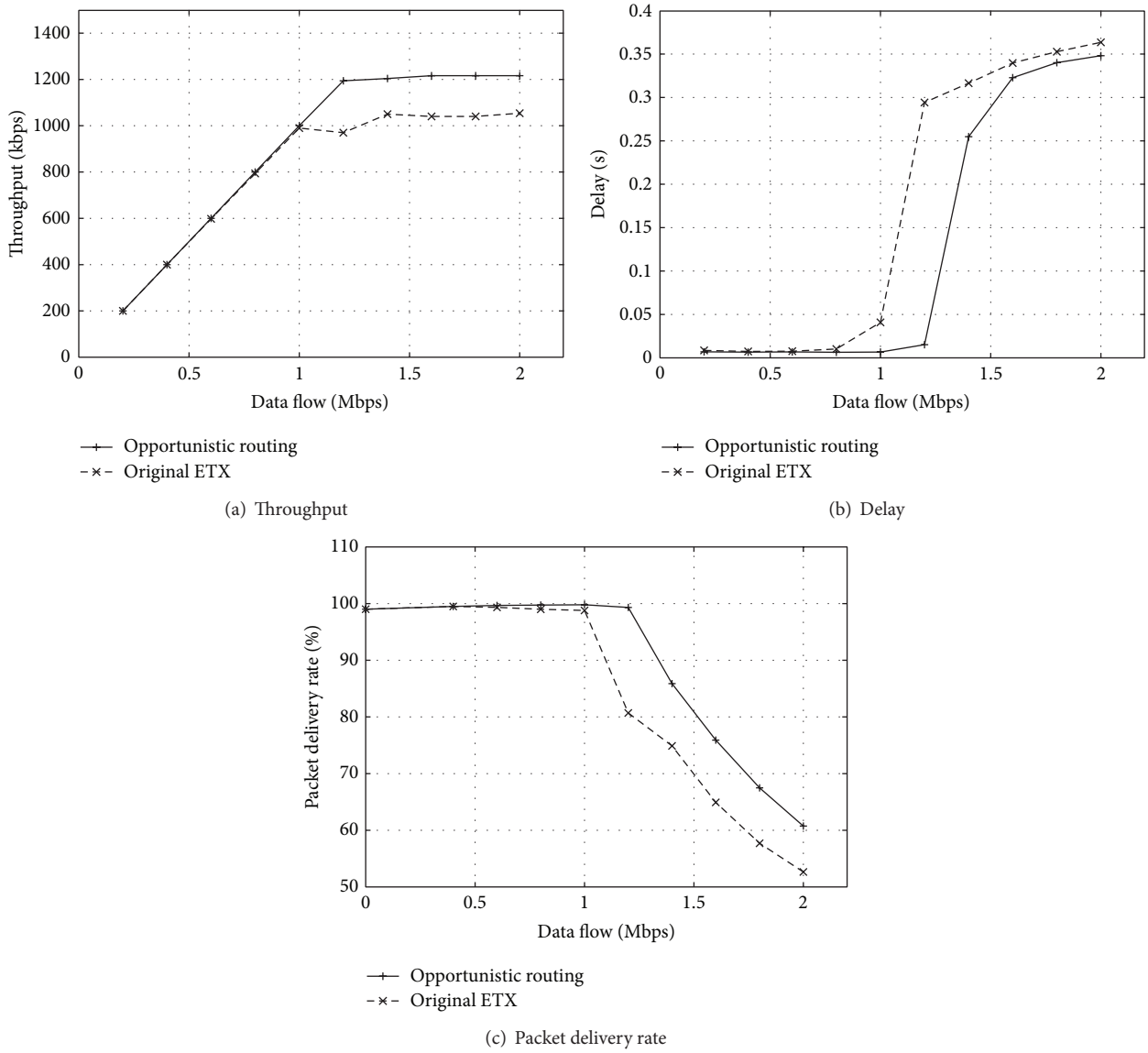
$$\frac{\mathrm{Th}_1}{\mathrm{Th}_2} = \frac{B/\left(\sum_{j=1}^{\lceil B/b \rceil} \kappa_j \xi_j\right)}{B/\left(\sum_{i=1}^{\lceil B/b \rceil} 1/\varepsilon_i\right)} = \frac{\mathrm{RTT}_{\min} E_{\min}}{\varepsilon_i \mathrm{RTT}_j E_j}. \quad (7)$$

5.2. Performance Evaluation of Opportunistic Routing. We implement the function of opportunistic routing mechanism in NS-2.34 simulator, and then we process another simulation to evaluate the routing performance of opportunistic routing. In simulation, we also used the simple topology in Figure 1. The simulation period is 50 seconds and we vary the data rate of source-destination pair between $n0$ and $n4$ which make the data rate change from 0.2 Mbps to 2.0 Mbps. Then, the source node sends the packets to the destination node using original ETX and opportunistic routing. The results of performance comparison between original ETX and vp-ETX different data rates are shown in Figure 6.

Figure 6(a) shows that the throughput of opportunistic routing is always higher than original ETX when the data rate is higher than 1 Mbps and reaches the highest throughput which is about 1.216 Mbps. And the throughput is increased about 20% when the data rate is higher than 1 Mbps. The throughputs of original ETX and opportunistic routing are almost the same when the data rate is lower than 1 Mbps. However, the delay of opportunistic routing is a little lower than original ETX which could be seen in Figure 6(b). And the advantage of delay of opportunistic routing is more obvious when the data rate is higher. So the real time of opportunistic routing is better when network is dense and data rate is higher. To the packet delivery rate, opportunistic routing always performs better than original ETX at every data rate which could be seen in Figure 6(c).

From the analysis of the results of Figure 6, we find that opportunistic routing could perform better when the data rate of WMN is higher. So opportunistic routing mechanism is extremely suitable for the scenario of WMN with heavy data traffic.

5.3. Analyses of Implementation. Mesh routers need to support this opportunistic routing mechanism which combines with short-term routing metric for its implementation. Routers need to support the metric of ETX and detect RTT for a group of specific links. Particularly, when it detects that there is only one link whose ETX is smaller than the threshold, the mesh router should ignore the mechanism of short-term routing metric. On the other hand, our routing



(a) Throughput

(b) Delay

(c) Packet delivery rate

Figure 6: Performance comparison with different data rates.

mechanism should be compatible with traditional mechanism with the metric of ETX and it will not enable the mechanism of short-term routing metric when $ETX_{threshold}$ takes 0. For the choice of $ETX_{threshold}$, the value of this threshold should be set according to the density of network. While the opportunistic mechanism cannot perform well in an extreme dense traffic network. The opportunistic mechanism avoids the bottleneck of transmission because of its avoidance of all packets transmitted over the same link simultaneously. Since the value of ETX changes at any time, the paths of packets transmitted may be different when the topology and ETX of network update, while the paths will not change much but will be fixed relatively when WMN adopts the opportunistic routing mechanism in some cases.

In practice, the coverage of IEEE802.11b is 300~500 meters and the speed of electromagnetic wave propagation is $2 * 10^8$ meters per second. Although, considering the processing delay, RTT of adjacent neighbor nodes is far less than 1 millisecond, for WMN based on IEEE 802.16a (WiMAX) [24], the coverage of signal could be tens of kilometers. So RTT between neighbor nodes will reach the level of millisecond. Considering the influence with mesh routers' performance and compatibility with WiMAX of frequent detection, the level of short-term routing metric's probe interval should be millisecond [22].

Mesh routers will detect RTT of each link in the candidate set every 10 ms (empirical value), so it brings extra messages to network and requirement of routers' storage and computation ability. In the process of detecting RTT to neighbor nodes, the mesh router of detecting will broadcast probe packets and the links in the candidate set will respond, and then the router calculates RTT according to the response packet. In the process of probing and responding, the size of packet should be as small as possible; because of the

smaller size of probe packet and longer probe interval, it does not bring too many messages to network. Since RTT is calculated by sender routers according to the current time and timestamp of response packet and candidate links are relatively limited, it will not bring too many mesh router operations and the performance of network transmission will not be influenced too much.

The opportunistic mechanism works across network and MAC layers. So it is responsible for routing and forwarding but not for order-preserving and error control of packets. The paths of packets transmitted may be different and packets arrive out of order if it adopts the opportunistic mechanism. For TCP flows, reliability of transmission can be assured through sequence number and congestion control mechanism, while, for UPD flows, it needs the process of order-preserving by application layer and it is the requirement of the opportunistic mechanism to upper layer protocols. It implies that the quality of link changes much when the packets of one date flow transmitted over different paths and the performance will decrease and even lead to the interruption of date flow if it still adopts original routing, while the opportunistic mechanism implements the transmission with higher efficiency and assures the quality of transmission through adjusting the routing of single packet in data flow. Moreover, this kind of mechanism avoids the situation that many data flows are transmitted over the same path and will not lead to failed transmission which is caused by the bottleneck of wireless links.

6. Conclusion

An opportunistic routing mechanism based on the routing metric of QoS-aware (ETX) is proposed in this paper. Through the simulation in this paper, ETX has proved that this kind of routing metric cannot reflect the quality of links quickly. However, our routing mechanism which combines with the routing metrics of short-term and long-term is proved theoretically to be able to overcome the weakness of ETX. Under the assurance of data flows' QoS, the link with highest performance will be chosen to transmit packets and this mechanism can better satisfy the requirement of wireless multihop network than the mechanism with ETX. This method also will change routing according to the quality of wireless links. Through analyses of performance and implementation, the opportunistic routing mechanism can increase the transmission efficiency of wireless links and improve the performance of wireless multihop network. In particular, in the scenario of heavy data traffic in WMN, opportunistic routing mechanism performs far better than original ETX mechanism.

We will find optimum values of $ETX_{threshold}$ and probe interval of RTT in the future research. The further simulation of the topology of dense network is needed to verify the advantages of our routing mechanism. Also to prove the efficiency of the proposed mechanism, a verification of test-bed in a real scenario will be our future work.

Notations

T_i:	Transmission time for network packets
E_i:	ETX of link i
RTT_i:	Round-trip time of link i
E_{min}:	Minimum ETX in candidate set
RTT_{min}:	Minimum RTT in candidate set
$ETX_{threshold}$:	A threshold to filter candidate links
B:	Size of data packets
b:	Size of data payload over the wireless link
R_i:	Transmission rate of packets
Th_i:	Throughput of data packets over link i.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] H. Cheng and G. Su, "Band width guaranteed routing with multiple links in multi-radio wireless mesh networks," in *Future Wireless Networks and Information Systems*, pp. 663–671, 2012.

[2] D. S. J. D. Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," *Wireless Networks*, vol. 11, no. 4, pp. 419–434, 2005.

[3] J. A. Cordero, E. Baccelli, and P. Jacquet, "OSPF over multi-hop ad hoc wireless communications," 2010.

[4] C. Hedrick, "Routing Information Protocol," RFC 1058, 1988.

[5] S. S. Ahmeda and E. A. Esseid, "Review of routing metrics and protocols for wireless mesh network," in *Proceedings of the 2nd Pacific-Asia Conference on Circuits, Communications and System (PACCS '10)*, vol. 1, pp. 27–30, Beijing, China, August 2010.

[6] P. K. Bedi, M. S. Aswal, and P. Kumar, "An improved hop-count metric for infrastructural wireless mesh network," *International Journal on Computer Science and Engineering*, vol. 3, no. 5, pp. 1757–1763, 2011.

[7] A. Adya, P. Bahl, J. Padhye, A. Wolman, and L. Zhou, "A multi-radio unification protocol for IEEE 802.11 wireless networks," in *Proceedings of the 1st International Conference on Broadband Networks (BroadNets '04)*, pp. 344–354, San Jose, Calif, USA, October 2004.

[8] R. Draves, J. Padhye, and B. Zill, "Comparison of routing metrics for static multi-hop wireless networks," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '04)*, pp. 133–144, Portland, Ore, USA, September 2004.

[9] J.D. Padhye, R. P. Draves Jr., and B. D. Zill, "System and method for link quality routing using a weighted cumulative expected transmission time metric," US Patent 7,616,575, 2009.

[10] Z. Liang and A. Y. Al-Dubai, "Routing metrics for wireless mesh networks: a survey," in *Recent Advances in Computer Science and Information Engineering*, pp. 311–316, 2012.

[11] D. Aguayo, J. Bicket, S. Biswas, G. Judd, and R. Morris, "Link-level measurements from an 802.11b mesh network," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 121–131, Portland, Ore, USA, September 2004.

[12] H. Mogaibel, M. Othman, S. Subramaniam, and N. A. W. A. Hamid, "Impact of the hybrid multi-channel multi-interface

wireless mesh network on ETX-based metrics performance," *Electrical Power Systems and Computers*, vol. 99, no. 3, pp. 147–160, 2011.

[13] C. E. Koksal and H. Balakrishnan, "Quality-aware routing metrics for time-varying wireless mesh networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 1984–1994, 2006.

[14] R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pp. 114–128, Philadelphia, Pa, USA, October 2004.

[15] P. Jacquet, P. Muhlethaler, T. Clausen, A. Laouiti, A. Qayyum, and L. Viennot, "Optimized link state routing protocol for ad hoc networks," in *Proceedings of the IEEE INMIC International Multi Topic Conference: Technology for the 21st Century*, p. 6268, 2001.

[16] D. Johnson and G. Hancke, "Comparison of two routing metrics in OLSR on a grid based mesh network," *Ad Hoc Networks*, vol. 7, no. 2, pp. 374–387, 2009.

[17] C. Houaidia , A. Van Den Bossche, H. Idoudi et al., "Link availability aware routing metric for wireless mesh networks," in *Proceedings of the IEEE International Conference on Computer Systems and Applications (AICCSA '13)*, pp. 1–4, 2013.

[18] B. Pinheiro, V. Nascimento, R. Lopes, E. Cerqueira, and A. Abelem, "A fuzzy queue-aware routing approach for wireless mesh networks," *Multimedia Tools and Applications*, vol. 61, no. 3, pp. 747–768, 2012.

[19] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc OnDemand Distance Vector (AODV) Routing RFC 3561," http://www.rfc-archive.org/getrfc.php?rfc=3561.

[20] X. Ni, L. Kun-Chan, and M. Robert, "On the performance of expected transmission count (etx) for wireless mesh networks," in *Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools*, p. 77, 2008.

[21] Network Simulator 2, http://www.isi.edu/nsnam/ns/.

[22] S. Paris, A. Capone, C. Nita-Rotaru, and F. Martignon, "A cross-layer reliability metric for wireless mesh networks with selfish participants," *ACM SIGMOBILE Mobile Computing and Communications*, vol. 14, no. 3, pp. 1–3, 2010.

[23] F. J. Ros, "UM-OLSR, an implementation of the OLSR (Optimized Link State Routing) protocol for the ns-2 network simulator," 2007, http://masimum.inf.um.es/fjrm/development/um-olsr/.

[24] W. Cordeiro, "OLSR-ETX module for NS-2," 2009, http://www.inf.ufrgs.br/~wlccordeiro/resources/olsr/README.html.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 350789, 8 pages
http://dx.doi.org/10.1155/2014/350789

Research Article

Node Deployment Algorithm Based on Viscous Fluid Model for Wireless Sensor Networks

Jiguang Chen and Huanyan Qian

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 21094, China

Correspondence should be addressed to Jiguang Chen; jiguang.chen@outlook.com

Received 29 April 2014; Accepted 26 June 2014; Published 14 July 2014

Academic Editor: Zhongmei Zhou

With the scale expands, traditional deployment algorithms are becoming increasingly complicated than before, which are no longer fit for sensor networks. In order to reduce the complexity, we propose a node deployment algorithm based on viscous fluid model. In wireless sensor networks, sensor nodes are abstracted as fluid particles. Similar to the diffusion and self-propagation behavior of fluid particles, sensor nodes realize deployment in unknown region following the motion rules of fluid. Simulation results show that our algorithm archives good coverage rate and homogeneity in large-scale sensor networks.

1. Introduction

As one of the basic problems in wireless sensor networks, deployment algorithm has attracted scholars' wide attention, while a series of algorithms have been put forward in allusion to different deployment demand. In this paper, with regard to the problem of large-scale wireless sensor network deployment, a scheme based on viscous fluid model is proposed.

A large number of researches have been done to solve various issues related to the deployment problem. There are practically three categories of methods: methods based on geometrical model, methods based on virtual potential field, and methods based on biological intelligence. Since fluid field is one kind of virtual potential field, the algorithm proposed in this paper belongs to methods based on virtual potential field.

The most classical method based on virtual potential field is proposed by Howard et al. [1]. They provided a solution to the problem by deploying a mobile sensor network in unknown dynamic environments. Moreover, they described a potential-field-based approach for deployment, in which nodes were treated as virtual particles, subject to virtual forces. These forces repel nodes from each other and from obstacles, ensuring that, from an initial compact configuration, nodes will spread out to maximize the coverage area of the network. In addition to these repulsive forces, nodes are also subject to a viscous friction force. This force is used to ensure that the network will eventually reach the state of static equilibrium; that is, all nodes will ultimately come to a complete stop. Similarly, the virtual force algorithm of [2] and the virtual spring force algorithm of [3] use both repulsive and attractive force components to maximize coverage and uniformity of a given number of sensors.

As we know, electrostatic field also is a kind of virtual potential field. In [4], Toumpis and Tassiulas imagined a scenario: the spatial distribution of sources and sinks is fixed, but we are free to place wireless nodes as we like. Then they raised a question: how to calculate the minimum number of nodes so as to support the traffic, as well as the associated placement of nodes that achieves this minimum. Apparently, it is plausible to solve this problem with regular approach. Yet, this may lead to many troubles and takes more time. By contrast, they solved the problem with electrostatic field easily and efficiently that is the advantage when we introduce electrostatic field in sensor networks.

Recently, mobile sensor network node deployment based with virtual potential field is drawing researchers' more and more attentions. In [5], the authors proposed an approach, in which Delaunay triangulation was formed with these nodes, while adjacent relationship was defined if two nodes were

connected in the Delaunay diagram. Force could only be exerted from those adjacent nodes within the communication range. Simulation results showed that the proposed approach had higher coverage rate and shorter convergence time than traditional virtual force algorithm. In [6], when Li and his companions were trying to overcome the connectivity maintenance and node stacking problems with traditional virtual force algorithm (VFA), they developed an extended virtual force-based approach to achieve the ideal deployment. Simulation results showed that the virtual force approach could effectively reach ideal deployment in mobile sensor networks with different ratio of communication range to sensing range. Furthermore, it achieved better performance in coverage rate, distance uniformity, and connectivity uniformity than previous VFA. In [7], they presented an energy-efficient self-deployment scheme to utilize the attractive force generated from the centroid of a sensor's local Voronoi polygon, as well as the repulsive force frequently used in self-deployment schemes with potential field. The simulation results showed that their scheme could achieve a higher coverage, leading to less sensor movements in shorter time than self-deployment schemes with traditional potential field.

In this paper, we present an adaptive deployment strategy that guarantees good coverage and uniformity, with only part of the deployment environment being given. Provided the approximate size of the deployment region, the algorithm can compute the number of nodes needed by setting appropriate parameters, thereby completing the deployment. The subsequent contents of this paper are arranged as follows: related works are introduced in Section 2. Section 3 briefly introduces knowledge related to fluid model. Section 4 is designed to establish mobile sensor network deployment model based on viscous fluid. Then, the model is to be solved to establish deployment algorithm. Section 5 takes advantage of computer software to simulate the deployment algorithm, as well as its performance. Finally, Section 6 summarizes the whole paper.

2. Related Work

There have been plenty of representative literatures [8–12] related to wireless sensor network deployment algorithm. Sergiou and Vassiliou [8] employed a macroscopic fluid dynamic model to estimate the maximum volume of traffic that may be carried out from the sources to the sink(s) of a WSN. Gribaudo et al. [9] claimed that the behavior of large-scale WSNs was complex and difficult to analyze. Thus, they developed an analytical model of the behavior of WSNs, based on a fluid approach. Actually, they represented WSNs by a continuous fluid entity distributed in the network area. Another work [10] proposed two gas models, one of which used a virtual force approach and the other adopted akinetic approach. Pac et al. [11, 12] built hydrodynamic model for mobile sensor network, taking the entire network as fluid and mobile nodes as microelements in fluid. As for this, the problem of node deployment is transformed into a problem of hydrodynamic governing equation. Relying on the self-diffusion property of fluid, nodes, being regarded as fluid microelements, may be diffused to the deployment area along with fluid, realizing automatic deployment. The author defined the process as "self-deployment."

The aforementioned literatures have made contributions to the deployment of mobile sensor network. Yet, they still need to be further strengthened, especially in allusion to large-scale sensor network. The deployment approach proposed in this paper is simple and pragmatic, leading to good adaptability in water, on the ground, and in the air. Literature [11, 12] provides our research with some ideologies. Based on previous researches, the paper puts forward a sensor network deployment scheme based on viscous fluid model.

Deployment algorithms mentioned in the above literatures are all based on virtual potential field. They share something in common that the deployment of mobile sensor nodes are considered as a coverage process. Nodes are affected by virtual force and eventually reach equilibrium from the initial position (randomly or prematurely set), thereby finishing the deployment of the entire network. Although these algorithms assume the deployment environment is unknown, claim themselves adaptive, and can complete the task before knowing the environment, they all implicitly presume that the size of deployment region is known, so is the number of nodes that need to be deployed. However, if the size of the deployment environment is not known a priori, these algorithms can only provide coverage to the size extent of the area that is previously fixed by the number of nodes to be deployed. Thus, a certain quality of service could not be guaranteed with these approaches. Therefore, the adaptivity and scalability of the deployment algorithms are dramatically limited.

3. Fluid Model

Hydrodynamics is a branch discipline of mechanics, which takes fluid as the study object, so as to research the motion of fluid, as well as the relationship and rule between acting forces. The principle of hydrodynamics may be applied in sensor networks, building a wireless sensor network based on flow field theory.

3.1. Viscous Fluid Motion Differential Equation. For continuous fluid, there are mainly four types of fluid models [13], where finite control volume model is divided into two types. Similarly, infinitesimal fluid micelle model is also comprised of two types. In this paper, nodes are regarded as fluid micelles that flow with fluid.

Randomly selecting an infinitesimal fluid micelle, and assuming that its volume microelement is dV, fluid micelle moves along filament line, while its velocity V is equal to the flow velocity along the filament line. Randomly selecting a spatial point $M(x, y, z)$ from the flow field, with the side length separately denoted by dx, dy, and dz, is shown in Figure 1. Assume that the velocity at point M at the moment is $v(x, y, z, t)$, while vectors of velocity v along the three coordinate directions are separately $v_x(x, y, z, t)$, $v_y(x, y, z, t)$, and $v_z(x, y, z, t)$, and fluid density is $\rho\ (x, y, z, t)$. In accordance with momentum conservation law, the accelerated velocity

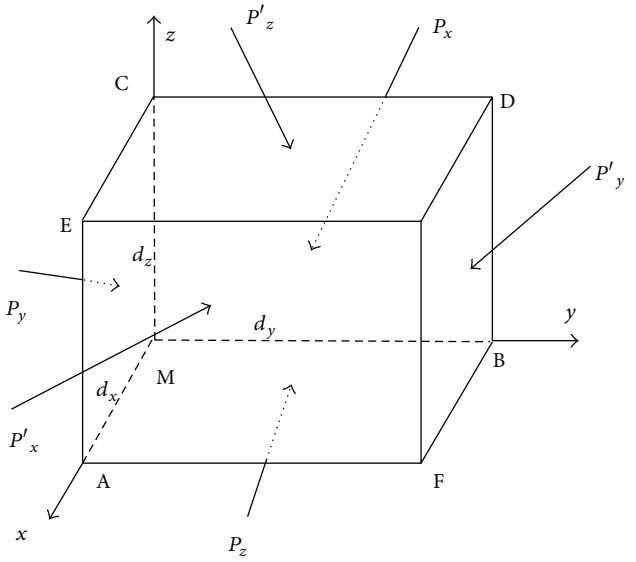


Figure 1: Fluid micelle force analysis.

shall be $a = dv/dt$. Along the three coordinate directions, the accelerated velocities are separated and described as $a_x = dv_x/dt$, $a_y = dv_y/dt$, and $a_z = dv_z/dt$.

External forces acting on the fluid micelle include mass force ρ, f, d_x, d_y, and d_z, as well as surface forces acting on the six surfaces. There is viscous effect between micelle and fluid around, so that the surface force is normally no longer perpendicular to their respective acting surface. P_x, P_y, and P_z are employed to separately describe surface forces on the three surfaces of MBDC, MCEA, and MAFB, so that $P_x = p_x d_y d_z$, $P_y = p_y d_y d_x$, and $P_z = p_x d_x d_y$. Decompose these surface forces as normal stress and shear stress, so that

$$\begin{aligned} P_x &= p_{xx} i + p_{xy} j + p_{xz} k, \\ P_y &= p_{yx} i + p_{yy} j + p_{yz} k, \\ P_z &= p_{zx} i + p_{zy} j + p_{zz} k. \end{aligned} \tag{1}$$

In the equation, $p_{xx}, p_{yz}, \ldots$ separately stands for stress components, while the first subscript refers to the normal direction of acting surface and the second subscript refers to the acting direction of stress. As for this, p_{xx}, p_{yy}, and p_{zz} separately represent normal stress components, while the rest $p_{xy}, p_{xz}, \ldots$ are shear stress components.

Surface forces acting on the micelle also include $P'_x = p'_x d_y d_z$, $P'_y = p'_y d_z d_x$, and $P'_z = p'_z d_x d_y$, on the three surfaces of GEAF, GFBD, and GDCE. These surface forces may be decomposed into their respective normal and shear components. Compared with the aforementioned three surfaces, the three acting surfaces are slightly changed in aspect of coordinate, that is, d_x, d_y, and d_z. Being expanded according to Taylor of function, and abandoning high-order small quantity, normal stress and shear stress on acting surface may be described as

$$\begin{aligned} p'_x &= \left(p_{xx} + \frac{\partial p_{xx}}{\partial x} dx\right) i + \left(p_{xy} + \frac{\partial p_{xy}}{\partial x} dx\right) j \\ &\quad + \left(p_{xz} + \frac{\partial p_{xz}}{\partial x} dx\right) k, \\ p'_y &= \left(p_{yx} + \frac{\partial p_{yx}}{\partial y} dy\right) i + \left(p_{yy} + \frac{\partial p_{yy}}{\partial y} dy\right) j \\ &\quad + \left(p_{yz} + \frac{\partial p_{yz}}{\partial y} dy\right) k, \\ p'_z &= \left(p_{zx} + \frac{\partial p_{zx}}{\partial z} dz\right) i + \left(p_{zy} + \frac{\partial p_{zy}}{\partial z} dz\right) j \\ &\quad + \left(p_{zz} + \frac{\partial p_{zz}}{\partial z} dz\right) k. \end{aligned} \tag{2}$$

According to Newton's second law, motion equation of micelle along the three coordinate directions may be figured out, while the motion equation along axis x shall be

$$\begin{aligned} \rho dx dy dz \frac{dv_x}{dt} &= f_x \rho dx dy dz - p_{xx} dy dz \\ &\quad + \left(p_{xx} + \frac{\partial p_{xx}}{\partial x} dx\right) dy dz - p_{yz} dz dx \\ &\quad + \left(p_{yx} + \frac{\partial p_{yx}}{\partial y} dy\right) dz dx - p_{zx} dx dy \\ &\quad + \left(p_{zx} + \frac{\partial p_{zx}}{\partial z} dz\right) dx dy. \end{aligned} \tag{3}$$

The above equation may be simplified as

$$\rho \frac{dv_x}{dt} = \rho f_x + \frac{\partial p_{xx}}{\partial x} + \frac{\partial p_{yx}}{\partial y} + \frac{\partial p_{zx}}{\partial z}. \tag{4}$$

Put the stress component in the above equation with generalized Newton's laws of internal friction, and simplify the expression:

$$\begin{aligned} \rho \frac{dv_x}{dt} &= \rho f_x - \frac{\partial}{\partial x}\left[-p - \frac{2}{3}\mu\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right) + 2\mu \frac{\partial v_x}{\partial x}\right] \\ &\quad + \frac{\partial}{\partial y}\left[\mu\left(\frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y}\right)\right] + \frac{\partial}{\partial z}\left[\mu\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial z}\right)\right]. \end{aligned} \tag{5}$$

By expanding the second term on the right side of the equal sign, the following expression is figured out upon simplification:

$$\begin{aligned} \rho \frac{dv_x}{dt} &= \rho f_x - \frac{\partial p}{\partial x} + \mu\left(\frac{\partial^2 v_x}{\partial x^2} + \frac{\partial^2 v_y}{\partial y^2} + \frac{\partial^2 v_z}{\partial z^2}\right) \\ &= \rho f_x - \frac{\partial p}{\partial x} + \mu \nabla^2 v_x. \end{aligned} \tag{6}$$

Expression (6) is viscous fluid momentum equation, that is, Navier-Stokes equation, an important differential equation researching practical fluid.

4. Sensor Network Deployment Based on Viscous Fluid Model

4.1. Viscous Fluid Model. The property of fluid resisting the relative motion between micelles is referred to as viscidity. Assuming that there is viscidity between nodes, such viscidity is utilized to control motion between nodes, so as to maintain the connectivity of network.

Navier-Stokes equation based on viscous flow is shown in

$$\frac{dU}{dt} = f - \frac{1}{\rho}\nabla p + \nu\nabla^2 U, \tag{7}$$

where du/dt stands for velocity derivative, that is, accelerated velocity; f refers to vector of unit acting force, ρ is fluid density, and p represents fluid pressure, while U stands for the velocity of an infinitesimal fluid element. In the end, ν is motion viscosity of fluid, and $\nu = \mu/\rho$.

In this paper, we are only to study the motion of two-dimensional fluid. As for this, only two-dimensional vector form of Navier-Stokes equation is adopted. In rectangular coordinate system, related components may be described as

$$\begin{aligned}\frac{du}{dt} &= f_x - \frac{1}{\rho}\frac{\partial p}{\partial x} + \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right),\\ \frac{dv}{dt} &= f_y - \frac{1}{\rho}\frac{\partial p}{\partial y} + \nu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right),\end{aligned} \tag{8}$$

where u and v separately refer to velocity components of an infinitesimal fluid element along the respective direction, while f_x and f_y are component forces per unit along x and y direction. Consider

$$\begin{aligned}\frac{du}{dt} &= \frac{\partial u_i^t}{\partial t} + u_i^t\frac{\partial u_i^t}{\partial x} + v_i^t\frac{\partial u_i^t}{\partial y},\\ \frac{dv}{dt} &= \frac{\partial v_i^t}{\partial t} + u_i^t\frac{\partial v_i^t}{\partial x} + v_i^t\frac{\partial v_i^t}{\partial y}.\end{aligned} \tag{9}$$

Putting (9) in (8), differential equation may be figured out as

$$\begin{aligned}\frac{\partial u_i^t}{\partial t} &= \frac{du}{dt} - \left(u_i^t\frac{\partial u_i^t}{\partial x} + v_i^t\frac{\partial u_i^t}{\partial y}\right),\\ \frac{\partial v_i^t}{\partial t} &= \frac{dv}{dt} - \left(u_i^t\frac{\partial v_i^t}{\partial x} + v_i^t\frac{\partial v_i^t}{\partial y}\right).\end{aligned} \tag{10}$$

In (10), subscript i and superscript t are applied to identify the values of parameters u, v, ρ, p, f, and ν of fluid element i at the moment of t.

4.2. Analogy Relationship between Flow Field and Sensor Network. When fluid model is applied to analyze the deployment process of wireless sensor network, nodes shall be assumed with some basic properties: nodes are able to move freely; each network node is able to acquire its position, velocity, and pressure; each sensor node has a sensing range with radius of Rs to sense the position of obstruction, and so forth. Each sensor node has a communication radius with distance of R.

In addition, it is also assumed that there is viscosity between nodes, while the viscosity is related to the distance between neighbor nodes. Larger distance leads to higher viscosity, and smaller distance leads to lower viscosity. Thus, in deployment process, overdispersion of nodes may be avoided, so as to ensure the connectivity of network.

Counter definitions of the above fluid variables in mobile sensor network may be described as follows.

(1) Velocity Vector. Velocity vector of nodes is similar to velocity vector of viscous fluid element. The velocity of node i at the moment of t may be descried as $V_i^t = (u_i^t, v_i^t)$.

(2) Density. Local density of the position, where each node is located, may be described by ρ_i. The definition is shown by

$$\rho_i = 1 + \frac{Rc^2}{\bar{r}_{ij}^{\,2}} \times n_i = 1 + \frac{Rc^2 n_i^2}{\left(\sum_{j\in N_i} r_{ij}\right)^2}. \tag{11}$$

Rc is communication radius of node i, while N_i refers to the set of a series of neighbor nodes j located within this range. n_i refers to the number of neighbor nodes within the communication range, and r_{ij} stands for the Euclidean distance from node i to neighbor nodes j within the communication range.

(3) Intensity of Pressure. Intensity of pressure is the main driving entity in flowing of fluid. For ideal gas, the status equation is shown by

$$p = \rho RT, \tag{12}$$

where ρ refers to local density, R is specified constant, and T represents absolute temperature.

(4) Spatial Derivative of Parameter. In (10), there are derivatives of velocity and pressure intensity, in allusion to spatial dimensionality of x and y. Grid-less method [14] is adopted to define first-order difference equation of flow variable, while the variable is identified by ξ. Consider

$$\begin{aligned}\frac{\partial \xi_i^t}{\partial x} &= \frac{1}{n_i^t}\sum_{j\in N_i^t}\frac{\left(\xi_j^t - \xi_i^t\right)}{r_{ij}}\cos\theta_{ij},\\ \frac{\partial \xi_i^t}{\partial y} &= \frac{1}{n_i^t}\sum_{j\in N_i^t}\frac{\left(\xi_j^t - \xi_i^t\right)}{r_{ij}}\sin\theta_{ij}.\end{aligned} \tag{13}$$

In (13), θ_{ij} refers to the polar angle if node j is in allusion to node i, while $\cos\theta_{ij}/r_{ij}$ and $\sin\theta_{ij}/r_{ij}$ are weight function, where $\theta_{ij} = \arctan(y/x)$, $r_{ij} = \sqrt{x^2 + y^2}$.

(5) Accelerated Velocity. In sensor networks, nodes are moving all the time. Yet, as in fluid models, the movement is not in

constant velocity, which may also present accelerated velocity that is similar to fluid elements, that is, du/dt, dv/dt.

(6) Viscosity. Assuming that viscosity is variable, such viscosity is related to the average distance between a certain node and its neighbor nodes. Larger average distance leads to higher viscosity, while smaller distance leads to lower viscosity. As for this, viscosity may be defined as follows:

$$\mu_i^t = \frac{\overline{r}_{ij}}{R_c}. \tag{14}$$

In the equation, μ_i^t refers to the viscosity of node i at the moment of t; $\overline{r}_{ij}$ is the average distance between the node and its neighbor nodes.

4.3. Solving of Equation. Velocity component may be briefly and approximately transformed into

$$u_i^{t+\Delta t} = u_i^t + \left(\frac{\partial u_i^t}{\partial t}\right)\Delta t. \tag{15}$$

Putting (10) into (15), after a minor time interval Δt, the value of velocity component may be described as follows:

$$\begin{aligned} u_i^{t+\Delta t} &= u_i^t + \left(\frac{du}{dt} - \left(u_i^t\frac{\partial u_i^t}{\partial x} + v_i^t\frac{\partial u_i^t}{\partial y}\right)\right)\Delta t, \\ v_i^{t+\Delta t} &= v_i^t + \left(\frac{dv}{dt} - \left(u_i^t\frac{\partial v_i^t}{\partial x} + v_i^t\frac{\partial v_i^t}{\partial y}\right)\right)\Delta t. \end{aligned} \tag{16}$$

In the formula, du/dt and dv/dt may be solved with the method of smoothed particle hydrodynamics (SPH) [15]. By converting the gradient term on the right side of equal sign in Navier-Stokes momentum equation with SPH approximation method, we may be able to get the following equation:

$$\begin{aligned} \frac{dv_i^\alpha}{dt} &= f^\alpha - \sum_{j=1}^{N} m_j\frac{p_i + p_j}{\rho_i\rho_j}\frac{\partial W_{ij}}{\partial x_i^\alpha} \\ &\quad + \sum_{j=1}^{N} m_j\frac{\mu_i\varepsilon_i^{\alpha\beta} + \mu_j\varepsilon_j{}^{\alpha\beta}}{\rho_i\rho_j}\frac{\partial W_{ij}}{\partial x_i^\alpha}. \end{aligned} \tag{17}$$

In (17), μ refers to the viscosity, ε is shear stress rate in fluid, f stands for acting force, ρ is the density, p represents pressure intensity, the Greek letter superscripts α and β are designed to identify coordinate direction, and W refers to smoothness index, as is shown in the following equation:

$$W(R,h) = \frac{2}{\pi h^2}\left(\frac{3}{16}R^2 - \frac{3}{4}R + \frac{3}{4}\right) \quad (0 \le R \le 2), \tag{18}$$

where $R = r/h = |x - x'|/h$, r is the distance between two points or particles on x and x', and h refers to the smooth distance, that is, communication radius. Moreover, ε approximate expression of particle i shall be

$$\begin{aligned} \varepsilon_i^{\alpha\beta} &= \sum_{j=1}^{N}\frac{m_j}{\rho_j}v_j^\beta\frac{\partial W_{ij}}{\partial x_i^\alpha} + \sum_{j=1}^{N}\frac{m_j}{\rho_j}v_j^\alpha\frac{\partial W_{ij}}{\partial x_i^\alpha} \\ &\quad - \left(\frac{2}{3}\sum_{j=1}^{N}\frac{m_j}{\rho_j}v_j\cdot\nabla_i W_{ij}\right)\delta^{\alpha\beta}. \end{aligned} \tag{19}$$

As for this, du/dt, dv/dt may be worked out. In the computation, du/dt, dv/dt, V_i^t, ρ_i, p, $\partial\xi_i^t/\partial x$, $\partial\xi_i^t/\partial y$ may be put into (19) accordingly. As for this, the value of velocity component may be figured out by iterative computation.

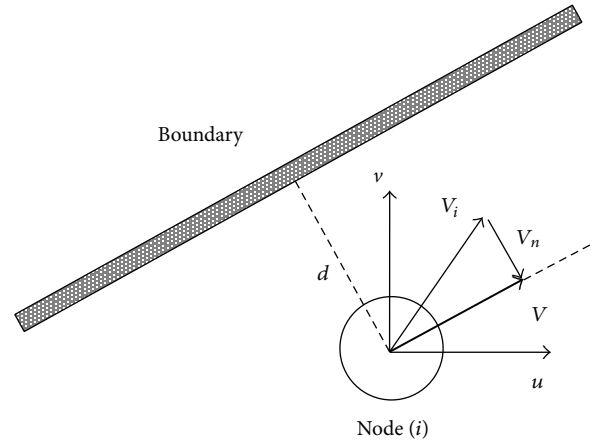


Figure 2: Object surface boundary condition.

4.4. Constraint Condition, Initial Condition, and Physical Boundary Condition

(1) Initial Value. Time iteration method in (16) needs to acquire the initial value of velocity component in advance, as well as current position before calculating the position of the next step.

(2) Physical Boundary Condition. The velocity of nodes sticking closely to object surface is tangent to object surface; that is, flow on object surface is tangent to object surface. As is shown by Figure 2, if the distance from nodes to obstruction or boundary is smaller than d, the moving velocity of node is changed.

(3) Constraint and Control Condition. Denoting the threshold value of velocity as V_{th}, the velocity of sensor node shall be controlled below the threshold.

4.5. Deployment Algorithm. Process of the algorithm is shown in Algorithm 1.

5. Simulation and Results

The section is designed to simulate the deployment algorithm under two different situations: with obstruction and without

(1) Partial parameters in NS equation are to be calculated according to the communication range, and distance between neighbor nodes.
(2) SPH method is adopted to work out the accelerated velocity of each node along each direction.
(3) Figuring out the velocity of all nodes;
(4) Checking if the node velocity is larger than the threshold value v_{th}; if so, assuming the velocity of node as v_{th};
(5) Working out the position of node in the next Step;
(6) Judging if the distance from node to obstruction or boundary is smaller than the threshold value d; if so, changing the velocity according to boundary condition;
Turning to Step 5;
(7) Position of mobile nodes
(8) Judging if network coverage rate meets the requirement; if not, turning to Step 1;
(9) Completion of the deployment

Algorithm 1: Node deployment algorithm based on flow field model.

obstruction, with initial status node deployment diagram, final status node deployment diagram, coverage rate changing curve, and uniformity changing curve presented.

5.1. Coverage. Generally, coverage can be considered as the measured service quality of a sensor network. In research on multirobot system, Gage firstly proposed the concept of coverage degree [16]. Literature [17] defines coverage rate as the specific value between the total coverage area of all nodes and the total area of the target region, as is shown by

$$\text{Coverage} = \frac{\bigcup_{i=1,\ldots,N} A_i}{A}, \tag{20}$$

where A_i is the area covered by the ith node, N is the total number of nodes, and A stands for the area of the region.

5.2. Uniformity. Good uniformity is a perfect standard to measure the service life of a network. Literature [17] defines uniformity as the standard deviation of distance between nodes. Smaller standard deviation leads to higher uniformity of network coverage. Consider

$$U = \frac{1}{N}\sum_{i=1}^{N} U_i,$$
$$U_i = \left(\frac{1}{K_i}\sum_{j=1}^{K_i}\left(D_{i,j} - M_i\right)^2\right)^{1/2}, \tag{21}$$

where N is the total number of nodes. K_i is the number of neighbors of the ith node, $D_{i,j}$ is the distance between ith and jth nodes, and M_i is the mean of internodal distances between the ith node and its neighbors.

In the calculation of the local uniformity U_i at the ith node, only neighboring nodes that reside within its communication range are considered. Uniformity measure is a local measure and is computed locally because each node has access to local information only. A smaller value of U means that nodes are more uniformly distributed.

Table 1: Simulation Parameters.

Parameters	Values
Region A	20 m × 20 m
Sensor numbers (N)	100
Communication radius (Rc)	2 m
Sensoring radius (Rs)	1 m
Distance threshold (d)	0.5 m
Simulation time step (Δt)	0.05
Velocity threshold (v_{th})	2
Pressure (p)	3

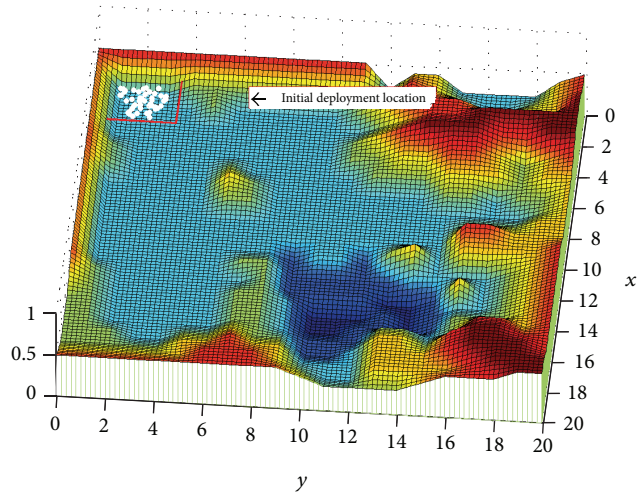


Figure 3: Node deployment environment and initial position.

5.3. Results. In practical application, owing to weather factor, battery failure, or other reasons, some nodes may be neutralized. Failure nodes may reduce the coverage rate of certain regions. When this happens, the balance of node deployment algorithm based on flow field model may be damaged, urging nodes to be relocated to recover these "exposed" regions, so as to regain a new balanced state.

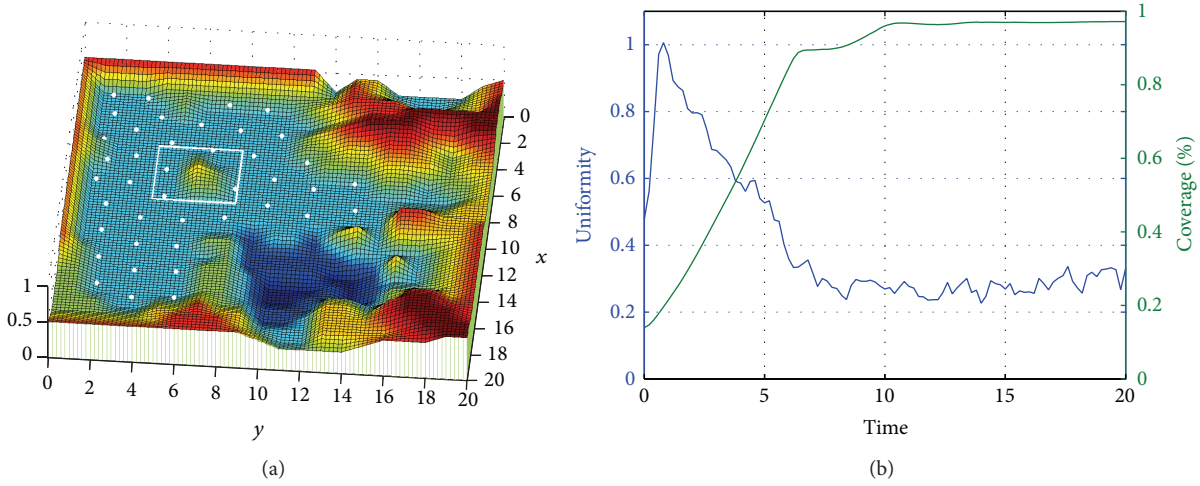


Figure 4: (a) Node position after deployment and (b) network coverage rate and uniformity Curve.

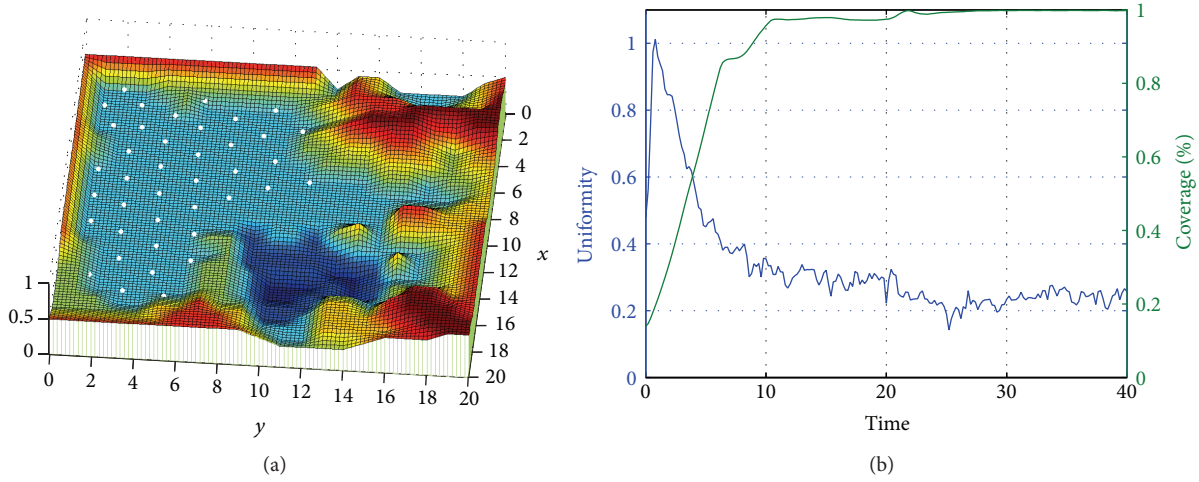


Figure 5: (a) Redeployment effect after the massif disappears and (b) network coverage rate and uniformity curve.

In this paper, 3D topographic map is applied to vividly describe the simulation process. Simulation parameters are shown in Table 1.

As is shown by Figure 3, this is a 3D topographic map simulated by computer software Matlab. In the map (Region A), there are mountainous area (in red), lake (in dark blue), and plain land. The initial deployment position of nodes is a certain corner in the map, that is, the range shown by red rectangle. The white dots are mobile nodes to be deployed.

Within the sensing radius R_s, nodes are able to detect the environment and to collect information. Similarly, they are also able to sense acting forces from neighbor nodes. The communication radius of nodes is larger than sensing radius R_s. Nodes are enabled to freely and mutually exchange information within the communication radius. The initial position of nodes is located at the lower left corner of the map. When the deployment begins, affected by unbalanced resultant force, all nodes will move to other positions in Region A.

Figure 4 shows the deployment result based on the algorithm proposed in this paper. According to the figure, nodes are distributed in plain lands, avoiding high mountains, massifs, lakes, and other obstructions. In the figure, the white rectangle shows a massif, while nodes are located around it. Seeing from the left plot, owing to the obstruction, nodes have to evade to complete the deployment. This is quite pragmatic in practical application. Dangers or inaccessible areas are avoided to reduce unnecessary node loss and waste, demonstrating the self-adaption nature of the algorithm. The right plot shows variation of coverage and uniformity in the deployment process. It may be discovered that, owing to the existence of obstruction, nodes are unable to fully cover the whole region, so that the maximum value of coverage rate is smaller than 1.

In Figure 5, massif mentioned in Figure 4 is manually removed to assess the self-adaptability of nodes. According to the figure, nodes' balance state is damaged, and nodes are relocated, reaching a new balance state. After the massif

disappears, there is no coverage gas left. Nodes around successfully recovered the area.

Seeing from the figure, there are nodes moving towards the place. After the adjustment, nodes regained the balance. Referring to coverage and uniformity curve in the right plot, we may clearly understand the process. When time = 20 (pointed by the arrow in the figure), the coverage rate increases from below 1 to 1. In the meanwhile, uniformity is also sharply improved. All these consequences resulted from the disappearance of massif.

After the redeployment (20 < time < 40), nodes reach the balance state again. The value of coverage increases to approximately 1, while the value of uniformity is stabilized around 0.2, which is slightly lower than the previous stable value −0.4. Thus, the disappearance of massif improves the uniformity.

Based on the above researches on deployment algorithm, this algorithm is properly simplified based on predecessors' research findings, eliminating the complexity of previous similar algorithms while preserving the property of self-adaption.

6. Conclusion

Deployment of large-scale sensor network is always a research hot spot in the field. How to save time and energy, to uniformly deploy nodes, so as to maximize service life of network? This is a goal pursued by all deployment algorithms. However, in practical application, problems and difficulties encountered are complicated and diversified. As for this, an extensible, self-adaptive, robust, and simple deployment algorithm is necessary.

In this paper, viscous fluid model is applied in deployment of wireless sensor network, with an extensible sensor network deployment algorithm proposed. This algorithm abstracts sensor nodes as fluid particles, while the parties abide by the motion rules of fluid. Nodes simulate the diffusion and self-propagating behavior of particles in fluid, realizing effective and ideal coverage range, so as to complete network deployment in unknown area. Shown by the simulation test, the algorithm shows good performance in various situations. For completely unknown deployment area, the algorithm proposed in this paper may fully bring to play its superiority of self-adaption, so as to reach the expected deployment effect. Thus, it is a robust and self-adaptive deployment algorithm.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. Howard, M. J. Matarić, and G. S. Sukhatme, "Mobile sensor network deployment using potential fields: a distributed, scalable solution to the area coverage problem," in *Distributed Autonomous Robotic Systems 5*, pp. 299–308, Springer, Tokyo, Japan, 2002.

[2] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization based on virtual forces," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications (INFOCOM '03)*, vol. 2, pp. 1293–1303, IEEE Societies, April 2003.

[3] B. Shucker and J. K. Bennett, "Scalable control of distributed robotic macrosensors," in *Distributed Autonomous Robotic Systems*, vol. 6, pp. 379–388, Springer, Berlin, Germany, 2007.

[4] S. Toumpis and L. Tassiulas, "Packetostatics: deployment of massively dense sensor networks as an electrostatics problem," in *Proceeding of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, vol. 4, pp. 2290–2301, IEEE, March 2005.

[5] X. Yu, W. Huang, J. Lan, and X. Qian, "A novel virtual force approach for node deployment in wireless sensor network," in *Proceedings of the 8th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS '12)*, pp. 359–363, Hangzhou, China, May 2012.

[6] J. Li, B. Zhang, L. Cui, and S. Chai, "An extended virtual force-based approach to distributed self-deployment in mobile sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2012, Article ID 417307, 14 pages, 2012.

[7] I. Larrabide, M. Kim, L. Augsburger, M. C. Villa-Uriol, D. Rüfenacht, and A. F. Frangi, "Fast virtual deployment of self-expandable stents: method and *in vitro* evaluation for intracranial aneurysmal stenting," *Medical Image Analysis*, vol. 16, no. 3, pp. 721–730, 2012.

[8] C. Sergiou and V. Vassiliou, "Estimating maximum traffic volume in wireless sensor networks using fluid dynamics principles," *IEEE Communications Letters*, vol. 17, no. 2, pp. 257–260, 2013.

[9] M. Gribaudo, C.-F. Chiasserini, R. Gaeta, M. Garetto, D. Manini, and M. Sereno, "A spatial fluid-based framework to analyze large-scale wireless sensor networks," in *Proceedings of the International Conference on Dependable Systems and Networks (DSN '05)*, pp. 694–703, July 2005.

[10] K. Wesley, D. Spears, W. Spears, and D. Thayer, "Two formal gas models for multi-agent sweeping and obstacle avoidance," in *Formal Approaches to Agent-Based Systems*, vol. 3228 of *Lecture Notes in Computer Science*, pp. 111–130, Springer, Berlin, Germany, 2005.

[11] M. R. Pac, A. M. Erkmen, and I. Erkmen, "Scalable self-deployment of mobile sensor networks: a fluid dynamics approach," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pp. 1446–1451, Beijing, China, October 2006.

[12] M. R. Pac, A. M. Erkmen, and I. Erkmen, "Towards fluent sensor networks: a scalable and robust self-deployment approach," in *Proceedings of the 1st NASA/ESA Conference on Adaptive Hardware and Systems (AHS '06)*, pp. 365–372, June 2006.

[13] D. Anderson J, *Computational Fluid Dynamics*, McGraw-Hill, New York, NY, USA, 1995.

[14] G. R. Liu and M. B. Liu, *Smoothed Particle Hydrodynamics: A Meshfree Particle Method*, World Scientific, 2003.

[15] J. J. Monaghan, "Smoothed particle hydrodynamics," *Reports on Progress in Physics*, vol. 68, no. 8, pp. 1703–1759, 2005.

[16] W. G. Douglas, "Command control for many-robot systems," *Unmanned Systems*, vol. 10, no. 4, pp. 28–34, 1992.

[17] N. Heo and P. K. Varshney, "A distributed self spreading algorithm for mobile wireless sensor networks," in *IEEE Wireless Communications and Networking (WCNC '03)*, vol. 3, pp. 1597–1602, 2003.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 835260, 11 pages
http://dx.doi.org/10.1155/2014/835260

Research Article

Utility-Oriented Placement of Actuator Nodes with a Collaborative Serving Scheme for Facilitated Business and Working Environments

Chi-Un Lei,[1] Woon Kian Chong,[2] and Ka Lok Man[3,4]

[1] *Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong Island, Hong Kong*
[2] *International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*
[3] *Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*
[4] *Yonsei University, Seoul 120-749, Republic of Korea*

Correspondence should be addressed to Chi-Un Lei; culei@eee.hku.hk

Received 27 March 2014; Accepted 17 June 2014; Published 2 July 2014

Academic Editor: Xiaoxuan Meng

Places to be served by cyber-physical systems (CPS) are usually distributed unevenly over the area. Thus, different areas usually have different importance and values of serving. In other words, serving power can be excessive or insufficient in practice. Therefore, actuator nodes (ANs) in CPS should be focused on serving around points of interest (POIs) with considerations of "service utility." In this paper, a utility-oriented AN placement framework with a collaborative serving scheme is proposed. Through spreading serving duties among correctly located ANs, deployment cost can be reduced, utility of ANs can be fully utilized, and the system longevity can be improved. The problem has been converted into a binary integer linear programming optimization problem. Service fading, 3D placements, multiscenario placements, and fault-tolerant placements have been modeled in the framework. An imitated example of a CPS deployment in a smart laboratory has been used for evaluations.

1. Introduction

Cyber-physical systems (CPS) are systems that contain numerous distributed, linked, and autonomously operated sensor nodes (SNs) and actuator nodes (ANs) [1]. Generally, CPS are used to gather adequate information about the physical environment, manipulate cyber/physical/information quantities, and, eventually, provide useful and prompt services for people [2–6]. In order to provide services, ANs interact with the physical environment by means of transducers and actuators.

In order to ensure coverage and credibility of the service, as well as minimizing resources consumed by nodes in CPS, nodes should be placed in correct positions through planning. Comparing to SNs, ANs are usually equipped with better capabilities of control, computation, and action as well as larger battery capacity. In addition, requirements and physical interaction of sensing and serving are different. Therefore, existing placement algorithms for SNs [7] are inapplicable to be directly adopted for ANs.

Placement of actor nodes (e.g., mobile robots and unmanned aerial vehicles) and mobile ANs has been explored in the literature [8–10]. Since these nodes can travel to the site of the event without obstructions, existing algorithms do not consider the uncertainty in the ability of actuators to deliver adequate service due to disturbance or fading in harsh environments. However, ordinary, small-size, and low-cost ANs are usually located in fixed positions and have different behaviors of motions and physical interactions. Therefore, algorithms for placement of actor nodes are also inapplicable for placements of ANs. Thus, specific placement frameworks for placement of ANs are needed.

Meanwhile, service coverage, system connectivity, and system longevity are typical metrics for node placement problems [7]. However, we believe that ANs in CPS should be focused on achieving a reliable and accurate serving around

points of interest (POIs) with considerations of a "service utility" metric, because of the following.

(i) Places/objects to be concerned/served (e.g., targets in the battlefield) or crucial occurrences of an event are usually distributed unevenly over the area. Thus different areas usually have different importance and values of serving [11].

(ii) Same type of ANs are often used in an application, but serving environments and requirements of each POI can be versatile and different [12]. As a result, serving power can be excessive or insufficient in practice. Thus, cost-effective approaches should be designed to spread serving duties among ANs.

(ii) Balanced serving workload of ANs can prevent overloading of particular ANs beyond their limit output performance. This can avoid aging/deteriorating of equipment on ANs and, thus, improve the system longevity.

These concerns are significant yet have not been explored thoughtfully by researchers. However, in the algorithm, partial serving of POIs and utilization of redundant residual utility in ANs have not been taken into account. It is undesirable due to setup and operation cost of unnecessary ANs. Thus, it is worthwhile to develop a CPS design framework that can fully utilize the capability of ANs via a collaborative serving scheme. However, this issue does not seem to have been explored in great detail by researchers; yet, it is significant not only to many indoor (business) environments, including offices, hospitals, laboratories, schools, business facilities, and factory workshops.

Based on the preliminary study [13], a relaxed utility-oriented AN placement framework is proposed in this paper. In the framework, utility of ANs is fully utilized and the fidelity of providing service to POIs is improved. To summarize, contributions of the framework are as follows.

(1) The utility-oriented placement problem has been converted into a binary integer linear programming problem. In addition, a recursive framework through the refinement of the search space has been proposed for finding a more suitable configuration.

(2) Serving requirements of POIs, serving capabilities of ANs, features of delivered service, and status of serving activities have been modeled in a unified framework through a generic description. In particular, geographical attributes and obstacles can be modeled as service fading, for a more realistic modeling for deployment.

(3) The proposed framework can be applied to placements for fault-tolerant serving, such that POIs are still partially served even if some ANs are damaged.

(4) The proposed framework can be used for multiscenario placements, such that the system can change its serving characteristics according to the dynamic nature of the environment, without relocating ANs or consuming unnecessary energy for actions.

(5) The proposed framework can be used for placements in three-dimensional (3D) fields, such that ANs can be placed on tall objects (e.g., buildings or trees) or for airborne/underwater serving.

(6) An imitated example of a CPS deployment in a smart laboratory has been used for evaluations of the proposed framework.

The paper is organized as follows. Search space and derived objective function as well as constraints for the single-step placement are presented in Section 2. Generalizations for recursive placement are shown in Section 3, respectively. Finally, the performance of the framework is evaluated by examples in Section 4.

2. Single-Step Placement of Actuator Nodes

The objective of the framework is to pick a set of ANs and assign them to serve a set of POIs. In other words, the framework has to determine the location and service coverage radius of deployed ANs and distribution of serving workload among ANs, with a minimized total cost for setup and operation. Here, we assume that ANs can serve different POIs with different amount of utility at the same time.

The placement problem is formulated as a binary integer linear programming (BILP) problem with a cost objective function and constraints that models different concerns in the deployment of ANs. For ease of explanations, the single-step optimization framework is first described in this section. Then, the recursive framework with a refined search space is described in Section 3.

2.1. Search Space in the Optimization. In the framework, I AN candidates can be selected to serve J POIs. For ease of explanations, A_i denotes the ith candidate of AN and P_j denotes the jth POI, where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. The Cartesian coordinates of A_i and P_j in 2D fields are $\{x_{A_i}, y_{A_i}\}$ and $\{x_{P_j}, y_{P_j}\}$, respectively. In 3D fields, the Cartesian coordinates of A_i and P_j are $\{x_{A_i}, y_{A_i}, z_{A_i}\}$ and $\{x_{P_i}, y_{P_i}, z_{P_i}\}$, respectively.

The placement problem without mesh discretization is an integer nonlinear programming problem. To simplify computations, the infinite search space of AN locations is replaced by a search space with preselected finite elements. All AN candidates are located at the grid points, in which the grid size depends upon the accuracy of placement desired. Since the search space grows exponentially with the increase in the number of points, irrelevant candidates are neglected in the optimization problem. This is further explained in Section 2.3.

Through the discretization, the problem becomes a BILP problem. Thus, branch and bound (BB) algorithm can be used to decompose the BILP problem into linear programming (LP) problems. Optimality of the BILP solution in a finite search space is guaranteed by combining the BB framework and the simplex algorithm in LP problems. However, BILP problems are NP-complete [14, 15]. Therefore, only suboptimal solutions can be efficiently obtained via heuristic methods, such as tabu search and simulated annealing. However, it is out of the scope of this paper.

2.2. Cost Objective Function. The objective of the framework is to design a CPS that can serve all POIs and minimize total cost for setup and operations. The cost can be described by the following expression:

$$\min \sum_{i=1}^{I} E \times Y_i + r_i, \tag{1}$$

where Y_i is the binary selection parameter of A_i; that is, $Y_i = 1$ indicates candidate A_i is selected in the placement. Furthermore, E is the setup cost of an AN and r_i is the service coverage radius of A_i. Generally $E := 100$, such that all POIs are served with the minimum number of ANs. Meanwhile, $\{r_i\}$ should be minimized because ANs with a smaller coverage radius can use less power to serve, which leads to a higher energy efficiency and longevity of ANs and the system. Coverage radius can be adjusted through control units and actuators on ANs.

In some situations, for example, early design stages, we only need to find out the minimum number of required ANs for serving ($I_{\min}$), while locations and serving coverage of ANs are unnecessary. In this case, expression (1) can be replaced by a simplified objective function for a quick cost analysis:

$$\min \sum_{i=1}^{I} Y_i. \tag{2}$$

In the second step, specifications of ANs can then be optimized for the best system performance and largest system longevity by the following objective function:

$$\min \sum_{i=1}^{I} r_i \quad \text{with} \sum_{i=1}^{I} Y_i = I_{\min}. \tag{3}$$

2.3. Constraints on the Coverage of Services. In the framework, inspired by Elfes model [16], a relaxed disc service coverage zone centered at A_i is adopted in the framework. In the framework, P_j can be served by A_i if it is within the coverage of A_i, while the serving efficiency depends on the distance between P_j and A_i (further discussions about service fading are given in Section 2.4). This restriction can be described by the following constraint:

$$r_i \geq d_{i,j} \times X_{i,j}, \tag{4}$$

where $X_{i,j}$ is the binary connection parameter between A_i and P_j; that is, $X_{i,j} = 1$ indicates P_j is served by A_i. $d_{i,j}$ is the Euclidean distance between A_i and P_j. In 2D fields, $d_{i,j} = \sqrt{(x_{A_i} - x_{P_j})^2 + (y_{A_i} - y_{P_j})^2}$, while in 3D fields, $d_{i,j} = \sqrt{(x_{A_i} - x_{P_j})^2 + (y_{A_i} - y_{P_j})^2 + (z_{A_i} - z_{P_j})^2}$.

In practice, ANs have a limited service coverage radius, which are based on the capability and power of circuits and devices on ANs. For simplicity, maximum service coverage radius of all ANs is $r_{\max}$. Furthermore, if A_i is not selected, $r_i := 0$. These requirements can be modeled as

$$0 \leq r_i \leq r_{\max} \times Y_i. \tag{5}$$

Meanwhile, if P_j is within the range of A_i, then P_j must be assigned to A_i. This can be modeled by the following constraint:

$$r_i - d_{i,j} \leq r_{\max} \times X_{i,j}. \tag{6}$$

In order to reduce the optimization problem, unnecessary AN candidates are neglected. For example, it is infeasible for an AN to serve a POI which is farther than its serving distance. Therefore, if $d_{i,j} > r_{\max}$ for any i and j; $X_{i,j} := 0$ and $q_{i,j} := 0$. Meanwhile, it is also not practical for an AN located at the same position of POIs or obstacles. Therefore, if A_i is located at the same position of POIs (i.e., $d_{i,j} = 0$) or obstacles, $Y_i := 0$.

2.4. Constraints on Consumption and Generation of Utilities with Collaborative Serving and Service Fading. Insufficient serving power to POIs may fail to provide guaranteed services and eventually deteriorate the performance of the system. Thus, it is necessary to consider the rate/amount of utilities that can be provided by ANs (utility budget) and rate/amount of utilities that are consumed by POIs. To be specific, each POI should receive a collection of required utility u_j by fusing service provided by multiple ANs. In other words, A_i contributes part of its utility budget to serve P_j, and P_j is served by contributions from several ANs. Comparing to [13], by collaborative serving, less ANs are needed to serve all POIs. An example is shown in Figure 1. Assume that after serving a POI, the residual utility of an AN is not able to serve another POI. However, if collaborative serving is allowed, the middle POI can collect enough utilities that are residual utility of its two surrounding ANs. Therefore, one AN is saved for serving. Collaborative serving can be described by the following expression:

$$C \times \left(\sum_{i=1}^{I} q_{i,j} \right) \geq u_j, \tag{7}$$

where u_j is the total consumed utility of P_j, $q_{i,j}$ is the proportion of utility budget of A_i that contributes to P_j, and C is the maximum utility budget of an AN that can serve to surrounding POIs. Meanwhile, serving capability of every AN is limited by its utility budget. Therefore, the following inequality is required in the problem:

$$\sum_{j=1}^{J} q_{i,j} \leq 1. \tag{8}$$

If part of the utility budget in A_i is used to serve P_j, A_i is selected for serving and it is responsible to serve P_j. These can be described by the following inequalities:

$$\sum_{j=1}^{J} q_{i,j} \geq Y_i,$$
$$X_{i,j} \geq q_{i,j}. \tag{9}$$

In practice, the strength of service power usually decays when the service is delivered from the AN to the POI. In other words, an AN can serve several POIs that are close to the

(a) (b)

Figure 1: Serving of POIs: (a) if collaborative serving is allowed and (b) if collaborative serving is not allowed.

AN, but it can only serve one POI that is far from the AN. In order to describe this phenomenon, (7) can be replaced by the following constraint:

$$C \times \left(\sum_{i=1}^{I} q_{i,j} \times \left(\frac{f_{i,j}}{d_{i,j} + g} \right)^2 \right) \geq u_j, \tag{10}$$

where $f_{i,j}$ is the factor for path loss of the delivered service from A_i to P_j and g is the factor for the operation loss of an AN. $\{f_{i,j}\}$ and g and $\{u_j\}$ and C can be determined through product specifications, theoretical estimations, or experimental measurements. In brief,

(i) $\{u_j\}$ depends on scenarios of serving. In particular, it depends on accuracy, reliability, relevance, timelessness, and usability of service. It also depends on whether the service is periodic or event-driven. In general, $\{u_j\}$ should be higher for critical POIs, such that allocated AN(s) will pay more attention to these POIs. For example, the POI may require a larger $\{u_j\}$ if there is a higher possibility of making a significant damage, due to the severity of crashes and shocks.

(ii) $\{f_{i,j}\}$ depends on features and decay rate of the service and the ambient environment as well as the disturbance between ANs and POIs (e.g., trees and walls). In general, ANs can serve targets that lie in their line of sight; therefore, $\{f_{i,j}\}$ should be larger; on the other hand, obstacles may make the service unreachable; therefore, $\{f_{i,j}\}$ should be smaller.

(iii) C and g depend on the battery capacity, power dissipation and characteristics of actuators, onboard circuitry, and peripherals on ANs.

The proposed generic constraint can be further generalized by including specific considerations [17, 18], such as battery models with energy harvesting capabilities [17]. However, the discussion of utility generalizations is beyond the scope of the paper.

2.5. Constraints for Fault-Tolerant Serving and Serving with Limited Recipients per AN. Generally, every POI is served by the minimum number of ANs. However, due to the impact from environment hazards and physical shock, ANs may be damaged or even failed to operate after deployment. In these dynamic situations, the system is expected to tolerate failures of some ANs and communication link as well as guaranteeing a proportion of serving ability to POIs. Introducing adequate redundant ANs can surely ensure the connectivity. However, it also requires more installed ANs, which can be undesirable due to the cost. In our framework, fault-tolerant serving can be done by ensuring each POI is served by at least K AN(s), such that at least POIs are partially served by ANs if minorities of ANs are broken. This design restriction can be described by the following connection constraint, for any P_j:

$$\sum_{i=0}^{I} X_{i,j} \geq K. \tag{11}$$

In the framework, we assume that ANs can serve unlimited POIs at the same time. However, sometimes this assumption becomes impractical because of the hardware capabilities of ANs. In these situations, we need to ensure that each AN can only serve at most L POIs. This design restriction can be described by the following connection constraint, for any A_i:

$$\sum_{j=0}^{J} X_{i,j} \leq L. \tag{12}$$

2.6. Multiscenario Placement through a Relaxation of Constraints. In the proposed framework, ANs are placed in fixed locations and serve POIs according to a single scenario. However, serving patterns and requirements can be changed according to the pattern of events as well as application-level interests. For example, requirements of air ventilation during daytime and night-time can be different. In such circumstances, a fixed arrangement is not capable to consider dynamic changes for a better delivery of services. Therefore, in order to improve the feasibility of meeting timeliness requirements, ANs' service should be rescheduled without node relocation. Through the proposed generalization, the framework can place ANs with considering multiple serving scenarios.

Assume that there are K serving scenarios, in order to distribute serving duties to ANs for each scenario correctly, $\{X_{i,j}\}$ and $\{q_{i,j}\}$ in (7)–(11) are replaced by $\{X_{i,j,k}\}$ and $\{q_{i,j,k}\}$, where $X_{i,j,k}$ is the binary connection parameter $X_{i,j}$ and $q_{i,j,k}$ is the proportion parameter $q_{i,j}$, for the kth scenario, respectively, for $k = 1, \ldots, K$. Meanwhile, in order to determine the coverage radius of ANs, $\{X_{i,j}\}$ are used as usual for (4)–(6). Furthermore, A_i is selected even if A_i is only selected in one of the scenarios; that is, $X_{i,j} = 1$ if $X_{i,j,k} = 1$ for any k, and $X_{i,j} = 0$ if $X_{i,j,k} = 0$ for all k. Therefore, determination of $\{X_{i,j}\}$ is enforced by the following linear inequality constraints, such that the generalized problem can still be solved by a BILP optimization process:

$$X_{i,j} \geq X_{i,j,k}, \quad \text{for } k = 1, \ldots, K, \tag{13}$$

$$X_{i,j} \leq \sum_{k=1}^{K} X_{i,j,k}. \tag{14}$$

(1) Construct initial search space $\{A_i\}$;
(2) **repeat**
(3) Determine $\{d_{i,j}\}$ via locations of $\{A_i\}$ and $\{P_j\}$
(4) **if** $I_{\min}$ is not given **then**
(5) Determine $\sum Y_i$ and $\{q_{i,j,k}\}$ through minimizing (1), subject to (4)–(14) for all i, j and K, via BILP;
(6) **else**
(7) Determine $\{r_i\}$ and $\{q_{i,j,k}\}$ through minimizing (3), subject to (4)–(14) for all i, j and K, via BILP, with given $I_{\min}$;
(8) **end if**
(9) Select $\{q_{i,j,k}\}$ and $\{r_i\}$ as the answer;
(10) Construct the refined search space $\{A_i\}$ from the determined answer;
(11) **until** $\left\|\{r_i\}_{\text{previous}} - \{r_i\}_{\text{current}}\right\|_2 / \left\|\{r_i\}_{\text{previous}}\right\|_2 \leq \varepsilon$

Algorithm 1: Pseudocodes of the recursive framework.

3. Recursive Placement via a Refined Search Space

The single-step placement framework described in Section 2 can be generalized into a recursive framework for a better placement. The algorithm pursues a divide-and-conquer strategy to split the overall placement problem into a series of placement problems with smaller size (in terms of area for placement), such that the single-step framework can be executed recursively.

We assume that if a set of ANs $\{A_{\text{selected}}\}$ is obtained through the single-step framework, a better placement can be obtained from the neighbourhood of $\{A_{\text{selected}}\}$. Here, neighbors of $\{A_{\text{selected}}\}$ are located within $[x_{A_{\text{selected}}} \pm (\Delta x_A/2), y_{A_{\text{selected}}} \pm (\Delta y_A/2)]$, where Δx_A and Δy_A are the resolution of the original search space. In the recursive framework, based on the location of $\{A_{\text{selected}}\}$, a new set of candidates can be generated. Meanwhile, through the new set of AN candidates, a new set of ANs with a better placement can be found. Thus, the placement can be repeated recursively until the replacement of ANs cannot further reduce the operation cost in terms of average $\{r_i\}$, which can be simplified as $\|\{r_i\}_{\text{previous}} - \{r_i\}_{\text{current}}\|_2 / \|\{r_i\}_{\text{previous}}\|_2 \leq \varepsilon$, where ε is the predefined tolerance. In that case, selected AN candidates and their positions are declared as the final answer and the algorithm terminates. In summary, pseudocodes of the recursive placement framework are shown in Algorithm 1.

4. Performance Evaluation of the Framework

In this section, examples are used to show the performance of the framework. Computations run in an optimization solver Lingo 11 [19] on a 6 GB-RAM Intel i7 3.4 GHz PC.

4.1. Placement in a 2D Laboratory Environment. The performance of the system is analysed via a simulated example of a school laboratory with dimensions 50 m × 33 m [20–22], as shown in Figure 2. In order to monitor and regulate ventilations and thermal comfort of the laboratory, a CPS with sensors and actuators (i.e., electrical fans) is installed. A total 15 points of interest (POIs) and their consumed utilities are assigned according to positions of vents and pattern of activities in the laboratory, as shown in Figure 3. In the example, service fading is excluded and single-step placement is used. Before the optimization, 100 candidates of ANs, which are located in areas bounded by outmost ANs, are generated. For every candidate, utility budget of an AN (C) is assumed to be 75 and $r_{\max}$ = 10 m. After computations, the framework suggests that seven ANs with average $\{r_i\}$ = 5.785 m are required to provide adequate requested services to all POIs. Results are shown in Figure 3(a). Figure shows that some ANs serve one critical POI and partially serve one noncritical POI, while others serve at most three noncritical POIs. Meanwhile, in some cases, one POI is served by two ANs. The problem contains 1036 variables (including 468 binary variables) and 4817 constraints, and the algorithm needs 30 minutes to obtain the optimal solution. Results with $r_{\max}$ = 1 m are also shown in Figure 3(c). Figure 3(c) shows that if $r_{\max}$ is small, ANs work independently without sharing their duties.

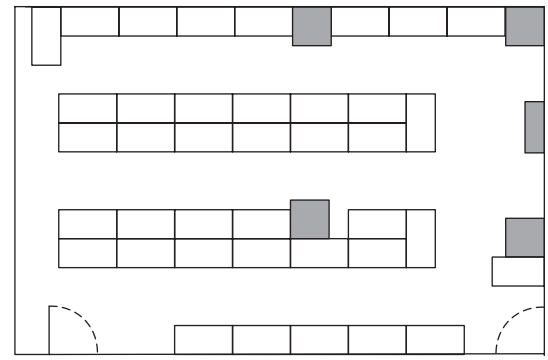


Figure 2: The floor-plan of the laboratory in the 2D field example. The grey squares and white rectangles denote wall structures and laboratory benches, respectively.

4.2. Placement via a Recursive Framework. The recursive framework is then applied to the example in Section 4.1. In the example, 100 AN candidates have been used for optimization in the first iteration. 25 (5 × 5) candidates have been generated for each AN. Since seven ANs are needed for placement, totally 175 candidates have been used for optimization after the first iteration. In every iteration,

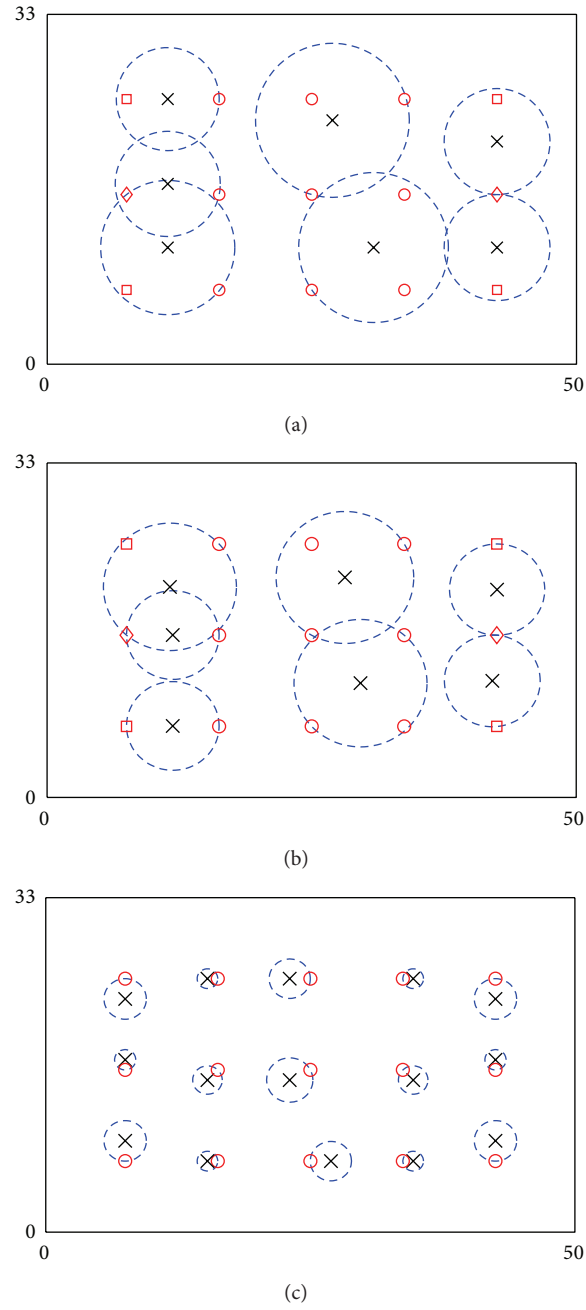


FIGURE 3: Locations of POIs and ANs in the 2D field example via the proposed framework: (a) optimization with $r_{max} = 10$ m, (b) optimization with $r_{max} = 10$ m after five iterations through recursive placement, and (c) optimization with $r_{max} = 1$ m. Crosses and dash circles denote ANs and service coverage of ANs. Circles, squares, and diamonds denote POIs with 25, 40, and 55 consumed utilities, respectively.

the algorithm terminates automatically after one hour if it cannot obtain an optimal solution. The obtained solution is then used to refine the search space for the next iteration. After five iterations, average coverage radius of ANs has been reduced by >8.98%. And the placement is shown in Figure 3(b). More details of results are shown in Table 1. Results show that recursive framework can improve the performance of ANs, with the expense of computation time.

4.3. Comparison with Other Strategies for Placements. Results of the proposed framework are compared with the following three strategies: (i) ANs are placed near the location of POIs one by one, with serving POIs with the highest utility consumption first and with utilization of residual utilities (Greedy approach, Strategy I), (ii) ANs are placed with considerations of geographical distribution of POIs only (Strategy II), and (iii) ANs are randomly placed with

Table 1: Results of the example via the recursive framework.

nth iteration	1	2	3	4	5
Average coverage radius (m)	5.7852	5.4523	5.2755	5.2706	5.2657

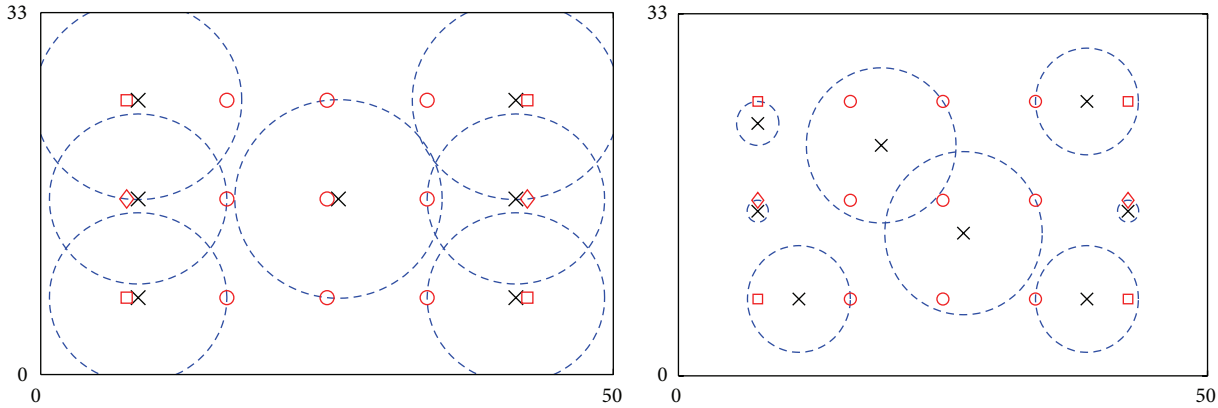


Figure 4: Locations of POIs and ANs in the 2D field example: (a) placement using Strategy I and (b) placement using Strategy II. Crosses and dash circles denote ANs and service coverage of ANs. Circles, squares and diamonds denote POIs with 25, 40, and 55 consumed utilities, respectively.

utilization of residual utilities (Strategy III). The example in Section 4.1 is used for evaluation. Results of the first two strategies are shown in Figure 4. Figures show that if Strategy I is used, seven ANs with average $\{r_i\} = 8.309$ m are needed to serve all POIs. Meanwhile, if Strategy II is used, the system serves all POIs by eight ANs with average $\{r_i\} = 3.891$ m. Furthermore, Strategy III has been tested ten times. Averagely, the system can serve all POIs by seven ANs with average $\{r_i\} = 11.546$ m. Therefore, the proposed framework outperforms the other three strategies in terms of the number of ANs (i.e., system cost) or average $\{r_i\}$ (i.e., operation efficiency and system longevity). This is because the proposed framework considers geographic distribution of POIs and utilization of residual utilities for placements at the same time. These considerations become even more significant if the number of ANs is limited. These situations are typical because ANs are usually expensive to deploy.

4.4. Placements with Different Configurations. Example in Section 4.1 is used for investigating the influence of various configurations. Results are shown in Table 2. From examples, we can arrive at a few conclusions about the result of placements as well as the efficiency of the framework.

(i) In general, the system requires fewer ANs as the utility budget (C) and maximum service coverage radius ($r_{\max}$) of ANs increase. For example, only one AN is needed to serve the whole room if the actuator has an extremely large C and $r_{\max}$. However, this assumption is usually not practical. Therefore, C and $r_{\max}$ should be configured carefully.

(ii) The space complexity of the optimization mainly depends on the number of POIs and number of AN candidates, while the time complexity mainly depends on the utility budget (C) and maximum service coverage radius ($r_{\max}$). This is because there are usually more combinations of duty distributions between ANs for a larger C or $r_{\max}$.

(iii) When the utility consumption is dominated by a few POIs or the utility budget of ANs is small, ANs tend to work independently without sharing job duty. On the other hand, when the utility consumption is evenly distributed, there is a higher possibility for ANs to share their workload. In this situation, the proposed framework can achieve a better performance.

(iv) If each AN can only serve at most two POIs due to limitations of hardware on ANs, through including constraint (12) for every ANs, the framework suggests that nine ANs with average $\{r_i\} = 4.238$ m are needed to serve all POIs.

4.5. Impact of Fading on the Delivery of Services. Example in Section 4.1 with different f and $g = 0.01$ is used for investigating the influence of service fading. Results are shown in Table 3. Results show that the number of required ANs increases as f decreases. It is because for a smaller f (i.e. a larger service power fading), POIs that are far from ANs consume more utility. In some situations, POIs that are too far from ANs cannot be served. As a result, additional ANs are added to serve these POIs. For example, if $f = 1.5$ m, workload cannot be shared among ANs; therefore, every POI is individually served by a nearby AN.

In the second example, an obstacle has been included, as shown in Figure 5. In the example, we assume the obstacle can effectively block the service and ANs can only serve POIs

Table 2: Results and computation complexity with respect to different problem configurations.

C	r_{max} (m)	Number of required ANs	Average $\{r_i\}$ (m)	CPU time (min.)
75	10	7	5.785	30
75	5	11	2.755	0.10
75	4.5	15	0.984	0.10
75	10	7	5.845	5
75	10	7	5.845	10
75	10	7	5.785	15
75	10	7	5.785	107 (Optimal)
50	10	10	5.288	60
25	10	20	2.5248	60
1000	25	1	21.865	0.58 (Optimal)

Table 3: Impact of service fading on the delivery of services with different f.

f (m)	Number of required ANs	Average coverage radius (m)	Average consumed utility (Scaled)	Average consumed utility (Original)
7.5	5	8.142	99.000	74.627
6.0	6	8.123	82.500	71.778
5.5	7	5.580	70.714	56.520
5.0	8	4.463	61.875	46.358
4.0	9	4.032	55.000	71.925
1.5	15	0.984	33.000	0.208

that lie in its line of sight. In order to model the obstacle, if the line of sight between A_i and P_j is blocked (i.e., the service is very difficult to be delivered from A_i to P_j), $f_{i,j} := 0.05$ m; otherwise, $f := 0.55$ m. Furthermore, some AN candidates are neglected since the obstacles have occupied spaces of some AN candidates. If the original placement is used, ANs are not able to serve some POIs. The new placement in Figure 5 shows that the system can still serve all POIs by placing ANs with similar configurations in different positions.

4.6. Multiscenario Placement. In this example, the system has to ensure the delivery of services in two scenarios: (i) the distribution of utility consumption in Section 4.1 is used as serving for office hours (Scenario A), and (ii) Figure 6(a) shows the distribution of serving for nonoffice hours (Scenario B). Figure 3(a) and Figure 6(a) show the placement of ANs if only Scenario A or B is considered, respectively. Meanwhile, Figure 6(b) shows the placement if both scenarios are considered for placements. The figure shows that 14 ANs with average $\{r_i\} = 5.174$ m are needed to serve all POIs. Therefore, the framework can be used for multiscenario placements without introducing extra ANs.

4.7. Fault-Tolerant Placement with a Large-Scale Example. In order to demonstrate the fault-tolerant placement, the example in Section 4.1 with fault-tolerant parameters $K = 2$ and $K = 3$ is used. Placements of ANs for these two situations are shown in Figure 7. Figures show that nine ANs with average $\{r_i\} = 7.5841$ m are necessary if each POI is served by at least two ANs (i.e., $K = 2$). Meanwhile, 14 ANs with average $\{r_i\} = 7.2357$ m are necessary if POIs are served by at least three ANs (i.e., $K = 3$). Figures show that in order to ensure POIs can be partially served by ANs, only a few extra ANs are needed. Therefore, the algorithm allows ANs to provide fault-tolerant service through a cost-effective approach.

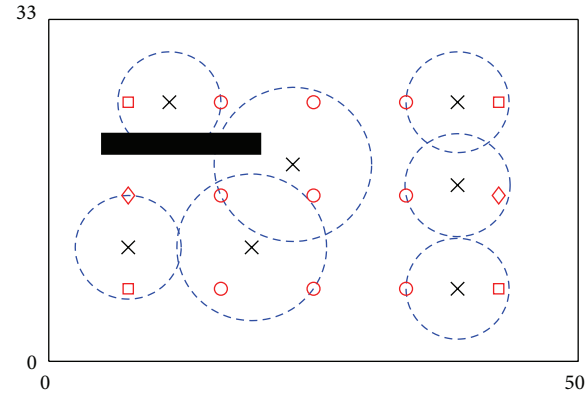


Figure 5: Locations of POIs and ANs in the 2D field example with the inclusion of an obstacle. Crosses and dash circles denote ANs and service coverage of ANs. Circles, squares, and diamonds denote POIs with 25, 40, and 55 consumed utilities, respectively. The black block denotes the obstacle.

The second example is a large-scale example with 21 POIs. Their location and consumed utility are randomly assigned. A search space with 400 candidates has been generated for optimization. If $K = 2$, the framework requires 130 minutes to obtain the optimal solution, and the result is shown in Figure 8. The example illustrates that the proposed framework can solve large-scale placement problems.

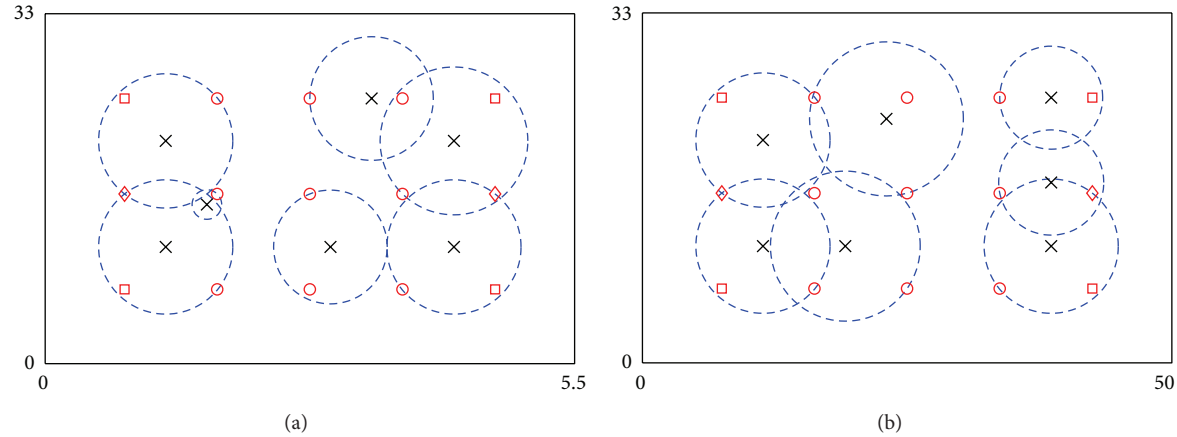


Figure 6: Locations of POIs and ANs in the 2D field example: (a) placement with Scenario B and (b) multiscenario placement with Scenarios A and B. Crosses and dash circles denote ANs and service coverage of ANs. Circles, squares, and diamonds denote POIs with 25, 40, and 55 consumed utilities in Scenario A and 35, 25, and 25 consumed utilities in Scenario B, respectively.

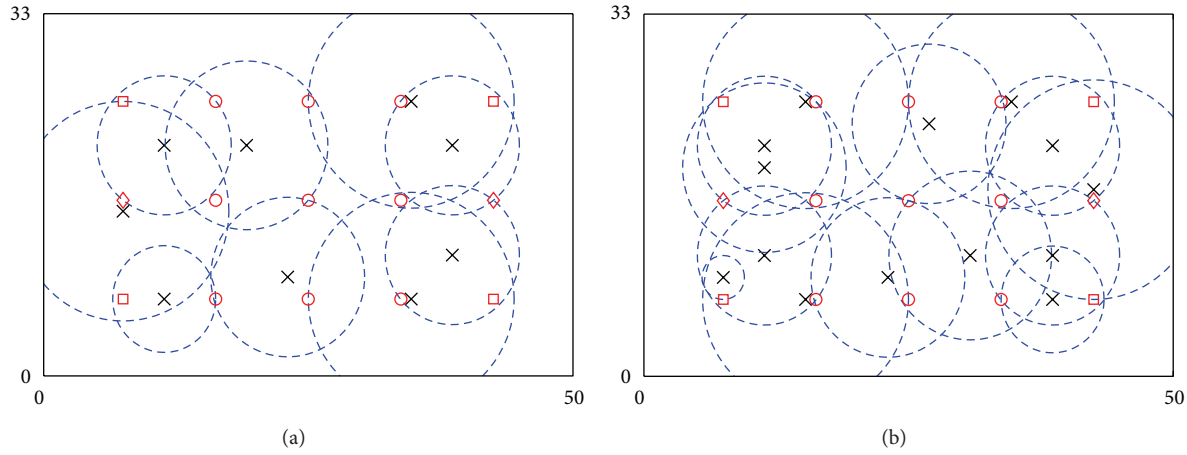


Figure 7: Locations of POIs and ANs in the 2D field example: (a) fault-tolerant placement with $K = 2$ and (b) fault-tolerant placement with $K = 3$. Crosses and dash circles denote ANs and service coverage of ANs. Circles, squares, and diamonds denote POIs with 25, 40, and 55 consumed utilities, respectively.

4.8. Placement in a 3D Field Environment. The performance of the framework is analysed via a simulated example of an arbitrary 3D environment with dimensions 80 m × 80 m × 80 m. A total of seven POIs are distributed randomly, with a random consumed utility ranging from 23 to 32. Totally, 512 candidates of ANs are distributed evenly in the 3D field. Through the optimization, six ANs with average $\{r_i\} = 7.594$ m are assigned to provide adequate requested services for all POIs. The framework needs 33 seconds to solve the problem. The example proves that the proposed framework can be used for placements of ANs in 3D fields.

5. Conclusion

An actuator node placement framework with collaborative sharing of utility has been presented. In particular, partial serving and utilization of redundant residual utility in ANs are allowed in the relaxed formulation. Service fading, 3D placements, multiscenario placements, and fault-tolerant placement have been modeled in the framework. The problem has been converted into a binary integer linear programming optimization problem, such that the optimal solution can be obtained in a discretized search space.

We believe that this research is not only confined to schools, hospitals, and factory workshops, but also to many similar business environments such as offices. These environments are important to create positive benefits to the society as well as business activities. If we are to develop smart societies, we need to look at systems that are easily deployable, noninvasive, economical, effective, and energy efficient. Our algorithm for actuator node placement shows we can satisfy all these conditions.

In the future, we will generalize the framework for duty scheduling of ANs or robots in CPS, which aims at awaking several ANs or robots to work while putting others in sleep mode, such that the system longevity can be extended.

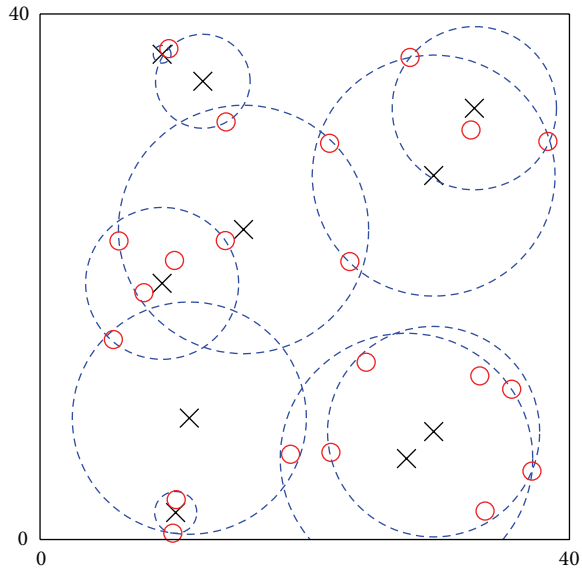


Figure 8: Placement of ANs in a large-scale example with $K = 2$. Crosses and dash circles denote ANs and service coverage of ANs. Circles, squares, and diamonds denote POIs with 25, 40, and 55 consumed utilities, respectively.

Further, to improve serving power, a smart power/battery management system will be potentially integrated into the framework.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is partially supported by the Research Development Fund (RDF-13-01-13) from Xian Jiaotong-Liverpool University, China.

References

[1] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: the next computing revolution," in *Proceeding of the 47th Design Automation Conference (DAC '10)*, pp. 731–736, New York, NY, USA, June 2010.

[2] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber-physical systems approach to data center modeling and control for energy efficiency," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 254–268, 2012.

[3] I. Lee, O. Sokolsky, S. Chen et al., "Challenges and research directions in medical cyber-physical systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 75–90, 2012.

[4] K. Lakshmanan, D. de Niz, R. Rajkumar, and G. Moreno, "Resource allocation in distributed mixed-criticality cyber-physical systems," in *Proceedings of the 30th IEEE International Conference on Distributed Computing Systems (ICDCS '10)*, pp. 169–178, Genova, Italy, June 2010.

[5] J. C. Eidson, E. A. Lee, S. Matic, S. A. Seshia, and J. Zou, "Distributed real-time software for cyber-physical systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 45–59, 2012.

[6] P. Derler, E. A. Lee, and A. Sangiovanni Vincentelli, "Modeling cyber-physical systems," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 13–28, 2012.

[7] M. Younis and K. Akkaya, "Strategies and techniques for node placement in wireless sensor networks: a survey," *Ad Hoc Networks*, vol. 6, no. 4, pp. 621–655, 2008.

[8] K. Akkaya and M. Younis, "C2AP: Coverage-aware and connectivity-constrained actor Positioning in wireless sensor and actor networks," in *Proceedings of the 27th IEEE International Performance Computing and Communications Conference (IPCCC '07)*, pp. 281–288, New Orleans, La, USA, April 2007.

[9] A. A. Abbasi, M. Younis, and K. Akkaya, "Movement-assisted connectivity restoration in wireless sensor and actor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 9, pp. 1366–1379, 2009.

[10] K. Akkaya, I. Guneydas, and A. Bicak, "Autonomous actor positioning in wireless sensor and actor networks using stable-matching," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 25, no. 6, pp. 439–464, 2010.

[11] Y. Li, C. Vu, C. Ai, G. Chen, and Y. Zhao, "Transforming complete coverage algorithms to partial coverage algorithms for wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 4, pp. 695–703, 2011.

[12] T. He, S. Krishnamurthy, J. A. Stankovic et al., "Energy-efficient surveillance system using wireless sensor networks," in *Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services*, pp. 270–283, June 2004.

[13] C.-U. Lei, "LUOPAN: light utility-oriented placement of actuator nodes in sensor/actuator networks," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 14)*, pp. 2563–2566, Melbourne, Australia, June 2014.

[14] C. H. Papadimitriou, "On the complexity of integer programming," *Journal of the Association for Computing Machinery*, vol. 28, no. 4, pp. 765–768, 1981.

[15] L. A. Wolsey, *Integer Programming*, John Wiley & Sons, New York, NY. USA, 1998.

[16] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[17] K. Ramachandran and B. Sikdar, "A population based approach to model the lifetime and energy distribution in battery constrained wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 4, pp. 576–586, 2010.

[18] S. Kellner, M. Pink, D. Meier, and E.-O. Blass, "Towards a realistic energy model for wireless sensor networks," in *Proceedings of the 5th IEEE Annual Conference on Wireless on Demand Network Systems and Services (WONS '08)*, pp. 97–100, Garmisch-Partenkirchen, Germany, January 2008.

[19] Official website of LINGO 14.0, http://www.lindo.com/products/lingo/.

[20] C.-U. Lei, H.-N. Liang, and K. L. Man, "Building a smart laboratory environment via a cyber-physical system," in *Proceedings of the IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE '13)*, pp. 243–246, 2013.

[21] C.-U. Lei, K. L. Man, H.-N. Liang, E. G. Lim, and K. Wan, "Building an intelligent laboratory environment via a cyber-physical system," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 109014, 9 pages, 2013.

[22] C. Lei, H. K.-H. So, E. Y. Lam, K. K. Wong, R. Y. Kwok, and C. K. Y. Chan, "Teaching introductory electrical engineering: project-based learning experience," in *Proceedings of the 1st IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE '12)*, pp. H1B–1–H1B–5, August 2012.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 415415, 11 pages
http://dx.doi.org/10.1155/2014/415415

Research Article

A Comparative Study of Routing Protocols of Heterogeneous Wireless Sensor Networks

Guangjie Han,[1,2] Xu Jiang,[1] Aihua Qian,[1] Joel J. P. C. Rodrigues,[3] and Long Cheng[4]

[1] *Department of Information & Communication Systems, Hohai University, Changzhou 213022, China*
[2] *Changzhou Key Laboratory of Photovoltaic System Integration and Production Equipment Technology, Changzhou 213022, China*
[3] *Institute of Telecommunications, University of Beira Interior, 6201-001 Covilhã, Portugal*
[4] *Department of Computer and Communication Engineering, Northeastern University, Qinhuangdao 066004, China*

Correspondence should be addressed to Guangjie Han; hanguangjie@gmail.com

Received 1 April 2014; Accepted 13 May 2014; Published 22 June 2014

Academic Editor: Jaime Lloret

Recently, heterogeneous wireless sensor network (HWSN) routing protocols have drawn more and more attention. Various HWSN routing protocols have been proposed to improve the performance of HWSNs. Among these protocols, hierarchical HWSN routing protocols can improve the performance of the network significantly. In this paper, we will evaluate three hierarchical HWSN protocols proposed recently—EDFCM, MCR, and EEPCA—together with two previous classical routing protocols—LEACH and SEP. We mainly focus on the round of the first node dies (also called the stable period) and the number of packets sent to sink, which is an important aspect to evaluate the monitoring ability of a protocol. We conduct a lot of experiments and simulations on Matlab to analyze the performance of the five routing protocols.

1. Introduction

1.1. Background and Motivation. Wireless sensor networks (WSNs) have been applied in many fields in recent years, because the sensor nodes can be deployed without any infrastructure and the network can monitor many dangerous or remote places that people cannot reach. At the same time, many of the latest researches related to WSNs mainly focused on routing, coverage, and localization [1]. WSNs are characterized with low-cost microsensor nodes with wireless capability, low power consumption for transmitting data, resource constraints, and battery energy limitation. Since sensor nodes have limited energy, it is urgent to introduce the energy-saving techniques in order to extend the lifetime of WSNs.

To pursue the effective routing protocols for WSNs, many researchers have done lots of studies recently and got the result that a scheme with hierarchy and clustering is promising in improving the scalability and extending the lifetime of WSNs. Low-energy adaptive clustering hierarchy (LEACH) [2] protocol is a classical protocol. Clustering is an efficient method to handle scalability problem and energy consumption challenge. For this reason, it is widely exploited in WSN applications [3].

To further prolong the lifetime of the network and make WSNs more suitable for various scenarios, some researchers proposed WSNs with heterogeneity [4]. Theoretically, we can divide HWSNs into two categories: one is that sensor nodes are deployed with different communication radius [5] and the other is that sensor nodes are deployed with different energy [6]. In fact, heterogeneous routing protocols are very common in WSNs routing protocols. Heterogeneous routing protocols should satisfy the following properties [7].

(i) *Balancing Energy Consumption.* The energies of the nodes in the network are different from each other when the nodes are deployed in the network for the first time. Because of restricted energy resource and large number of deployed sensor nodes, changing the battery for the nodes is a very tough work and sometimes is impossible in some particular scenarios. Then, we deploy some nodes with more energy in the network to act as the center of data aggregation,

processing, and transmission so that the energy dissipation of the whole network can be balanced.

(ii) *The Coordination of Communications.* The communication environment in some sensing areas is unfavorable due to the obstacles, so deploying the nodes with different communication radius is sometimes necessary.

(iii) *Effectiveness for Computation and Storage.* The computational and storage capability of a sensor node is very limited. In some protocols, nodes have to regularly act as the aggregation and relay nodes, and it is necessary for these nodes to have better computational and storage ability than the other nodes to meet this requirement.

1.2. Contributions. In comparison with the homogeneous WSNs, the latest proposed HWSN routing protocols have tried to extend the lifetime of network, prolong the stable period, and achieve the reliable data transmission by deploying the sensor nodes with different capabilities. However, the proposed protocols often cannot balance the energy consumption of sensor nodes efficiently. It is vital to minimize and balance the energy consumption among sensor nodes in a network to improve its performance. In this paper, we will compare the performance of three proposed HWSN protocols recently together with two classical protocols in the same HWSN model. These five protocols are all cluster-based. This paper aims to analyze which protocol can outperform others through comparisons under various scenarios, which will lead us to propose more energy-efficient protocols in HWSNs.

1.3. Paper Organization. The remainder of this paper is organized as follows. In Section 2, related works are briefly introduced. In Section 3, the protocols we compare are presented in detail. In Section 4, we will show you the specific HWSN model in a Matlab platform. In Section 5, simulations of the protocols and the results are given. In Section 6, we summarize the paper and give the future work.

2. Related Work

Recently, many WSN routing protocols have been proposed to improve the performance of the network [8, 9]. They can be divided into two categories: cluster-based protocols [6, 10, 11] and plain-based protocols [5, 12, 13]. As we all know that LEACH is proposed based on the homogeneous WSNs, while, in the practical applications, heterogeneity of nodes cannot be avoided. Proposing the protocol which is suitable for HWSNs is needed. When LEACH is utilized in HWSNs, every sensor node has to select a random number. If the number is less than a threshold $T(n)$, the sensor node becomes a CH for the current round. The threshold is set as follows:

$$T(n) = \begin{cases} \dfrac{p}{1 - p * (r \bmod 1/p)}, & \text{if } s \in G \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where p is the proper percentage of CH nodes in the whole WSN, r is the current round of election, and G is the set of nodes that are not elected to be CH. Many LEACH-type schemes are applied in homogeneous WSNs. In homogenous sensor networks, sensor nodes cannot adapt well to the presence of heterogeneity when the network is in operation. As a result, these nodes which consume more energy will die first, and as a result the LEACH-type protocols turn out to be unstable. On the whole, there are lots of specific WSN applications that could highly benefit from being equipped with a percentage of the nodes which have more initial energy than the normal nodes, because these nodes make sure that there is more stable or dependable feedback from the network. And, in some cases, the stable period is a very important concern.

The proposed cluster-based routing protocols to handle heterogeneity in WSNs mainly focus on three aspects: (1) electing the cluster head by the energy prediction scheme; (2) saving energy consumption by multihop between cluster head and sink; (3) using the evolutionary algorithms.

2.1. The Energy Prediction Scheme. To get better performance, stable election protocol (SEP) [6] is proposed to maintain the hierarchical routing in the HWSNs where two types of nodes have their own election probability. Distributed energy-efficient clustering (DEEC) assumes that a WSN with two types of nodes of different initial energy levels is a two-level heterogeneous network, and the one with three types of nodes of different initial energy levels is a three-level heterogeneous network [14]. In DEEC, the probability for a node to be a CH is based on the ratio between the residual energy of the node and the average energy of the whole network. So the node with more initial energy and residual energy is more likely to be elected as a CH. Other prediction-based cluster schemes include energy dissipation forecast and clustering management (EDFCM) [15], which is an improvement of DEEC, reliable routing based on energy prediction (REP) [16], and energy-efficient prediction clustering algorithm (EEPCA) [17].

2.2. Multihop Transmission. In [18], Younis and Fahmy proposed the hybrid energy-efficient distributed (HEED) clustering algorithm for HWSNs. HEED combines communication range limits and intracluster communication consumption information to improve LEACH protocol. Every sensor node has the initial probability to become a tentative CH depending on its residual energy, and the final CH is selected according to the consumption information. In HEED, the cluster heads are randomly deployed in the sensing area, which makes HEED a cluster-based protocol whose CHs are dynamically selected. HEED has the following advantages: (1) maximizing network lifetime; (2) minimizing control overhead; (3) improving the stability of data transmission; (4) selecting well-distributed cluster heads and well-knit clusters. However, HEED fails to take the balanced energy dissipation among CHs into account. Those CHs near the sink consume energy more quickly than others and they would die first, which causes the energy hole around the

Table 1: Recent proposed routing protocols.

Protocols	Prediction related	Data transmission	Evolutionary related	Energy efficiency
LEACH	No	Single-hop	No	Poor
SEP	No	Single-hop	No	Good
DEEC	Yes	Single-hop	No	Good
EDFCM	Yes	Single-hop	No	Good
REP	Yes	Single-hop	No	Good
EEPCA	Yes	Single-hop	No	Very good
HEED	No	Single-hop	No	Good
EHEED	No	Multihop	No	Good
EEHC	No	Single-hop	No	Very good
MCR	No	Multihop	No	Very good
EAERP	No	Single-hop	Yes	Good
ERP	No	Single-hop	Yes	Good
SAERP	No	Single-hop	Yes	Good

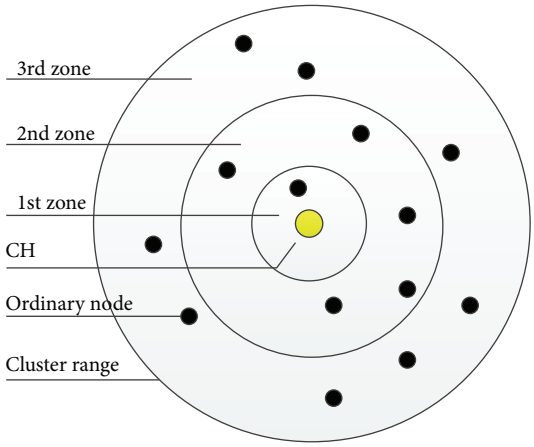


Figure 1: An example of cluster reorganization.

sink [19]. This energy hole is a common phenomenon in this kind of networks where there are many CHs transmitting the information to only one sink node. To maximize the lifetime of the network, Senouci et al. proposed the EHEED (extended HEED) [20]. The procedure of selecting CHs is the same as that of HEED, but the way to save energy of EHEED is based on two lemmas which are proved by [21] so as to build a multihop path to the sink. The main idea of these two lemmas is trying to find a proper relay node A from the ordinary nodes within the cluster which resides between node B and node C during the data transmission phase. To find the proper node A, the communication range of the cluster is evenly divided into three zones to compute the parameters. The example of this kind of cluster is shown in Figure 1.

In [22], Kumar et al. proposed a stable election clustering protocol called energy-efficient heterogeneous clustered (EEHC) scheme in the heterogeneous model. The nodes in the network are divided into three categories according to their initial energy: the normal nodes, the advanced nodes, and the super nodes. Apparently, the normal nodes have the least energy, the advanced nodes have more energy than the normal ones, and the super nodes have the highest level of energy. EEHC is based on SEP, and the three types of nodes in EEHC have their own election probability to be CHs within a fixed time to keep stable. In [11], Kumar et al. improved EEHC further and proposed a multihop clustering protocol called MCR. In MCR, the multihop path is built to reduce the energy consumption.

2.3. The Improvement of Evolutionary Algorithms. Researchers combined the cluster scheme with the biologically inspired routing scheme, and they proposed the evolutionary algorithms (EAs). The EAs are used to handle the cluster-based problem to optimize energy consumption and prolong lifetime of network with heterogeneity, such as energy-aware evolutionary routing protocol (EAERP) [23], evolutionary-based clustered routing protocol (ERP) [24], and stable-aware evolutionary routing protocol (SAERP) [25]. The evolutionary-based routing protocol EAERP redesigned some significant features of EAs, which can assure longer stable period and extend the lifetime with efficient energy dissipation. The protocol ERP overcame the shortcomings of hierarchical clustering-algorithm-based genetic algorithm [26] by uniting the clustering aspects of cohesion and separation error, and then a new fitness function was proposed based on these two clustering aspects. The fitness function is the primary factor used to minimize network energy consumption. SAERP combined the main idea of SEP and EAs, and SAERP mainly aimed at increasing the stability of the network. So these routing schemes which are inspired genetically demonstrated their advantages in prolonging the lifetime of HWSNs.

Table 1 summarizes all the routing protocols above, and some performances of them are compared simply.

3. Typical Protocols for HWSNs

In this section, three latest typical cluster-based HWSN protocols are introduced in detail. They are energy dissipation forecast and clustering management (EDFCM), multihop communication routing (MCR) protocol, and energy-efficient prediction clustering algorithm (EEPCA). These

three protocols are all hierarchical, and they represent three kinds of cluster-based techniques which are used for heterogeneous wireless sensor networks to prolong the lifetime of the network.

3.1. EDFCM. In [15], Zhou et al. proposed a new model for HWSNs with new energy and computation heterogeneity. By using a mathematical method, the authors acquire the energy dissipation model and the priority percentage of cluster heads in HWSNs. In addition, to improve the cluster-based scheme in LEACH-type protocols, a new type of energy-efficient protocol EDFCM which can guarantee the reliable transmission in HWSNs is proposed. CHs selection of this protocol is based on an energy dissipation forecast method and a clustering management method. EDFCM takes the remaining energy and consumption rate of all nodes into account.

3.1.1. The Algorithm of Cluster Head Selection. To predict the energy consumption in the next round, the average energy consumption of CHs of the two types of nodes in previous round is used. If a node has higher forecasted residual energy which is based on the previous prediction value of energy consumption, it will be selected as a CH. In EDFCM, these two types of nodes are set to be type 0 and type 1 with different levels of energy. The weighted probabilities for the two types of nodes to be selected as CHs are defined as

$$P_i(r+1) = \begin{cases} \dfrac{p}{1+\alpha m} \times \dfrac{E_i(r) - E_{Pr\_T0}(r)}{\overline{E}(r+1)}, \\ \qquad \text{if type 0} \\ \dfrac{p}{1+\alpha m} \times (1+\alpha) \dfrac{E_i(r) - E_{Pr\_T1}(r)}{\overline{E}(r+1)}, \\ \qquad \text{if type 1,} \end{cases} \tag{2}$$

where $E_i(r)$ is the residual energy of node i in round r, $E_{Pr\_T0}(r)$ and $E_{Pr\_T1}(r)$ are the average energy dissipations of these two types of cluster heads in the r round, respectively, $\overline{E}(r+1)$ is the average energy of nodes in $r+1$ round, and

$$\overline{E}(r+1) = \frac{1}{N} \times E_{\text{total}} \times \left(1 - \frac{r+1}{R}\right), \tag{3}$$

where E_{total} is the total initial energy of all nodes in the network and R is an estimated round of lifetime of the whole network, which is defined as

$$\begin{aligned} E_{\text{total}} &= N \times E_0 \times (1+\alpha m), \\ R &= \frac{E_{\text{total}}}{E_{\text{round\_total}}}, \end{aligned} \tag{4}$$

where $E_{\text{round\_total}}$ is the total energy consumption in a round.

3.1.2. Operation Mechanism of EDFCM. The operation of EDFCM protocol includes two stages: cluster formation and data collection. At the beginning of cluster formation stage, the information of $2R$ (R refers to the communication radius of a normal node and $2R$ is that of a cluster node) and $\overline{E}(r+1)$ is stored in each node's memory. During the cluster head formation stage, the weighted function of the CH selection probability is calculated firstly. Then, the management nodes make sure that the percentage of CHs is equal to the predefined priority percentage p in each round. At the same time, data collection also contains two substages: the stage of sending data and the stage of sending the information about the current status of energy consumption. EDFCM is a single-hop communication method to transmit the data to sink which means that the CHs communicate with sink directly.

3.2. MCR. In [22], Kumar et al. proposed an energy-efficient heterogeneous clustered (EEHC) scheme for WSNs. In this scheme, EEHC first calculates the optimal cluster numbers based on the side length of the sensing area and the total number of sensor nodes; then, according to the concept of SEP, the clustering algorithm contains two phases: the setup phase and the stable phase. In the setup phase, three different kinds of weighted probability formulas are defined for three kinds of the sensor nodes to elect their own CHs. After the CHs election, the other nodes choose a cluster and join in it. One CH takes the responsibility to transmit the data packets with a single-hop to sink node. The performance of the proposed EEHC system is better than LEACH and SEP in terms of reliability and lifetime. Based on their previous researches, in 2011, Kumar et al. proposed an energy-efficient multihop communication routing (MCR) protocol. MCR provides load balancing, lifetime enhancement, stability, and energy efficiency for the given HWSNs. MCR first calculates the optimal number of the CHs k_{opt} in the network based on the side length of the sensing area, node numbers, and the transmitter amplifier's multiple.

3.2.1. The CH Election Weighted Probabilities. Protocol MCR uses both single-hop transmission and multihop transmission in the network. CHs are picked based on the same weighted probability formulas which are used in EEHC. Cluster member nodes communicate with the CH by using single-hop communication and CH communicates with the sink through multihop communication by choosing the proper CH nearest to the sink as the next hop. In MCR, normal nodes, advanced nodes, and super nodes are deployed randomly together in the sensing area to create the HWSN. The advanced nodes have more initial energy than the normal nodes, and the super nodes have more initial energy than the advanced nodes. The authors consider that m_0 percentage of m nodes are super nodes which initially have β times more initial energy than the normal nodes and the $n * m * (1-m)$ fraction of total nodes are advanced nodes which initially have α times more initial energy than the normal nodes, and the remaining $(1-m)$ percentage of total nodes is normal nodes. n is the number of total sensor nodes. E_0 is defined as the initial energy of the normal node; then, initial energy of each super node and each advanced node should be $E_0(1+\beta)$ and $E_0(1+\alpha)$, respectively.

As described above, the total energy of the whole HWSN setting can be $E_{\text{total}} = nE_0(1 + m(\alpha - m_0(\alpha - \beta)))$. As we can see from E_{total}, the total initial energy is increased

$1 + m(\alpha - m_0(\alpha - \beta))$ times compared with the homogeneous network. To make sure that the election of CHs of the network is stable, which means making these three kinds of nodes elect CHs separately, the new optimal epoch is defined as

$$\frac{1}{p_{\text{opt}}} \times \left(1 + m\left(\alpha - m_0\left(\alpha - \beta\right)\right)\right). \tag{5}$$

Then, the weighted probabilities of three kinds of nodes to become CHs are as follows:

$$\begin{aligned} p_{\text{normal}} &= \frac{p_{\text{opt}}}{1 + m\left(\alpha - m_0\left(\alpha - \beta\right)\right)}, \\ p_{\text{advanced}} &= \frac{p_{\text{opt}}}{1 + m\left(\alpha - m_0\left(\alpha - \beta\right)\right)} \times \left(1 + \alpha\right), \\ p_{\text{super}} &= \frac{p_{\text{opt}}}{1 + m\left(\alpha - m_0\left(\alpha - \beta\right)\right)} \times \left(1 + \beta\right). \end{aligned} \tag{6}$$

By the above formulas, the authors can get threshold to elect the CHs for normal nodes, advanced nodes, and super nodes, respectively.

3.2.2. Cluster Formation, Route Selection, and Data Transmission Phases. In cluster formation phase, non-CH nodes join the nearest CH simply by detecting the RSSI that depends on the received signal from the CHs. After the nodes have completely joined the clusters, a TDMA slot is needed for every cluster, and every CH node sends the TDMA slot to its member nodes to tell them when they can transmit the data. In route selection phase, a CH node aggregates the data from the member nodes and then transmits the data to the sink over a multihop path. Because the shortest path will have the lowest energy cost, a CH node chooses another CH as the next hop whose distance to sink is the shortest. In the data transmission phase, a CH node collects and aggregates the data from its member nodes in the fixed TDMA slot. After this, the CH transmits the data to the sink over the previously built multihop path in the route selection phase.

3.3. EEPCA. It is vital to reduce energy consumption and prolong network lifetime in designing an energy-efficient WSN. In [17], Peng et al. put forward a research on the existing cluster-based schemes for HWSNs and then proposed an energy-predicting clustering algorithm named energy-efficient prediction clustering algorithm (EEPCA). A CH in EEPCA is elected from the sensor nodes by using this algorithm mainly depending on energy consumption and communication cost; thus, the nodes with higher residual energy and lower communication cost are more likely to become a CH than the other nodes. Then, the energy of the network should be consumed uniformly. A prediction model for energy dissipation is also built for this algorithm to be more energy efficient.

3.3.1. Calculation of the Distance between Nodes. The energy consumed by node i transmitting a message to node j is defined as E_i^{tran}; at the same time, node j detects the received data strength with energy $E_{j,i}^{\text{rec}}$. If the distance between node i and node j is $d_{i,j}$, then the relationship between E_i^{tran} and $E_{j,i}^{\text{rec}}$ is shown as follows:

$$E_{j,i}^{\text{rec}} = \frac{K}{d_{i,j}^{\theta}} \times E_i^{\text{tran}}, \tag{7}$$

where K is a constant and θ is the distance-energy gradient that changes from 1 to 6 depending on the application environment.

3.3.2. Cluster Head Selection. Due to the burden of communications and processing various data, CH consumes a great deal of energy compared with the cluster member nodes. Thus, the nodes with more residual energy should have higher probability to become a CH. And it is the same for the other nodes to become a CH in the next round.

The probability p_i of becoming a CH of every node is changing in every round according to its current residual energy [14]. The authors first calculate the optimal number of cluster heads K_{opt}, and then the proportion is

$$p_{\text{opt}} = \frac{K_{\text{opt}}}{N}, \tag{8}$$

where N is the total number of nodes.

The average energy of the nodes within node i's communication range is

$$w(E)_i = \frac{E_i}{\sum_{j-1}^{n}\left(E_j/n\right)}, \tag{9}$$

where n is the number of nodes within node i's communication range.

To predict the energy dissipation more precisely, the author divides the communication range of a node into two sublevels. Level one contains those nodes whose distance to the center node is smaller than d_0, while level two contains those nodes whose distance to the center node is larger than d_0, and d_0 is a predefined constant.

If the number of nodes in level one is m_1 and the number of nodes in level two is m_2, then the average energy consumption of every round within every node's communication range is $\overline{E}_{i-\text{round}}$ and the predicted energy consumption of every node in every round is $\overline{E}_{\text{consume}}$, respectively.

Then, the communication cost factor is as follows:

$$w(C)_i = \frac{\overline{E}_{\text{consume}}}{\overline{E}_{i-\text{round}}}. \tag{10}$$

After integrating $w(E)_i$ and $w(C)_i$, the probability of node i to be elected as a cluster head is

$$p_i = p_{\text{opt}}\left(aw(E)_i + bw(C)_i\right), \tag{11}$$

where $a + b = 1$. Here, a and b will be set to be 0.5 while changing the other parameters in our later simulations to see the performance of EEPCA.

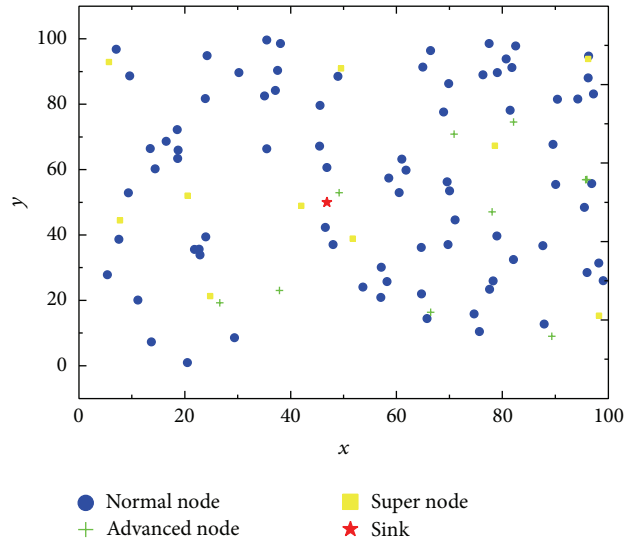


Figure 2: The network model with heterogeneity in three energy levels.

Table 2: Basic parameters used in simulations.

Parameter	Value
Sensing area	100 m ∗ 100 m
Sink location	(50 m, 50 m)
Number of nodes N	100
Priority percentage p	0.1
Initial energy of normal node	0.5 J
α	1.0
β	1.2
m	0.2
m_0	0.5
R	25 m
Data packet size	4000 bits
E_{elec}	50 nJ/bit
E_{fs}	10 pJ/(bit∗m^2)
E_{mp}	0.0013 pJ/(bit∗m^4)
r_{max}	5000

At last, a new threshold formula T_i for node i is similar to LEACH protocol, as shown in the following:

$$T(i) = \begin{cases} \dfrac{p_i}{1 - p_i \left(r \bmod (1/p_i)\right)} \\ \times \left[\begin{array}{c} (aw(E)_i + bw(C)_i) \\ + r_s \operatorname{div} \left(\dfrac{1}{p_i}\right) \\ \times \left(1 - aw(E)_i + bw(C)_i\right) \end{array} \right], & \text{if } i \in G \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

where r_s is the number of rounds that a node fails to be selected as the cluster head.

4. Network Model

4.1. Node Deployment. In this paper, different kinds of nodes with different energy but the same sensing radius and communication radius are deployed in the heterogeneous network. The basic model of the network is shown in Figure 2.

As we can observe from Figure 2, there are three types of nodes deployed in the network: normal nodes, advanced nodes, and super nodes, and they are shown in different colors and shapes. The difference between these three types of nodes is their initial energy. Sink is located at the center of the network, and the other sensor nodes are deployed randomly in the network area.

4.2. Energy Dissipation. In our study, we use the similar energy dissipation model which is proposed in [2]. The radio energy dissipation model is illustrated in Figure 3. When a node transmits L bit message over a distance d to another node, the energy consumed by the radio is defined as

$$E_{Tx}(L, d) = \begin{cases} L \times E_{elec} + L \times E_{fs} \times d^2, & \text{if } d < d_0 \\ L \times E_{elec} + L \times E_{fs} \times d^4, & \text{if } d \geq d_0, \end{cases} \quad (13)$$

where E_{fs} and E_{mp} depend on the transmitter amplifier model, d_0 is equal to $\sqrt{E_{fs}/E_{mp}}$, and the energy dissipation is defined as

$$E_{Rx}(L) = E_{Rx-elec}(L) = L \times E_{elec}. \quad (14)$$

4.3. Simulation Setup. As shown in Table 2, sensor nodes are distributed in an area of 100 m∗100 m, and sink is located at the center of sensing area, and the number of nodes N is 100. The advanced node has α times more energy than the normal node, and the super node has β times more energy than the normal node. Priority percentage p is calculated theoretically according to the previous work. The fraction m is the fraction of the number of heterogeneous nodes of all nodes, and m_0 is the fraction of super nodes of all the heterogeneous nodes. R is the sensing radius of single node and r_{max} is the total round of network or the running time of network. The parameters are the basis. We can change some of them to create the different simulation environments in our later experiments. There are three kinds of nodes with three energy levels. In simulations, we consider advanced node and super node in EDFCM to be type 0 and the normal node to be type 1. We also consider advanced node and super node to the same type of node in SEP. The details will be given in the following part. To evaluate the performance of the algorithms we introduced in this paper, we conduct extensive simulation experiments on Matlab.

5. Simulation and Performance Analysis

Simulations are run to compare the performance of the protocols in five scenarios in terms of the round of the first node dies and packets that sink receives. The former one refers to the stable period of the network which is very important in some occasions and the latter one refers to the monitoring ability which is also a critical factor in some

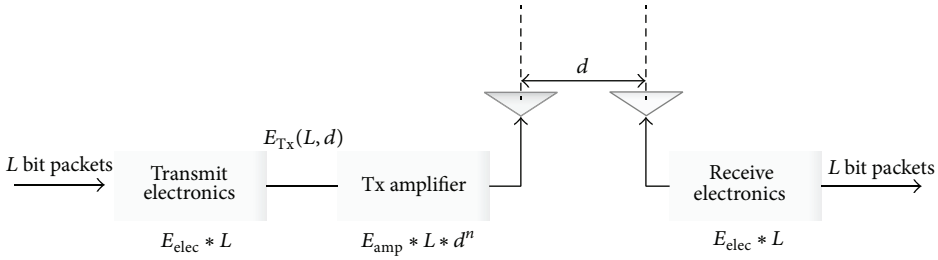


Figure 3: Radio energy dissipation model.

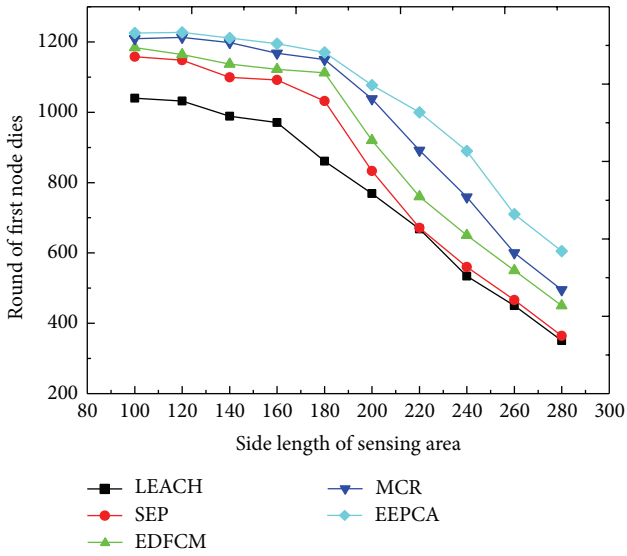


Figure 4: Round of first node dies with varying side length of sensing area.

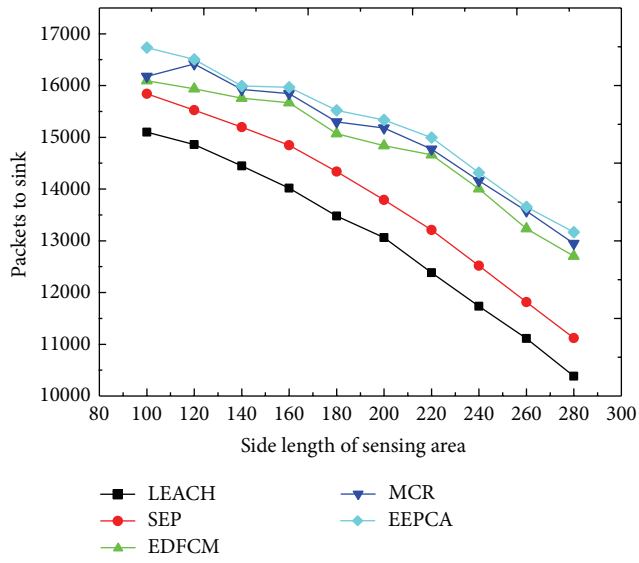


Figure 5: Packets to sink with varying side length of sensing area.

WSN applications. Furthermore, we put forward another two scenarios to compare the lifetime of network.

As shown in Figures 4 and 5, the packets that the sink receives and the round of the first node dies are decreasing with the increasing of the side length of the sensing area which is changing from 100 m to 280 m at the step of 20 m. Because of the increasing of side length, the density of nodes in the area is decreasing, which results in the distance between two nodes getting farther. One node has to consume more energy to transmit data to the neighbors. As a result, the energy dissipation of a single node and the network become higher. And the time when the first node dies becomes earlier correspondingly, which leads to the decrease of lifetime of the network as well as the number of packets received by the sink. We can also observe from the two figures that EEPCA, MCR, and EDFCM have better performance than the two former protocols, SEP and LEACH. EEPCA can make the energy of nodes uniformly consumed in the network, so it has higher energy efficiency than MCR and EDFCM. MCR utilizes a multihop way to transmit the data from CH to sink at the data transmission phase, and we know from other articles that most of the energy is used to transmit data from CH to sink. In MCR, three types of nodes have their own election probability to be stably selected as CHs, but MCR cannot uniformly consume the energy like EEPCA. EDFCM limits the number of CHs during the whole process; when the number of CHs is beyond the threshold, EDFCM randomly chooses some of CHs and turns them into a non-CH and when the number of CHs is below the threshold EDFCM also chooses some nodes with more energy to be CH. An energy prediction method is also introduced to predict every node's probability to decide which is most likely to be CH in the next round. However, EDFCM transmits the data from CH to sink by single-hop; thus, CH consumes more energy than MCR.

In Figures 6 and 7, we change the number of nodes from 100 to 190 at the step of 10 to see how the five protocols work, and the other parameters remain the same as shown in Table 2. In Figure 6, we can observe that, with the increasing number of nodes, the time of the first node dies almost stays in the same levels among these five protocols. This is because even though the number of nodes increases, the average energy consumption of communication and data transmission in the cluster is almost the same and the average energy consumption in one node changes a little. We can also discover that EEPCA, MCR, and EDFCM have better performance in stable period than LEACH and SEP, because they can make the nodes dissipate their energy uniformly

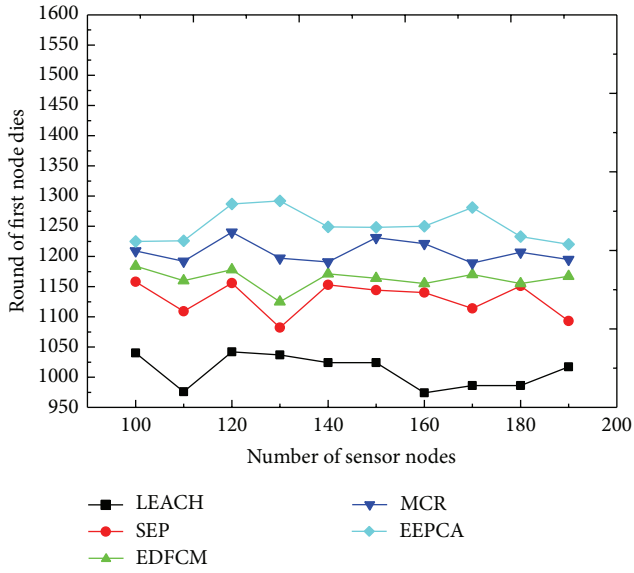


Figure 6: Round of first node dies with varying number of sensor nodes.

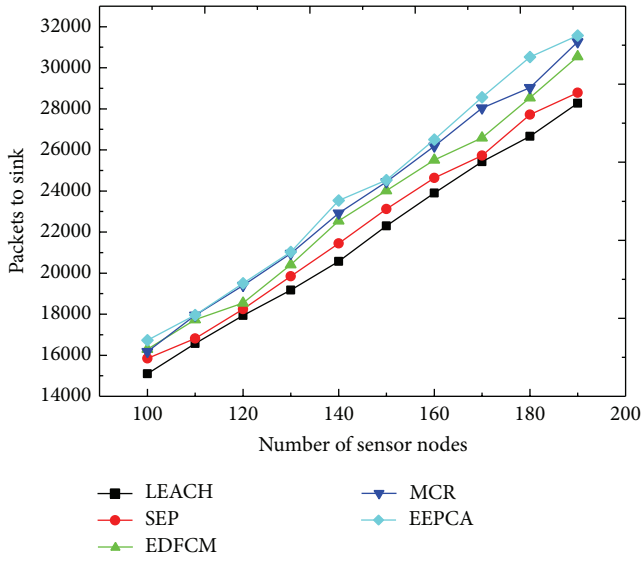


Figure 7: Packets to sink with varying number of sensor nodes.

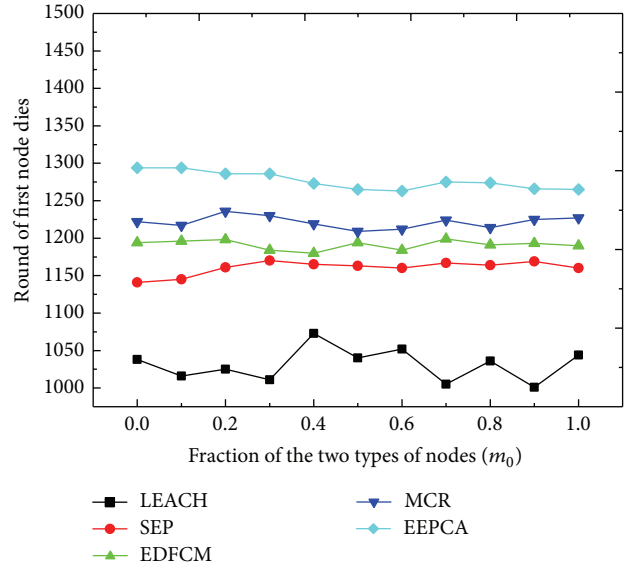


Figure 8: Round of first node dies with varying fraction of the two types of nodes.

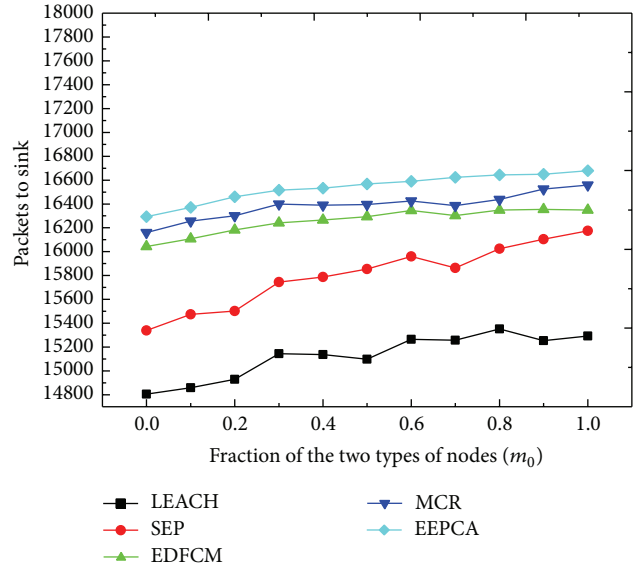


Figure 9: Packets to sink with varying fraction of the two types of nodes.

and balance the energy consumption to provide longer stable period.

In Figure 7, the reason why the packets sent to sink are increasing is that the increasing number of sensing nodes creates more sensing data, and, at the same time, they elect more CHs to transmit these data packets to sink. We can also observe from Figure 7 that EEPCA has better network monitoring quality than the other algorithms.

In Figures 8 and 9, we change the fraction m_0 from 0 to 1 at the step of 0.1. m_0 is the fraction of the super nodes in the total heterogeneous nodes, and the other parameters stay the same as shown in Table 2. We can observe from Figure 8 that the round of first node dies of these five protocols almost stays at the same level under different m_0. This is because the first node dies usually occur to the normal node, and the fraction m does not change, which means that the number of normal nodes does not change. The energy consumed in communication and data transmission among the normal nodes remains almost the same, and the stable period almost stays at the same level. In Figure 9, with the increase of m_0, the number of advanced nodes decreases, but the number of super nodes increases; at the same time, the total energy of heterogeneous nodes increases, which results in the increase of the energy of the entire network. With more energy, more nodes can survive for a longer time, which makes them transmit more packets to sink. That is why the number of

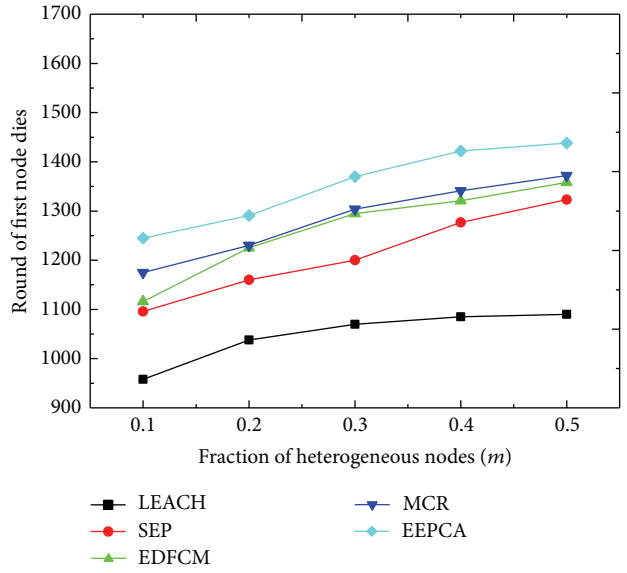


Figure 10: Round of first node dies with varying fraction of heterogeneous nodes.

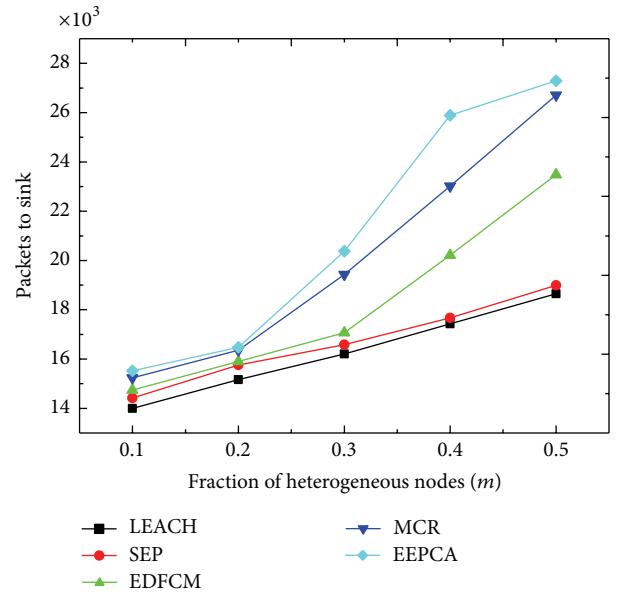


Figure 11: Packets to sink with varying fraction of heterogeneous nodes.

packets sent to sink increases with the increase of m_0. And, for the same reason we discussed above, EEPCA has the best network monitoring quality in these protocols, and the result shows that MCR is better than EDFCM.

In Figures 10 and 11, we change the parameter m from 0.1 to 0.5 at the step of 0.1. m is the fraction of heterogeneous nodes and the other parameters stay the same as shown in Table 2. In Figure 10, we can observe that the round of first node dies increases slightly when m increases. This is because the increase of m means more heterogeneous and less normal nodes. These five protocols are all cluster-based, and the main idea of them is to elect the node which can best manage the cluster as a CH. So the node with more energy has the priority to be a cluster head. As we discussed above, the first node dies usually occur to the normal node, while the normal node mainly acts as the cluster member rather than a CH, which has a smaller energy consumption rate than the advanced node and super node. In all the five algorithms, the first node tends to die later with the increase of m. In Figure 11, it is apparent that the packets that the sink receives are increasing. Because the rising of m causes the total energy of network to rise, then the nodes have more time to collect and transmit data packets.

In Figures 12 and 13, we compare the performance under different values of p which is the priority percentage of CHs of all sensor nodes. As we can observe from the figure, the p changes from 0.1 to 0.5 at the step of 0.1. Round of first node dies of these five protocols does not change much under different values of p, while the packets that the sink receives increase a lot with the increasing of p. In Figure 12, the reason why the stable period of every protocol stays at almost the same level is that, at the beginning, cluster heads are mainly elected from the advanced nodes and super nodes because they have more energy to be capable of managing the clusters and they spend the most of the energy during

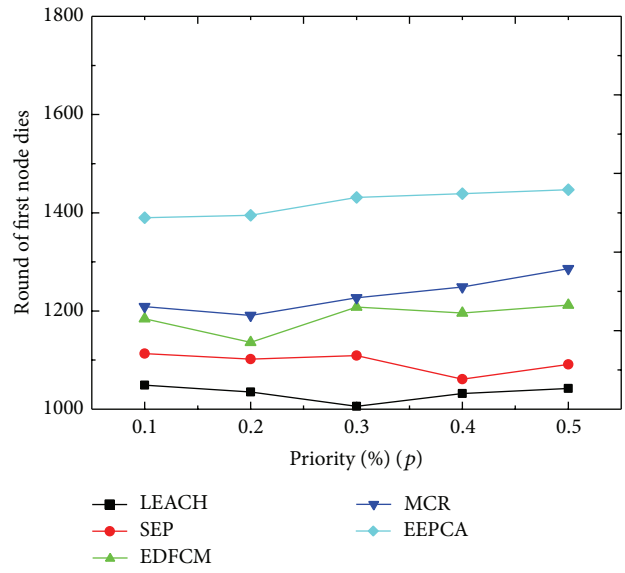


Figure 12: Round of first node dies with varying priority percentage.

the data gathering and transmitting phase. In contrast, most of the normal nodes have very little chance of being elected as a CH, and they need not consume that much energy correspondingly. Even though the increase of p leads to the increase of the probability of the normal nodes to be a cluster head, those heterogeneous nodes are still the main part of cluster heads. In Figure 13, we can observe that the amount of packets sent to sink is rising with the increase of priority percentage p, because the increase of p enables the nodes to elect more CHs to send the sensed data packets to sink.

In Figure 14, we set $m = 0$, $m_0 = 0$, $\alpha = 0$, and $\beta = 0$; m is the fraction of total number of heterogeneous

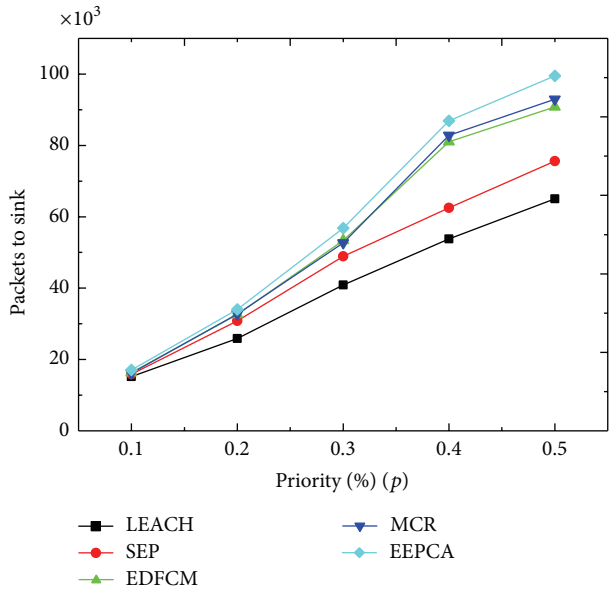


Figure 13: Packets to sink with varying priority percentage.

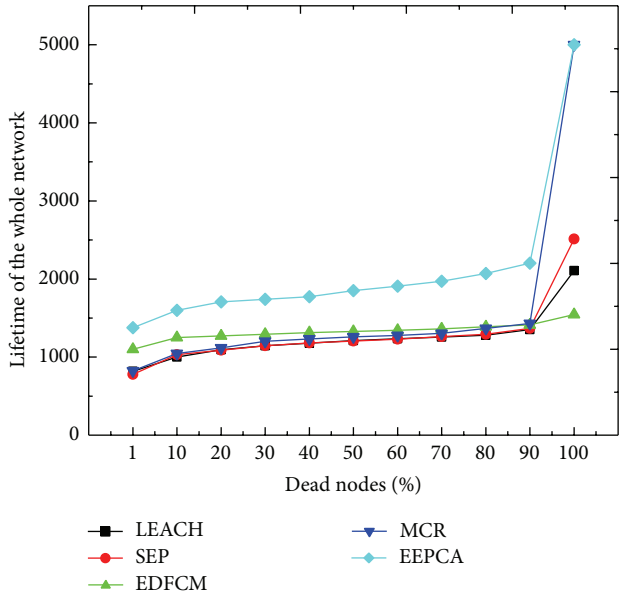


Figure 14: Lifetime of the whole network.

nodes of all nodes, m_0 is the fraction of super nodes in the fraction m, α is the energy multiple which means that advanced node has α times more energy than the normal node, and β is the energy multiple which means that super node has β times more energy than the normal node. Figure 14 describes a homogeneous circumstance, and we can observe that heterogeneous cluster-based protocols have a longer lifetime than LEACH, and the former can also be applied in the homogeneous circumstance. Because EEPCA has a good ability of balancing energy consumption, it can achieve a longer stable period and lifetime no matter whether the network is homogeneous or heterogeneous. MCR uses

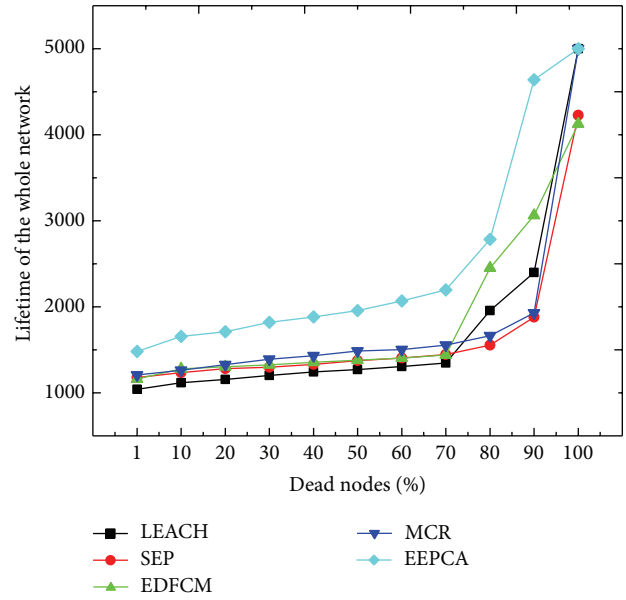


Figure 15: Lifetime of the whole network.

both multihop and stable election to save energy. In fully distributed manner, EDFCM elects the CHs by using one-step energy consumption prediction, but a CH consumes much more energy when transmitting the packets to sink by single-hop, so its performance is not so much better than the former two protocols but much better than LECAH and SEP.

In Figure 15, $m = 0.2$, $m_0 = 0.5$, $\alpha = 1$, and $\beta = 1.2$, and there is no doubt that heterogeneous cluster-based protocols have the better performance, because these heterogeneous protocols have the ability to manage the clusters and their member nodes and can better balance the energy consumption of the nodes in the whole network.

6. Conclusions and Future Work

Simulation results show that the characteristics of HWSN algorithms are better than the homogeneous ones in terms of both the round of the first node dies and the number of packets sent to sink. As mentioned above, these heterogeneous cluster-based protocols have the ability to manage the clusters and their member nodes and can better balance the energy consumption of the nodes in the whole network. Moreover, the multihop path among CHs to sink is a very important concern to save energy during the data transmission. Our further work will mainly focus on how to further balance the energy consumption of every node by using the unequal clusters and on the moving heterogeneous sensor nodes. Furthermore, the energy whole problem is to be relieved in the network.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work has been partially supported by the Natural Science Foundation of Jiangsu Province of China (no. BK20131137). Joel J. P. C. Rodrigues's work has been supported by Instituto de Telecomunicações, Next Generation Networks and Applications Group (NetGNA), Covilhã Delegation, and by National Funding from the FCT, Fundação para a Ciência e a Tecnologia, through the Pest-OE/EEI/LA0008/2013 Project.

References

[1] G. Han, H. Xu, T. Q. Duong, J. Jiang, and T. Hara, "Localization algorithms of wireless sensor networks: a survey," *Telecommunication Systems*, vol. 52, no. 4, pp. 2419–2436, 2013.

[2] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, 2002.

[3] C. Sevgi and A. Kocyigit, "On determining cluster size of randomly deployed heterogeneous WSNs," *IEEE Communications Letters*, vol. 12, no. 4, pp. 232–234, 2008.

[4] E. J. Duarte-Melo and M. Liu, "Analysis of energy consumption and lifetime of heterogeneous wireless sensor networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM'02)*, pp. 21–25, IEEE Press, Taipei, Taiwan, November 2002.

[5] S.-S. Wang and Z.-P. Chen, "LCM: a link-aware clustering mechanism for energy-efficient routing in wireless sensor networks," *IEEE Sensors Journal*, vol. 13, no. 2, pp. 728–736, 2013.

[6] G. Smaragdakis, I. Matta, and A. Bestavros, "SEP: a stable election protocol for clustered heterogeneous wireless sensor networks," in *Proceedings of the 2nd International Workshop on Sensor and Actor Network Protocols and Applications (SANPA '04)*, pp. 251–261, Boston university Computer Science Department, 2004.

[7] V. Katiyar, N. Chand, and S. Soni, "A survey on clustering algorithms for heterogeneous wireless sensor networks," *International Journal of Advanced Networking and Applications*, vol. 2, no. 4, pp. 273–287, 2011.

[8] K. Lin, J. J. P. C. Rodrigues, H. Ge, N. Xiong, and X. Liang, "Energy efficiency QoS assurance routing in wireless multimedia sensor networks," *IEEE Systems Journal*, vol. 5, no. 4, pp. 495–505, 2011.

[9] S. Tyagi and N. Kumar, "A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 623–645, 2013.

[10] J. Yu, Y. Qi, G. Wang, and X. Gu, "A cluster-based routing protocol for wireless sensor networks with nonuniform node distribution," *AEU-International Journal of Electronics and Communications*, vol. 66, no. 1, pp. 54–61, 2012.

[11] D. Kumar, T. C. Aseri, and R. B. Patel, "Multi-hop communication routing (MCR) protocol for heterogeneous wireless sensor networks," *International Journal of Information Technology, Communications and Convergence*, vol. 1, no. 2, pp. 130–145, 2011.

[12] X. Chen, Z. Dai, W. Li et al., "ProHet: a probabilistic routing protocol with assured delivery rate in wireless heterogeneous sensor networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1524–1531, 2013.

[13] W. Jing and Y. Liu, "Routing protocol based on link reliability for WSN," *Physics Procedia*, vol. 33, pp. 410–416, 2012.

[14] L. Qing, Q. Zhu, and M. Wang, "Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks," *Computer Communications*, vol. 29, no. 12, pp. 2230–2237, 2006.

[15] H. Zhou, Y. Wu, Y. Hu, and G. Xie, "A novel stable selection and reliable transmission protocol for clustered heterogeneous wireless sensor networks," *Computer Communications*, vol. 33, no. 15, pp. 1843–1849, 2010.

[16] K. Lin and M. Chen, "Reliable routing based on energy prediction for wireless multimedia sensor networks," in *Proceedings of the 53rd IEEE Global Communications Conference (GLOBECOM '10)*, pp. 1–5, December 2010.

[17] J. Peng, T. Liu, H. Li, and B. Guo, "Energy-efficient prediction clustering algorithm for multilevel heterogeneous wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 678214, 8 pages, 2013.

[18] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Transactions on Mobile Computing*, vol. 3, no. 4, pp. 366–379, 2004.

[19] T. Liu, Q. Li, and P. Liang, "An energy-balancing clustering approach for gradient-based routing in wireless sensor networks," *Computer Communications*, vol. 35, no. 17, pp. 2150–2161, 2012.

[20] M. R. Senouci, A. Mellouk, H. Senouci, and A. Aissani, "Performance evaluation of network lifetime spatial-temporal distribution for WSN routing protocols," *Journal of Network and Computer Applications*, vol. 35, no. 4, pp. 1317–1328, 2012.

[21] I. Stojmenovic and X. Lin, "Power-aware localized routing in wireless networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 12, no. 11, pp. 1122–1133, 2001.

[22] D. Kumar, T. C. Aseri, and R. B. Patel, "EEHC: energy efficient heterogeneous clustered scheme for wireless sensor networks," *Computer Communications*, vol. 32, no. 4, pp. 662–667, 2009.

[23] E. A. Khalil and B. A. Attea, "Energy-aware evolutionary routing protocol for dynamic clustering of wireless sensor networks," *Swarm and Evolutionary Computation*, vol. 1, no. 4, pp. 195–203, 2011.

[24] B. A. Attea and E. A. Khalil, "A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks," *Applied Soft Computing Journal*, vol. 12, no. 7, pp. 1950–1957, 2012.

[25] E. A. Khalil and B. A. Attea, "Stable-aware evolutionary routing protocol for wireless sensor networks," *Wireless Personal Communications*, vol. 69, no. 4, pp. 1799–1817, 2013.

[26] R. V. Kulkarni, A. Förster, and G. K. Venayagamoorthy, "Computational intelligence in wireless sensor networks: a survey," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 1, pp. 68–96, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 926020, 11 pages
http://dx.doi.org/10.1155/2014/926020

Research Article

A Fast Density-Based Clustering Algorithm for Real-Time Internet of Things Stream

Amineh Amini, Hadi Saboohi, Teh Ying Wah, and Tutut Herawan

Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Correspondence should be addressed to Amineh Amini; amini@siswa.um.edu.my

Received 10 April 2014; Accepted 18 May 2014; Published 19 June 2014

Academic Editor: Xudong Zhu

Data streams are continuously generated over time from Internet of Things (IoT) devices. The faster all of this data is analyzed, its hidden trends and patterns discovered, and new strategies created, the faster action can be taken, creating greater value for organizations. Density-based method is a prominent class in clustering data streams. It has the ability to detect arbitrary shape clusters, to handle outlier, and it does not need the number of clusters in advance. Therefore, density-based clustering algorithm is a proper choice for clustering IoT streams. Recently, several density-based algorithms have been proposed for clustering data streams. However, density-based clustering in limited time is still a challenging issue. In this paper, we propose a density-based clustering algorithm for IoT streams. The method has fast processing time to be applicable in real-time application of IoT devices. Experimental results show that the proposed approach obtains high quality results with low computation time on real and synthetic datasets.

1. Introduction

Using RFID and conventional sensors in the base of the data collection mechanisms in Internet of Things (IoT) makes the volume of the collected data intensively large. In many cases, the communications and data transfers between the objects are required to enable smart analytics. Such communications and transfers require both bandwidth and energy consumption, which are usually limited resources in real scenarios. Furthermore, the analytics required for such applications is often real-time, and therefore it requires the design of methods which can provide real-time insights [1–3]. Data mining techniques are very useful for this kind of analytics. However, since the generated data is considered as stream, we modify the multilayer data mining model for Internet of Things (IoT) from [4] to a multilayer data stream mining model for IoT. The model is illustrated in Figure 1.

Mining data stream is relatively a new area of research in the data mining community. It became more prominent in many applications such as monitoring environmental sensors, social network analysis, real-time detection of anomalies in computer network traffic, and web searches [5, 6].

Clustering is a remarkable task in mining data stream [6]. However, data stream clustering needs some important requirements due to data streams' characteristics such as clustering in limited memory and time with single pass over the evolving data streams and also handling noisy data [7–9].

There are different methods for clustering data streams. In clustering methods, data are categorized based on the similarities among objects. The similarity is determined based on distance or density [5]. The distance-based method [10] leads to form only spherical shapes. On the other hand, density-based method [11] has the ability to detect any shape cluster and they are useful for identifying the noise.

In the last few years, many proposals to extend density-based clustering for data stream have been presented [12]. Density-based data stream clusterings are mainly grouped as density grid-based method and density microclustering method.

The density grid-based clustering [13] quantizes the data space into a number of density grids that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time, which is independent of the number of data points, yet

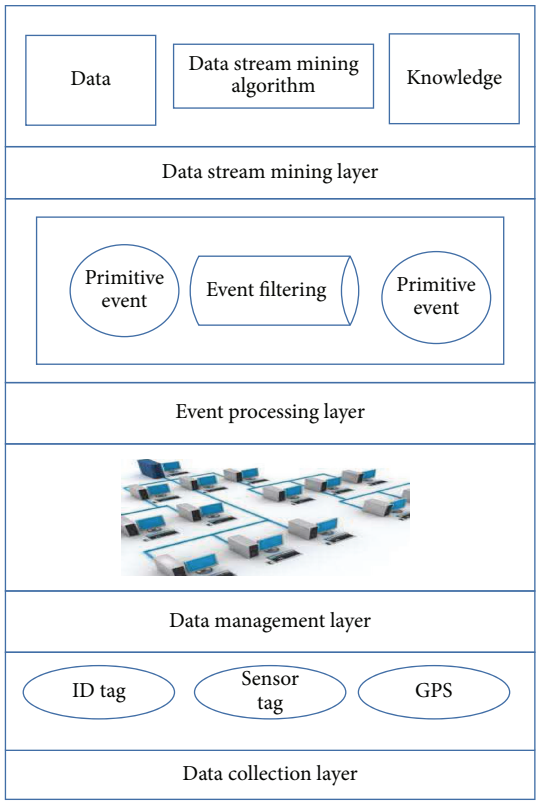


Figure 1: Multilayer data stream mining model for Internet of Things (adopted from [4]).

dependent on only the number of cells. However, they may have lower quality and accuracy of the clusters despite the fast processing time of the technique [5]. Some of density grid-based clustering algorithms are D-Stream [14], MR-Stream [9], and ExCC [15].

On the other hand, in density-based microclustering [16], microclusters keep summary information about data and clustering is performed on this synopsis information. Microcluster [10] is a temporal extension of cluster feature (CF), that is, a summarization triple maintained about a cluster. Density-based microclustering methods keep summary of clusters in microclusters and form final clusters from them. They have better quality compared to grid-based ones but need more computation time. Some of the density microclustering algorithms include DenStream [14], FlockStream [17], and SOStream [18].

To mitigate the problem of density microclustering methods, we propose a hybrid density-based method for clustering evolving data streams. Our proposed method uses the advantages of both density grid-based and microclustering methods. We refer to our algorithm as HDC-Stream (hybrid density-based clustering for data stream). HDC-Stream has three steps: in step one, the new data point is either mapped to the gird or merged to an existing minicluster. Minicluster is a concept similar to microcluster which is formed from a grid cell. Second step prunes miniclusters and grids in each pruning time. Last step forms the final clusters from the pruned miniclusters using a modified DBSCAN algorithm.

The main contributions of HDC-Stream are summarized as follows.

(1) In HDC-Stream, instead of searching list of outlier microclusters to find the suitable one, it maps the new data point into the grid cell which saves computation time. This reduces the number of comparisons from $o(\text{mi})$ in finding outlier microclusters to $o(1)$ which is the mapping time. mi is the number of miniclusters.

(2) In HDC-Stream, instead of forming a new microcluster for a new data point, which is not placed in any existing microcluster and may be a seed of outlier, the new data point is mapped and kept in the grid until the grid density reaches a predefined threshold. In this case, it is converted to a minicluster.

(3) The experimental results also show that it outperforms two of the well-known existing density microclustering and density grid-based clustering methods in terms of quality and execution time. Furthermore, the experimental results show that HDC-Stream obtains clusters of high quality even when the noise is present.

The remainder of this paper is organized as follows: Section 2 surveys related work. Section 3 introduces basic definitions. In Section 4, we explain in detail the HDC-Stream algorithm. We analyze the HDC-Stream algorithm using synthetic and real datasets in Section 5. Section 6 discusses the advantages of the proposed method. We conclude the paper in Section 7.

2. Related Work

Clustering is an important task in data stream mining. Recently, a plenty of clustering algorithms have been developed for data streams. These clustering algorithms can be generally grouped into the four following main categories [5].

A partitioning-based clustering algorithm tries to find the best partitioning for data points in which intraclass similarity is maximum and interclass similarity is minimum. Two of the well-known extensions of k-means [19, 20] on data streams are STREAM [7] and CluStream [10]. Hierarchical clustering algorithms work by decomposing data objects into a tree of clusters. BIRCH [10] and ClusTree [8] are examples of hierarchical clustering family. Grid-based clustering is independent of the distribution of data objects. In fact, it partitions the data space into a number of cells, which forms the grids. Grid-based clustering has fast processing time since it is not dependent on the number of data objects. D-Stream [14], MR-Stream [9], and ExCC [15] are grid-based clusterings over data stream.

Density-based clustering algorithms have been developed to discover clusters with arbitrary shapes. They find clusters based on the dense areas in a shape. If two points are close enough and the region around them is dense, then these two data points join and contribute to construction of

a cluster. DBSCAN [21], OPTICS [22], and DENCLUE [23] are examples of this approach.

Due to data streams' characteristics, the traditional density-based clustering is not applicable. Recently, many density-based clustering algorithms are extended for data streams. The main idea in these algorithms is using density-based method in the clustering process and at the same time overcoming the constraints, which are put by data stream's nature. Density-based clustering algorithms are categorized into two broad groups called density microclustering and density grid-based clustering algorithms. A comprehensive survey on density-based clustering algorithm on data stream is presented in [12].

DenStream [24] is a density microclustering algorithm for evolving data stream. The algorithm extends the micro-cluster [10] concept and introduces the outlier and potential microclusters to distinguish between outliers and the real data. It has online and offline phases. In the online phase, the microclusters are formed and the offline phase performs macroclustering on the microclusters. FlockStream [17] is an extension of DenStream using a bioinspired model. It is based on flocking model [25] in which agents are microclusters and they work independently but form clusters together. It considers an agent for each data point which is mapped in the virtual space. Agents move in their predefined visibility range for a fixed time. If they visit another agent, they join to form a cluster in case they are similar to each other. It merges the online and offline phases since the agents form the clusters at any time. In FlockStream, searching for the similar agents is a time consuming process. SOStream (self-organizing density-based clustering over data stream) [18] detects structures within fast evolving data streams by automatically adapting the threshold for density-based clustering. SOStream dynamically creates, merges, and removes clusters in an online manner. It uses competitive learning as introduced for SOMs (self-organizing maps) [26] which is a time consuming method for clustering data stream. Density microclusterings are effective in terms of quality and they can capture the evolution of clusters effectively. However, they have high computation time in finding suitable microclusters.

The other important category is density grid-based method. D-Stream [27] is a density grid-based clustering algorithm in which the data points are mapped to the corresponding grids and the grids are clustered based on their density. It adjusts the clusters in real-time and captures the evolving behavior of data streams and has techniques for handling the outliers. MR-Stream [9] is another clustering algorithm which has the ability to cluster data stream at multiple resolutions. The algorithm partitions the data space into cells and a tree-like data structure which keeps the space partitioning. The tree data structure keeps the data clustering in different resolutions. Each node has the summary information about its parent and children. The algorithm improves the performance of clustering by determining the right time to generate the clusters. D-Stream and MR-Stream algorithms cannot work properly for high dimensional data stream [12]. ExCC (exclusive and complete clustering) [15] is a density grid-based clustering for heterogeneous data stream. The algorithm maps the numerical attributes to the grid and the categorical attributes are assigned granularities according to distinct values in respective domain sets. ExCC introduces fast and slow stream based on the average arrival time of the data points in the data stream. The algorithm detects noise in the offline phase using wait and watch policy. For detecting real outliers, it keeps the data points in the hold queue, which is kept separately for each dimension. The hold queue strategy needs more memory and processing time since it is defined for each dimension. Density grid-based clustering has lower quality since it depends on the granularity of clustering. On the other hand, they can handle the outlier effectively. The computation time is high for high dimensional data.

3. Basic Definitions of HDC-Stream

Definition 1 (ϵ-neighborhood of a point). The neighborhood is within a radius of ϵ. Neighborhood of point p is denoted by $N_\epsilon(p)$:

$$N_\epsilon(p) = \{q \in D \operatorname{dist}(p,q) \leq \epsilon\}, \tag{1}$$

where $\operatorname{dist}(p,q)$ is an Euclidean distance between p and q.

Definition 2 (MinPts). MinPts is the minimum number of data points around a data point p in the ϵ-neighborhood of p.

Definition 3 (data point weight value). For each data point in the data stream, we consider a weight which decreases over time. The initial value of data point is 1. The weight of data point x (with d dimensions) in time t_c is defined based on the weight in t_p as follows ($t_c > t_p$):

$$w(x,t_c) = w(x,t_p) f(t_c - t_p), \tag{2}$$

where function f is a fading function. The fading function [28] that we use in HDC-Stream is defined as $f(t) = 2^{-\lambda t}$, where $\lambda > 0$.

Definition 4 (grid weight). For a grid g at current time t_c, the grid weight is defined based on sum of data points' weights which are mapped to it:

$$w(g,t_c) = \sum_{x \in g} 2^{-\lambda(t_c - t_x)}. \tag{3}$$

According to the work presented in [27], we update the grid weight in t_c with the last updated value t_p as follows:

$$w_g(t_c,x) = 2^{-\lambda(t_c - t_p)} * w_g(t_p) + 1. \tag{4}$$

The total weight of all the grids in data space S is $w(S,t) = \sum_{x \in S(t)} w(x,t)$ which is less than $1/(1 - 2^{-\lambda})$. Moreover, we have

$$\lim_{t \to \infty} \sum_{x \in S(t)} w(x,t) = \frac{1}{1 - 2^{-\lambda}}. \tag{5}$$

It means that sum of all data points' weights has an upper bound of $1/(1 - 2^{-\lambda})$. The number of grids equals N, which

is $N = \prod_{i=1}^{d} P_i$, and every ith dimension is divided into P_i partitions. Therefore, the average density of each grid is $1/N(1-2^{-\lambda})$.

Definition 5 (core point). It is defined as an object for which its overall weight of all ϵ-neighborhood data points is at least a value $1/N(1-\lambda)$.

Definition 6 (dense grid). At time t, for a grid g, we call it a dense grid if $w_g(t) \geq \alpha/N(1-2^{-\lambda})$.

Definition 7 (sparse grid). At time t, for a grid g, we call it a sparse grid if $w_g(t) < \alpha/N(1-2^{-\lambda})$.

Because the overall weight cannot be more than $1/(1-\lambda)$, α is a controlling threshold.

Definition 8 (minicluster (MIC)). A MIC at time t is defined as MIC(w, c, r) for a group of very close data points $p_{i1}, \ldots, p_{in}$ with timestamps $T_{i1}, \ldots, T_{in}$ as follows:

$$w_{\mathrm{MIC}} = w_g(t), \quad w_g(t) \geq \frac{\alpha}{N\left(1-2^{-\lambda}\right)},$$
$$\mathrm{center}_{\mathrm{MIC}} = \frac{\sum_{j=1}^{n} 2^{-\lambda(t-T_{ij})}\left(p_{ij}\right)}{w_{\mathrm{MIC}}}, \tag{6}$$
$$\mathrm{radius}_{\mathrm{MIC}} = \frac{\sum_{j=1}^{n} 2^{-\lambda(t-T_{ij})}\mathrm{distance}\left(p_{ij}, c_{\mathrm{MIC}}\right)}{w_{\mathrm{MIC}}},$$

where distance$(\mathrm{center}_{\mathrm{MIC}}, p_{ij})$ is an Euclidean distance between the center of minicluster and the data points in that grid cell.

Definition 9 (grid synopsis). Is a tuple GS(n_g, t_p, w_g) where n_g is the number of data points, t_p is the last timestamp and w_g is the grid weight.

Definition 10 (outlier weight threshold (OWT)). This threshold is considered for the sparse grids which do not receive any data for long. In fact, these grids do not have any chance to be converted to dense grids and consequently to MIC. If the grid weight is less than this threshold, it can safely be deleted from the grid list (in the outlier buffer) [14]. If the last updated time of grid g is t_p, then, at current time t_c, the outlier weight threshold is defined as follows $(t_c > t_p)$:

$$\mathrm{OWT}\left(t_c, t_p\right) = \frac{\alpha}{N}\sum_{i=0}^{t_c-t_p} 2^{-\lambda i} = \frac{\alpha\left(1-2^{-\lambda(t_c-t_p+1)}\right)}{N\left(1-2^{-\lambda(t_p)}\right)}. \tag{7}$$

Definition 11 (pruning time). We check all MICs' weights as well as the weights of all grid cells in a time we call it t_{pt}. t_{pt} is the minimum time for a MIC in timestamp t_1 to be converted to an outlier in t_2 $(t_2 > t_1)$ which is described as follows:

Lemma 12.

$$t_{pt} = \log_{\lambda}^{\alpha/(\alpha-N(1-2^{-\lambda}))}. \tag{8}$$

Proof.

$$w_{\mathrm{MIC}}\left(t_2\right) = 2^{-\lambda(t_2-t_1)} * w_{\mathrm{MIC}}\left(t_1\right) + 1,$$
$$\frac{\alpha}{N\left(1-2^{-\lambda}\right)} = 2^{-\lambda(t_2-t_1)}\frac{\alpha}{N\left(1-2^{-\lambda}\right)} + 1, \qquad t_{pt} = t_2 - t_1,$$
$$t_{pt} = \log_{\lambda}^{\alpha/(\alpha-N(1-2^{-\lambda}))}. \tag{9}$$

□

4. HDC-Stream Algorithm

HDC-Stream is a hybrid density-based clustering algorithm for evolving data streams. The overall architecture of HDC-Stream algorithm is outlined in Algorithm 1. It has an online-offline component. For a data stream, at each timestamp, the online component of HDC-Stream continuously reads a new data record and either adds it to an existing minicluster or maps it to the grid. In pruning time, HDC-Stream periodically removes real outliers. The offline component generates the final clusters on demand by the user. The procedure adopted in this algorithm is divided into three steps as follows. The steps are also illustrated in Figure 2.

(1) Merging or papping (MM-Step): the new data point is added to an existing minicluster or mapped to the grid (lines 5–18 of Algorithm 1).

(2) Pruning grids and miniclusters (PGM-Step): the grids cells as well as miniclusters' weights are periodically checked in pruning time. The periods are defined based on the minimum time for a minicluster to be converted to an outlier. The grids and the miniclusters with the weights less than a threshold are discarded, and the memory space is released (lines 19–33 of Algorithm 1).

(3) Forming final clusters (FFC-Step): final clusters are formed based on miniclusters which are pruned. Each minicluster is clustered as a virtual point using a modified DBSCAN (lines 34–36 of Algorithm 1).

The steps are explained as follows.

4.1. MM-Step of HDC-Stream. When a new data point arrives (Figure 3), we get the following.

(i) HDC-Stream finds the nearest MIC to the new data point.

(ii) If the new data point's distance to the nearest MIC is less than r_{MIC}, it will be added to that particular MIC.

(iii) Otherwise, the data point has to be mapped into the grid in the outlier buffer.

  (a) If the number of data points in grid n_g reaches MinPts, then we check the grid weight w_g.

    (1) If the grid weight w_g is higher than the dense grid threshold, then we form a new MIC out of the data points in this grid.

    (2) The related grid g of the new MIC is discarded from the grid list.

Input: a data stream, MinPts, λ, and α
Output: arbitrary shape clusters

(1) $t_{pt} = \log_{\lambda}^{\alpha/(\alpha - N(1-2^{-\lambda}))}$
(2) $t_c = 0$
(3) **while** not end of stream **do**
(4) Read data point x from Data Stream
{∗ ∗ ∗ ∗ ∗ MM-Step ∗ ∗ ∗ ∗ ∗}
(5) Find the nearest mini-cluster MIC to x
(6) **if** distance $(x, \text{center}_{\text{MIC}}) < r_{\text{MIC}}$ **then**
(7) Merge x to the MIC
(8) **else**
(9) Map the new data point x to the grid
(10) $n_g = n_g + 1; w_g = 2^{-\lambda(t_c - t_p)} w_g(t_p) + 1; t_p = t_c$
(11) Update GS(n_g, t_p, w_g)
(12) **if** $n_g \geq$ MinPts and $w_g \geq \dfrac{\alpha}{N(1-2^{-\lambda})}$ **then**
(13) $w_{\text{MIC}} = w_g$
(14) $c_{\text{MIC}} = \dfrac{\sum_{i=1}^{n} f(t_c - T_i)(p_i)}{w_{\text{MIC}}}$
(15) $r_{\text{MIC}} = \dfrac{\sum_{i=1}^{n} f(t_c - T_i)\text{distance}(p_{ij}, c_{\text{MIC}})}{w_{\text{MIC}}}$
(16) Remove grid g from the grid list
(17) **end if**
(18) **end if**
{∗ ∗ ∗ ∗ ∗ PGM-Step ∗ ∗ ∗ ∗ ∗}
(19) **if** $t \bmod t_{pt} == 0$ **then**
(20) **for all** grid g **do**
(21) $\text{OWT}(t_c, t_p) = \dfrac{\alpha(1 - 2^{-\lambda(t_c - t_p + 1)})}{N(1 - 2^{-\lambda(t_p)})}$
(22) **if** $w_g <$ OWT **then**
(23) Remove grid g from the grid list
(24) **end if**
(25) **end for**
(26) **for all** {MIC} **do**
(27) **if** $w_{\text{MIC}} < \dfrac{\alpha}{N(1-2^{-\lambda})}$ **then**
(28) Remove MIC from {MIC}
(29) **end if**
(30) **end for**
(31) **end if**
(32) $t_c = t_c + 1$
(33) **end while**
{∗ ∗ ∗ ∗ ∗ FCC-Step ∗ ∗ ∗ ∗ ∗}
(34) **if** the clustering request is arrived **then**
(35) Generate clusters using a modified DBSCAN
(36) **end if**

Algorithm 1: HDC-Stream (DS, MinPts, λ, and α).

4.2. PGM-Step of HDC-Stream. For each MIC, if no new point is added, its weight will gradually decay. Furthermore, there are some grids which do not receive data points for a long time and become sporadic. These kinds of MIC and grid cells should be removed from the miniclusters and the grid list, respectively. The decision for removing grids and miniclusters is made based on a comparison of their weights and a specified threshold. Therefore, PGM-Step is performed in each t_{pt} which is defined in Definition 11.

4.3. FCC-Step of HDC-Stream. When a clustering request arrives, a variant of DBSCAN algorithm is applied on the set of the online maintained miniclusters to get the clustering result. Each minicluster MIC is considered as a virtual point located at the center of MIC with the weight w_{MIC}. We adopt the concept of density connectivity from [21], in order to determine the final clusters. All the density-connected MICs form a cluster. The variant of DBSCAN algorithm includes two parameters: ϵ and MinPts.

Definition 13 (directly density-reachable). A MIC_p is directly density-reachable from a MIC_q with respect to ϵ and MinPts if $\text{dist}(\text{Center}_{\text{MIC}_p}, \text{Center}_{\text{MIC}_q}) < r_{\text{MIC}_p} + r_{\text{MIC}_q}$. $\text{Dist}(\text{Center}_{\text{MIC}_p}, \text{Center}_{\text{MIC}_q})$ is the Euclidean distance between the centers of MIC_p and MIC_q.

Definition 14 (density-reachable). A MIC_p is density-reachable from a MIC_q with respect to ϵ and MinPts if there is a chain of miniclusters $\text{MIC}_1, \ldots, \text{MIC}_n$, such that $\text{MIC}_1 = \text{MIC}_q$ and $\text{MIC}_n = \text{MIC}_p$ ($\text{MIC}_{p_{i+1}}$ is directly density reachable from MIC_{p_i}).

Definition 15 (density-connected). A MIC_p is density-connected to a MIC_q with respect to ϵ and MinPts if there is a minicluster MIC_k such that both MIC_p and MIC_q are density-reachable from MIC_k with respect to ϵ and MinPts.

5. Experimental Evaluation

In this section, we present the evaluation of HDC-Stream with respect to two existing well-known methods DenStream and D-Stream. We have implemented HDC-Stream as well as the comparative methods in Java. All experiments were conducted on a 2.5 GHz machine with 4 GB memory, running on Mac OS X. In this section, firstly, we describe the datasets and then evaluation measures used for the evaluation of the HDC-Stream algorithm. Detailed experiments on real and synthetic datasets are discussed as well.

5.1. Datasets. For evaluation purposes, the clustering quality, scalability, and sensitivity of the HDC-Stream algorithm on both real and synthetic datasets are used. We generated three synthetic datasets DS1, DS2, and DS3 which are depicted in Figures 4(a), 4(b), and 4(c), respectively. DS1 has 10000 data points with 5% noise. DS2 has 10000 data points with 4% noise, and DS3 has 10000 data points with 5% noise. Eventually, we generated an evolving data stream (EDS) by randomly selecting one of the datasets (DS1, DS2, and DS3) 10 times. For each iteration, the chosen dataset forms a 10000-point part of the data stream, so the total length of the evolving data stream is 100000.

The real dataset used is KDD CUP99 Network Intrusion Detection dataset (all 34 continuous attributes out of the total 42 available attributes are used) [29]. The dataset comes from the 1998 DARPA Intrusion Detection. It contains training

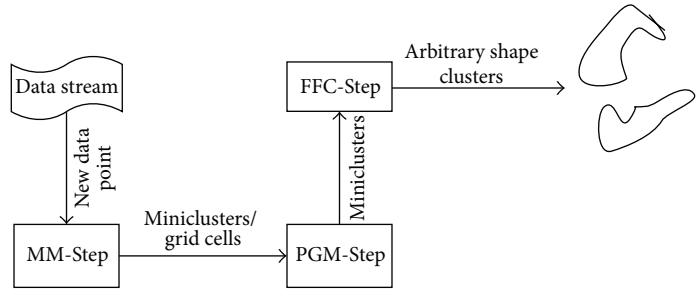


Figure 2: Overall view of HDC-Stream algorithm.

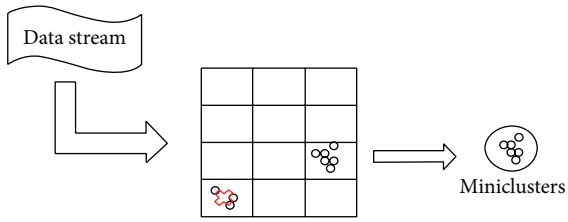


Figure 3: MM-Step of HDC-Stream algorithm.

data consisting of 7 weeks of network-based intrusions inserted in the normal data and 2 weeks of network-based intrusions and normal data for a total of 4,999,000 connection records described by 42 characteristics. KDD CUP99 has been used in [14, 17, 24, 27] and it is converted into data stream by taking the data input order as the order of streaming.

5.2. Evaluation Metrics. Cluster validity is an important issue in cluster analysis. Its objective is to assess clustering results of the proposed algorithm by comparing existing well-known clustering algorithms. In the following, we adopt two popular measures, purity and normalized mutual information (NMI), in order to evaluate the quality of HDC-Stream.

5.2.1. Purity. The clustering quality is evaluated by the average purity of clusters which is defined as follows:

$$\text{purity} = \frac{\sum_{i=1}^{K} \left(\left|C_i^d\right| / \left|C_i\right|\right)}{K} * 100\%, \tag{10}$$

where K is number of clusters, $|C_i^d|$ is the number of points with the dominant class label in cluster i, and $|C_i|$ is the number of points in cluster i. The purity is calculated only for the points arriving in a predefined window (H), since the weight of points diminishes continuously.

5.2.2. Normalized Mutual Information (NMI). The normalized mutual information (NMI) is a well-known information theoretic measure that assesses how similar two clusterings are. Given the true clustering $A = \{A_1, \ldots, A_k\}$ and the grouping $B = \{B_1, \ldots, B_h\}$ obtained by a clustering method, let C be the confusion matrix whose element C_{ij} is the number of records of cluster i of A that are also in the cluster j of B. The normalized mutual information, NMI(A, B), is defined as

$$\text{NMI}(A, B) = \frac{-2 \sum_{i=1}^{c_A} \sum_{j=1}^{c_B} C_{ij} \log\left(c_{ij} N / C_i \cdot C_j\right)}{\sum_{i=1}^{c_A} C_i \log\left(C_i / N\right) + \sum_{j=1}^{c_B} C_j \log\left(C_j / N\right)}, \tag{11}$$

where $c_A(c_B)$ is the number of groups in the partition $A(B)$, $C_i(C_j)$ is the sum of the elements of C in row i (column j), and N is the number of data points. If $A = B$, NMI(A, B) = 1, and if A and B are completely different, NMI(A, B) = 0.

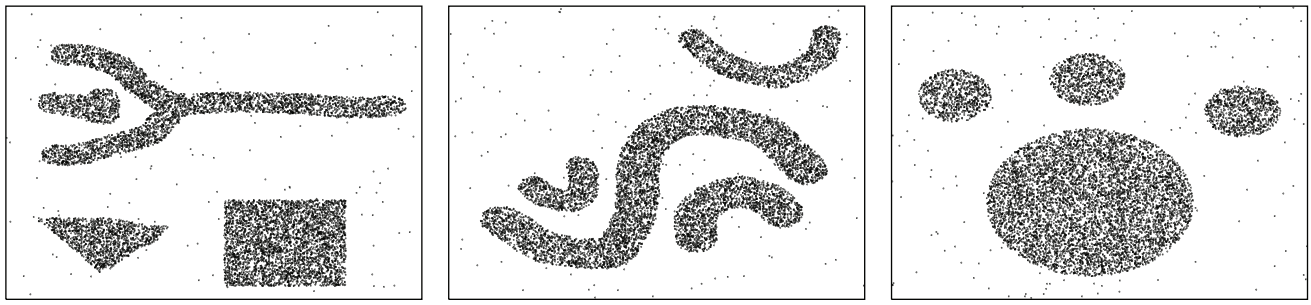
The parameters of HDC-Stream adopt the following settings: decay factor $\lambda = 0.25$, minimum number of points MinPts = 30, and $\alpha = 0.8$. The parameters for DenStream and D-Stream are chosen to be the same as those adopted in [24] and [14], respectively.

5.3. Evaluation of HDC-Stream on Synthetic Datasets. Figure 5 shows the purity results of HDC-Stream compared to DenStream and D-Stream on EDS data stream. In Figure 5(a), the stream speed is set to 2000 points per time unit and horizon $H = 1$. HDC-Stream shows a good clustering quality. Its clustering purity is higher than 97%. We also set the stream speed at 2000 points per time unit and horizon $H = 10$ for EDS. Figure 5(b) shows similar results too. We conclude that HDC-Stream achieves much higher clustering quality than DenStream and D-Stream in two different horizons. For example, in horizon $H = 1$, time unit 50, HDC-Stream has 98% while DenStream and D-Stream have purity values as 82% and 78%, respectively.

The same is observed from the normalized mutual information aspect. In fact, Figure 6 shows the NMI values obtained by three methods. We repeated the experiments with the same horizon and stream speed (Figures 6(a) and 6(b)). The results show a noticeable high NMI score for HDC-Stream. In fact, its value approaches 1 for both horizons. It also proves that DenStream has better NMI compared to D-Stream.

We noted very good clustering quality of HDC-Stream, D-Stream, and DenStream when no noise is present in the dataset. In fact, purity values are always higher than 98% and all methods are insensitive to the horizon length.

5.4. Evaluation of HDC-Stream for Real Datasets. The comparison results among HDC-Stream and both DenStream



(a) Dataset DS1—10000 data points, 3% noise (b) Dataset DS2—10000 data points, 4% noise (c) Dataset DS3—10000 data points, 5% noise

Figure 4: Synthetic datasets.

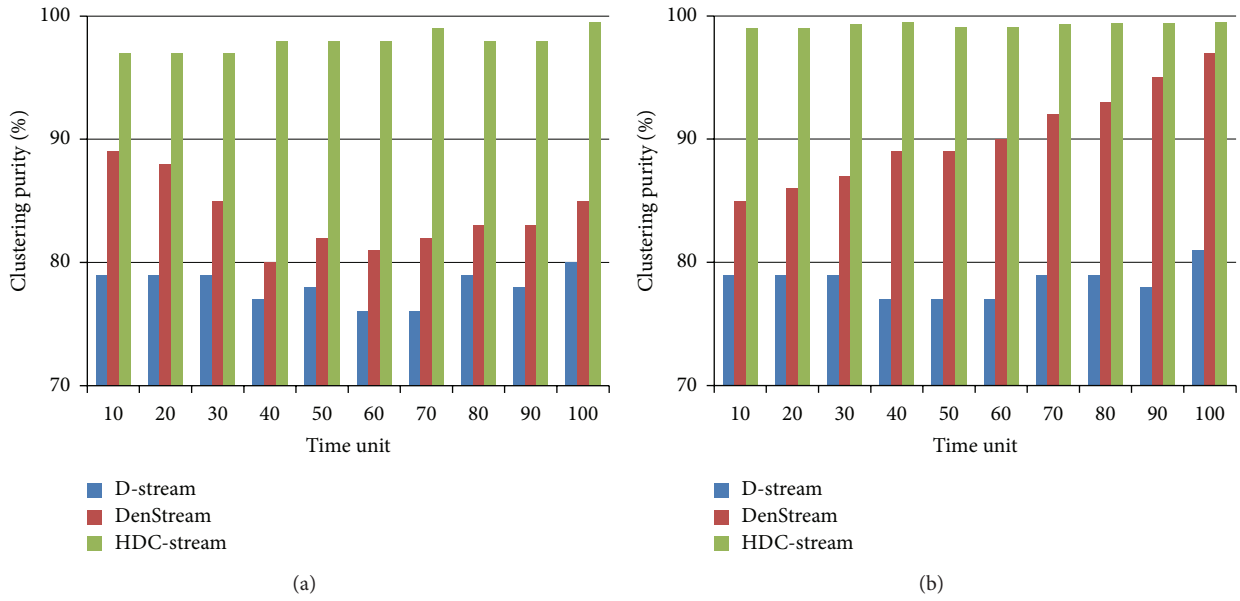


Figure 5: Cluster purity of HDC-Stream for EDS with (a) horizon = 1 and stream speed = 2000 and (b) horizon = 5 and stream speed = 2000.

and D-Stream on the Network Intrusion dataset are shown in Figure 7. The evaluation is defined based on the selected time units when the attacks happen on horizons 2 and 5, whereas the stream speed is 1000. For instance, in horizon $H = 5$ and stream speed 1000, there are 99 teardrop attacks, 182 ipsweep attacks, 618 neptune attacks, and 4097 normal connections. HDC-Stream clearly outperforms DenStream and specifically D-Stream. The purity of HDC-Stream is always above 91%. For example, at time 55, the purity of HDC-Stream is about 95% which is higher than both DenStream (86%) and D-Stream (76%).

We show the normalized mutual information results on Network Intrusion Detection dataset in Figure 8. The results have been determined by setting the horizon to 1 and 5, whereas the stream speed is 1000 (Figures 8(a) and 8(b)). The values of normalized mutual information for HDC-Stream approach 1 for both horizons. It reveals that HDC-Stream detects the true class labels of data more accurately than DenStream and D-Stream do.

5.5. Scalability Results

5.5.1. Execution Time. The execution time of HDC-Stream is influenced by the number of data points processed at each time unit, that is, the stream speed. Figure 9 shows the execution time in seconds on Network Intrusion Detection dataset for HDC-Stream compared to DenStream and D-Stream, when the stream speed augments from 1000 to 10,000 data items.

DenStream has higher processing time due to its merging task which is time consuming. HDC-Stream has lower execution time compared to the others. The execution time of other methods increases linearly with respect to the stream speed.

5.5.2. Memory Usage. Memory usage of HDC-Stream is $o(\text{mi} + g)$ which is the total number of miniclusters and grids.

5.6. Sensitivity Analysis. An important parameter of HDC-Stream is λ. It controls the importance of historical data. We

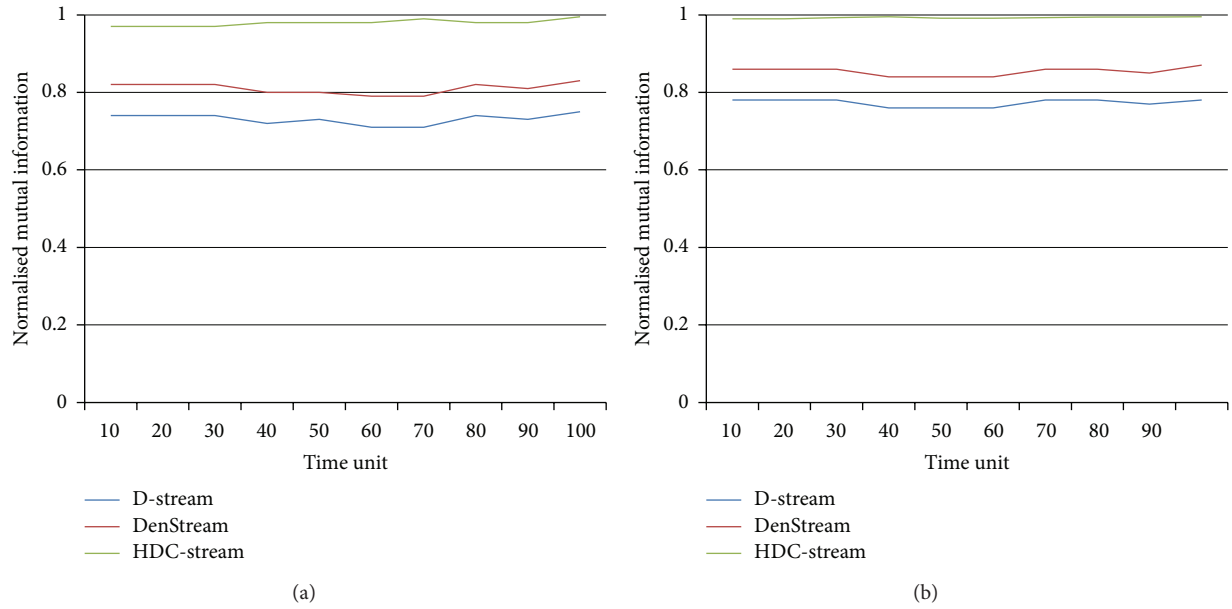


Figure 6: Normalised mutual information of HDC-Stream for EDS with (a) horizon = 1 and stream speed = 2000 and (b) horizon = 5 and stream speed = 2000.

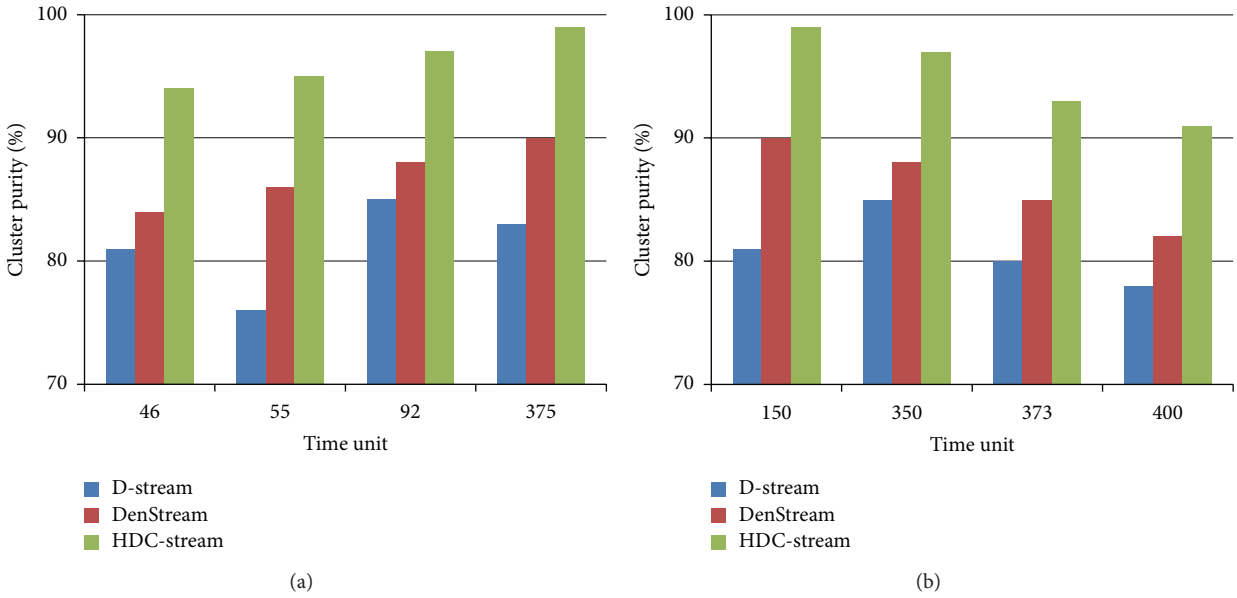


Figure 7: Cluster purity of HDC-Stream for Network Intrusion Detection dataset with (a) horizon = 2 and stream speed = 1000 and (b) horizon = 5 and stream speed = 1000.

test the quality of clustering on different values of λ ranging from 0.0078 to 1 (Figure 10). When λ is too small or too large, the clustering quality becomes poor. For example, when $\lambda = 0.0078$, the purity is about 75%, and, when $\lambda = 0.5$, the points decay soon after their arrival, and only a small number of recent points contribute to the final results. So the result is not very good. However, the quality of HDC-Stream is still higher than that of DenStream and D-Stream. It is proved that if λ varies from 0.0625 to 0.25, the clustering quality is quite good, stable, and always above 96%.

6. Discussion

We proposed a hybrid method for clustering evolving data streams which has high quality and low computation time compared to existing methods. The algorithm clusters data streams in three distinctive steps. In existing methods such as DenStream, when a new data point arrives, it takes time to search in two lists of microclusters including potentials and outliers in order to find the suitable microcluster. If it is unable to find a microcluster, DenStream forms a new

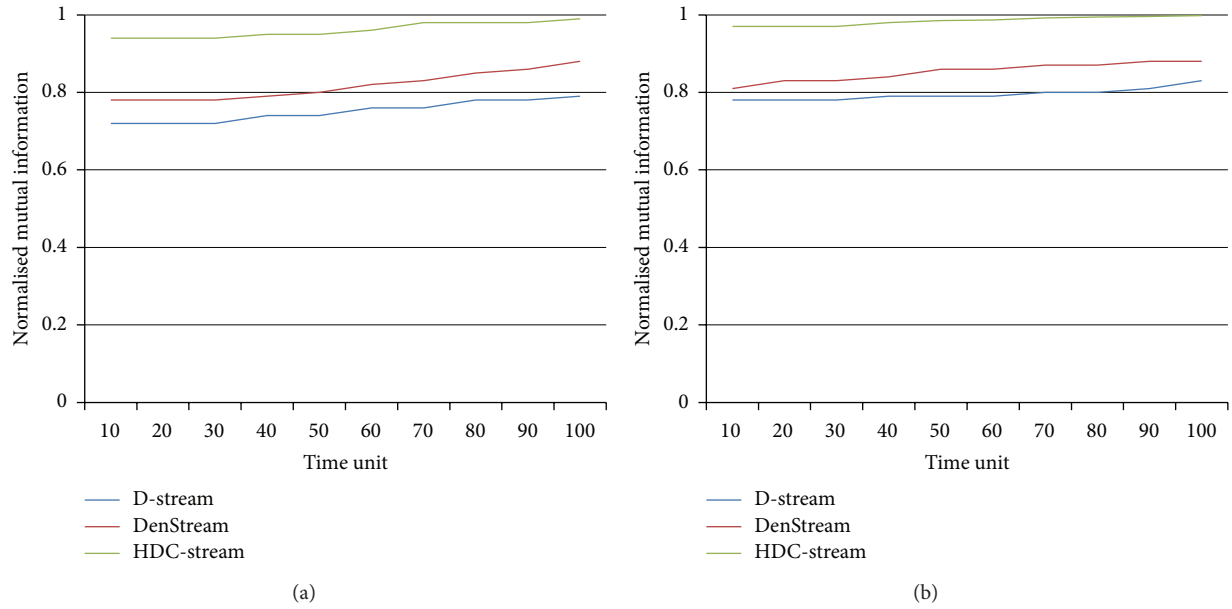


Figure 8: Normalised mutual information of HDC-Stream on Network Intrusion Detection dataset with (a) horizon = 1 and stream speed = 1000, (b) horizon = 5 and stream speed = 1000.

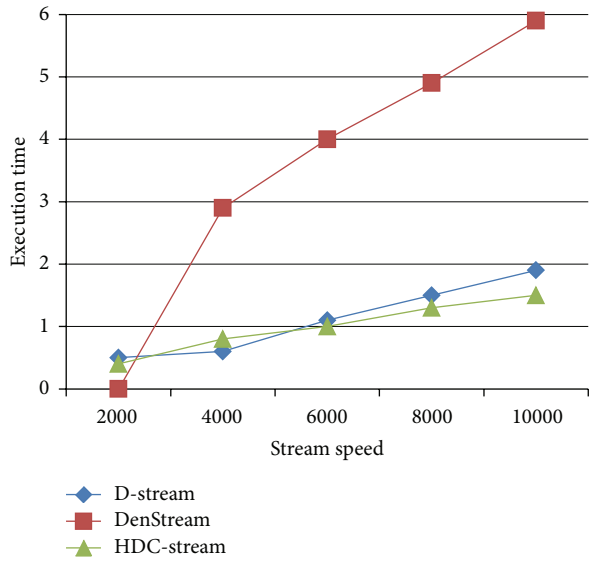


Figure 9: Execution time for increasing stream lengths on Network Intrusion Detection dataset.

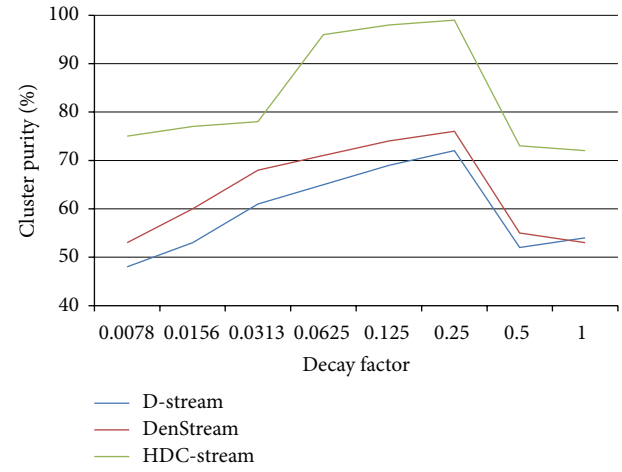


Figure 10: Cluster quality versus decay factor.

microcluster for that data point which may be a seed of an outlier, hence leading to a low clustering quality result. However, HDC-Stream only searches in potential list and if it cannot find the suitable microcluster, the data point is mapped to the grid, which keeps the outlier buffer. We reduced the time complexity of clustering algorithm using grid-based clustering. The grid-based method allows us to decrease merging time complexity from $o(\text{mi})$ to $o(1)$. We implemented the grid list in a 2-3-4 tree data structure which makes search and update faster. The size of the grid list is $o(\log_{1/\lambda} N)$ and the time required for search and update in the grid list is $o(\log \log_{1/\lambda} N)$. Consider

$$o(\text{MM-Step}) = o(\text{mi}) + o\left(\log_{1/\lambda} N\right) + o(1),$$
$$o(\text{PGM-step}) = o\left(\log_{1/\lambda} N\right) + o(\text{mi}). \quad (12)$$

We reduced the number of comparisons; therefore, time complexity for merging to minicluster list is $o(\text{mi})$; in which the number of mi is less than number of microclusters in DenStream, since, in that algorithm, there are two lists to keep potential and outlier microclusters. Furthermore, we increased the clustering quality by forming miniclusters from the data points that are surely not outliers. When the grid density reaches the specified threshold, the data points inside

that grid form a minicluster. Therefore, we do not need to form a minicluster for a newly arrived data if it cannot be placed in any minicluster. The quality is also increased since miniclusters are never formed from an outlier.

Finally, the evaluation results prove that using a hybrid method for clustering evolving data streams improves the clustering quality results and reduces the computation time.

7. Conclusion

In this paper, we proposed a hybrid density-based clustering algorithm for Internet of Things (IoT) streams. Our hybrid algorithm has three steps in which the new data point is either mapped to grid or merged to an existing minicluster, the outliers are removed, and finally arbitrary shape clusters are formed using miniclusters by a modified DBSCAN. Our method is a hybrid one, which uses density grid-based clustering and density microclustering to improve the computation time and quality. The evaluation results on synthetic and real datasets show that it has high quality with low computation time for merging. However, HDC-Stream is not suitable to be used in distributed environments.

Our future work will focus on the improvement of HDC-Stream as a distributed density-based data stream clustering algorithm.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research is supported by High Impact Research (HIR) Grant, University of Malaya, no. UM.C/625/HIR/MOHE/SC/13/2 from Ministry of Higher Education.

References

[1] C. Aggarwal, N. Ashish, and A. Sheth, "The internet of things: a survey from the data-centric perspective," in *Managing and Mining Sensor Data*, C. C. Aggarwal, Ed., pp. 383–428, Springer, New York, NY, USA, 2013.

[2] C.-W. Tsai, C.-F. Lai, M.-C. Chiang, and L. Yang, "Data mining for internet of things: a survey," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 77–97, 2014.

[3] R. Lacuesta, G. Palacios-Navarro, C. Cetina, L. Peñalver, and J. Lloret, "Internet of things: where to be is to trust," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, article 203, pp. 1–6, 2012.

[4] S. Bin, L. Yuan, and W. Xiaoyi, "Research on data mining models for the internet of things," in *Proceedings of the 2nd International Conference on Image Analysis and Signal Processing (IASP '10)*, pp. 127–132, April 2010.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 3rd edition, 2011.

[6] C. C. Aggarwal and C. K. Reddy, Eds., *Data Clustering: Algorithms and Applications*, CRC Press, 2013.

[7] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams," in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS '00)*, pp. 359–366, IEEE Computer Society, Washington, DC, USA, November 2000.

[8] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The ClusTree: indexing micro-clusters for anytime stream mining," *Knowledge and Information Systems*, vol. 29, no. 2, pp. 249–272, 2011.

[9] L. Wan, W. K. Ng, X. H. Dang, P. S. Yu, and K. Zhang, "Density-based clustering of data streams at multiple resolutions," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 3, article 14, pp. 1–28, 2009.

[10] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proceedings of the 29th International Conference on Very Large Data Bases*, pp. 81–92, VLDB Endowment, 2003.

[11] M. Ester, H. P. Kriegel, J. Sander, M. Wimmer, and X. Xu, "Incremental clustering for mining in a data warehousing environment," in *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB '98)*, pp. 323–333, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 1998.

[12] A. Amini, Y. W. Teh, and H. Saboohi, "On density-based data streams clustering algorithms: a survey," *Journal of Computer Science and Technology*, vol. 29, no. 1, pp. 116–141, 2014.

[13] A. Amini, T. Y. Wah, M. R. Saybani, and S. R. A. S. Yazdi, "A study of density-grid based clustering algorithms on data streams," in *Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '11)*, pp. 1652–1656, IEEE, Shanghai, China, July 2011.

[14] L. Tu and Y. Chen, "Stream data clustering based on grid density and attraction," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 3, article 12, pp. 1–27, 2009.

[15] V. Bhatnagar, S. Kaur, and S. Chakravarthy, "Clustering data streams using grid-based synopsis," *Knowledge and Information Systems*, pp. 1–26, 2013.

[16] A. Amini and T. Y. Wah, "Density micro-clustering algorithms on data streams: a review," in *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS '11)*, pp. 410–414, Hong Kong, March 2011.

[17] A. Forestiero, C. Pizzuti, and G. Spezzano, "A single pass algorithm for clustering evolving data streams based on swarm intelligence," *Data Mining and Knowledge Discovery*, vol. 26, no. 1, pp. 1–26, 2013.

[18] C. Isaksson, M. H. Dunham, and M. Hahsler, "Sostream: self organizing density-based clustering over data stream," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed., vol. 7376 of *Lecture Notes in Computer Science*, pp. 264–278, Springer, Berlin, Germany, 2012.

[19] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1, pp. 281–297, University of California Press, 1967.

[20] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. IT-28, no. 2, pp. 129–137, 1982.

[21] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96)*, pp. 226–231, AAAI Press, Portland, Oregon, 1996.

[22] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," *ACM SIGMOD Record*, vol. 28, no. 2, pp. 49–60, 1999.

[23] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD '98)*, pp. 58–65, 1998.

[24] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *Proceedings of the 6th SIAM International Conference on Data Mining*, pp. 328–339, April 2006.

[25] J. F. Kennedy, J. Kennedy, and R. C. Eberhart, *Swarm Intelligence*, Morgan Kaufmann Publishers, 2001.

[26] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[27] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*, pp. 133–142, ACM, New York, NY, USA, August 2007.

[28] W. Ng and M. Dash, "Discovery of frequent patterns in transactional data streams," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems II*, vol. 6380 of *Lecture Notes in Computer Science*, pp. 1–30, Springer, Berlin, Germany, 2010.

[29] S. Rosset and A. Inger, "Kdd-cup 99: knowledge discovery in a charitable organization's donor database," *SIGKDD Explorations*, vol. 1, no. 2, pp. 85–90, 2000.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 161874, 11 pages
http://dx.doi.org/10.1155/2014/161874

Research Article

Smart HVAC Control in IoT: Energy Consumption Minimization with User Comfort Constraints

Jordi Serra, David Pubill, Angelos Antonopoulos, and Christos Verikoukis

Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), 08860 Castelldefels, Spain

Correspondence should be addressed to Jordi Serra; jordi.serra@cttc.es

Received 10 April 2014; Accepted 16 May 2014; Published 18 June 2014

Academic Editor: Xudong Zhu

Smart grid is one of the main applications of the Internet of Things (IoT) paradigm. Within this context, this paper addresses the efficient energy consumption management of heating, ventilation, and air conditioning (HVAC) systems in smart grids with variable energy price. To that end, first, we propose an energy scheduling method that minimizes the energy consumption cost for a particular time interval, taking into account the energy price and a set of comfort constraints, that is, a range of temperatures according to user's preferences for a given room. Then, we propose an energy scheduler where the user may select to relax the temperature constraints to save more energy. Moreover, thanks to the IoT paradigm, the user may interact remotely with the HVAC control system. In particular, the user may decide remotely the temperature of comfort, while the temperature and energy consumption information is sent through Internet and displayed at the end user's device. The proposed algorithms have been implemented in a real testbed, highlighting the potential gains that can be achieved in terms of both energy and cost.

1. Introduction

The Internet of Things (IoT) paves the way for the connection of sensors, actuators, and other objects to the Internet, permitting the perception of the world, as well as the interaction with it, in an unprecedented manner. In addition, IoT will foster a huge number of new applications, for example, environmental monitoring, healthcare, and efficient management of energy in smart homes [1], potentially generating important economic benefits [2]. Actually, the US National Intelligence Council considers IoT as one of the six disruptive civil technologies with potential impact on US national power [3]. As a result, the concept of IoT, in terms of architectural aspects, protocol stacks, applications, and conceptual visions, has recently started to be studied [4–7].

Smart grid is considered as one of the main IoT applications and it has attracted a great interest during the last few years [1, 8, 9]. The smart grid is envisioned as the evolution of the current energy grid, which faces important challenges, such as blackouts caused by peaks of energy demand that exceed the energy grid capacity [10]. A proposed approach to alleviate this problem is to incentivize the consumers to defer or reschedule their energy consumption to different time intervals with lower expected power demand. These incentives are based on smart (or dynamic) pricing tariffs that consider a variable energy price [11]. For instance, in real-time pricing (RTP) tariffs, the price of the energy will be higher at certain time periods, where the energy consumption is expected to be higher, for example, during the afternoon or in cold days. Other types of smart pricing tariffs are critical-peak pricing (CPP) or time-of-use pricing (ToUP) [11–13]. Energy scheduling algorithms are the state-of-the-art methods to manage the energy consumption of loads within a smart pricing framework [11, 12, 14–16]. These techniques assume a specific smart pricing tariff and various time periods. For each of these time intervals, the scheduler determines the operational power of each appliance to minimize the energy consumption cost. It is worth mentioning that the appliances that can be controlled by the energy scheduler can be categorized into three classes: (i) nonshiftable, which do not admit any change on their consumption profile, (ii) time-shiftable, which tolerate postponing their operation, but not their consumption profile, and (iii) power-shiftable, whose operational power can be changed.

Regarding the power-shiftable loads, heating, ventilation, and air conditioning (HVAC) modules are considered as

the most energy demanding appliances in home buildings [17, 18]. According to studies, they represent the 43% of residential energy consumption in the USA and the 61% in UK and Canada [18]. Apparently, the significant energy consumption of the HVAC systems, along with their direct influence on the user's well-being, highlights the necessity for effective HVAC management algorithms that reduce the power consumption in the home buildings, taking into account the end-user's comfort.

In this paper, we propose two HVAC energy scheduling methods in an IoT framework, where the users are able to interact remotely with the HVAC control system. In particular, the users may retrieve information about the temperature and the energy consumption at various spots of the building under control, while they are also able to remotely configure the temperature in given places. Our contribution can be summarized as follows.

(i) We propose a dynamic energy scheduler with comfort constraints (DES-CC), which considers both the smart pricing tariffs and the user's comfort, in order to select the most energy efficient configuration of HVACs that satisfies the user's needs. We formulate an optimization problem of HVAC control by predicting the temperature that a given set of HVAC modules would cause in different locations. We result in a boolean quadratic optimization problem which, although not convex, can be solved via an exhaustive search when the number of variables (i.e., HVAC modules in our case) is low. In case that a large number of HVAC modules are considered, semidefinite relaxation techniques can be applied [19].

(ii) Taking into account the energy efficiency priority, we propose a dynamic energy scheduler with comfort constraints relaxation (DES-CCR), where the user relaxes their comfort constraints, allowing a higher degree of flexibility for the system to further reduce the energy consumption. In this case, the problem is reformulated and the user's comfort (i.e., temperature) is set as a penalty in the objective function instead of constraint.

(iii) We have designed and developed a real testbed to evaluate the performance of the proposed algorithms, demonstrating the potential financial and energy gains that can be achieved.

The remainder of the paper is organized as follows. Section 2 provides a brief review of the related work in this field. Section 3 describes the general network architecture and the system model under study. Section 4 introduces the two HVAC schedulers in a smart pricing and comfort constraint context. Section 5 provides the description of the testbed and the experimental results. Finally, Section 6 concludes the paper.

2. Related Work

The energy cost management of HVAC systems has recently attracted the research attention. In [20], the energy cost is studied as a function of the parameters that control the air and water subsystems and an evolutionary programming method is proposed to save energy. Moreover, in [21], a dynamic threshold controls the energy consumption and it varies according to the user satisfaction, which also depends on a thermal model. However, neither [20] nor [21] explicitly consider a dynamic pricing cost. In [22], smart pricing is considered in the energy cost optimization, but the user comfort is not explicitly incorporated in the algorithm, as the authors consider that the HVAC is turned on/off when the indoor temperature is outside the margin of comfort. Recently, in [13, 18], both energy scheduling of HVAC under smart pricing and the user comfort are taken into account. In [13], Nguyen et al. propose the construction of a lookup table (LUT) of room temperatures that depends on (i) the past temperatures, (ii) the outdoor temperature, and (iii) the HVAC power. The authors claim that the LUT is built during a training period (that takes place only once) and permits to assess the temperature of comfort for a given operation of the HVAC energy scheduler. However, this heuristic approach seems hardly applicable in general scenarios. In [18], a linear energy cost function is considered, although quadratic or two-step piecewise linear functions are more common in practice [11], while user's comfort is measured only at a specific location. It is also worth noting that none of the aforementioned works considers an IoT framework.

Unlike [20–22], in our proposed energy scheduling methods, both the smart pricing tariffs and the user comfort are taken into account. Moreover, the temperature of comfort is measured at several building positions by different sensor nodes that form a wireless sensor network (WSN), thus providing a more accurate measure of comfort compared to [18]. Furthermore, compared to [13], we adopt a more analytical and less heuristic model to assess the user comfort in the HVAC energy cost optimization. In particular, our model considers the time varying nature of the real thermal conditions, without requiring a training period. Moreover, our model adaptively updates the past temperature measurements for each time period, whereas the model in [13] is only carried out once to construct the LUT for particular indoor and outdoor conditions. Finally, unlike most of the above references, our methods are validated in a real scenario.

3. Network Model

3.1. General Architecture. Figure 1 presents the overall architecture which is used to evaluate the proposed energy schedulers in an IoT context. It consists of the following elements:

(i) a set of HVAC modules;

(ii) a set of actuators that control the HVAC modules;

(iii) a WSN, which sends measurements of temperature and energy consumption to a gateway;

(iv) a gateway (GW) that incorporates the proposed energy scheduling methods and connects the local network to the Internet. That is, it contains a web

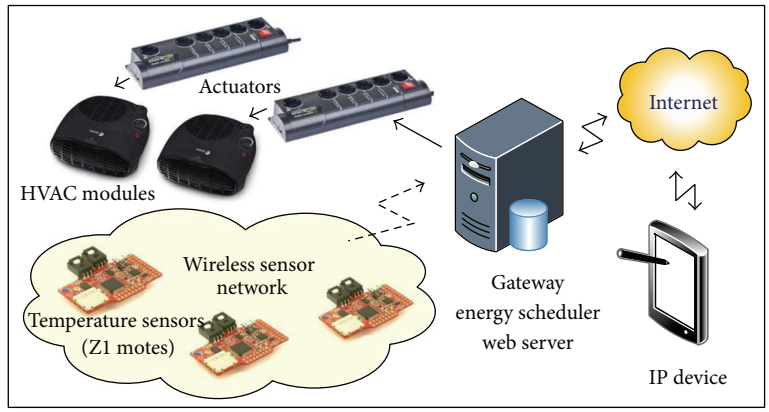


Figure 1: Overall architecture of the proposed HVAC energy scheduler in the IoT context.

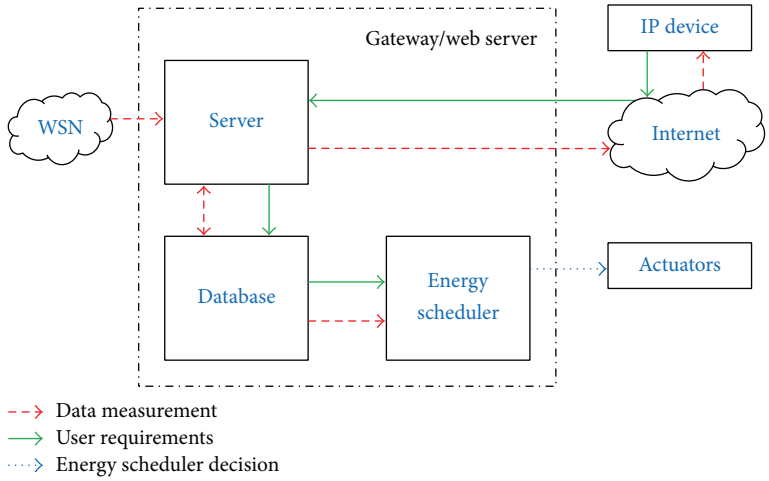


Figure 2: System model of the gateway.

server and a database to store data received at the GW from the WSN or the internet;

(v) an embedded IP device (e.g., tablet or smartphone) with an interface to interact with the HVAC energy scheduler. It also displays both the temperature and the energy consumption in the building measured by the WSN.

The functionality and flow of information of the proposed architecture is explained as follows. The temperature is measured at several locations by means of the WSN. Then, the measurements are periodically sent to the gateway, where the energy scheduling algorithm is implemented. This algorithm selects the combination of the active HVAC modules that minimizes the energy cost for given comfort constraints and energy price during a particular time period. These decisions are sent, through shell commands, to programmable surge protectors (actuators), which actuate the HVAC modules. The HVAC modules modify the room temperature according to the decisions taken by the energy scheduler.

Moreover, the gateway hosts a database to store the measurements of temperature and energy consumption. These measurements can be accessed by a remote Internet user. More specifically, they are displayed at the user's IP device, as the gateway implements a web server which manages the communication between the remote user and the local database. This is illustrated in more detail in Figure 2, where the connections between the most relevant blocks are shown. Furthermore, users are allowed to interact with the energy scheduler through their IP devices, by setting the upper and lower bounds of the temperature of comfort.

3.2. System Model. The system model for the proposed HVAC energy scheduler is depicted in more detail in Figure 3. In particular, the energy scheduler is implemented within the gateway and it interacts with the following modules. First, a WSN composed of M sensor nodes S_i, $1 \leqslant i \leqslant M$, which sense the temperature and transmit the measurements to the energy scheduler, through a WSN sink node. Second, a set of K HVAC modules that are controlled by the energy scheduler through a set of actuators A_k, $1 \leqslant k \leqslant K$. Moreover, the inputs that the energy scheduler requires are described as follows.

(i) The measurements taken by the WSN nodes. For each time interval, N measurements are taken by each node when a given configuration of HVAC modules

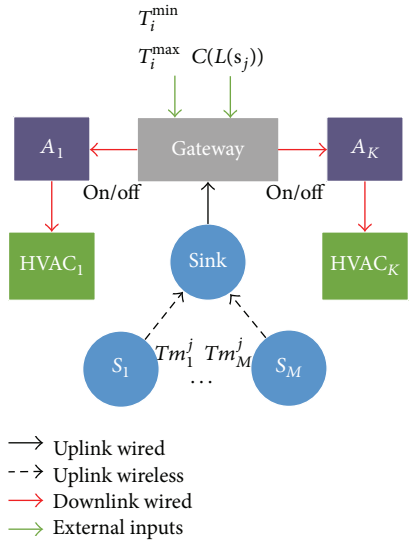


Figure 3: Detailed system model of the HVAC energy scheduler.

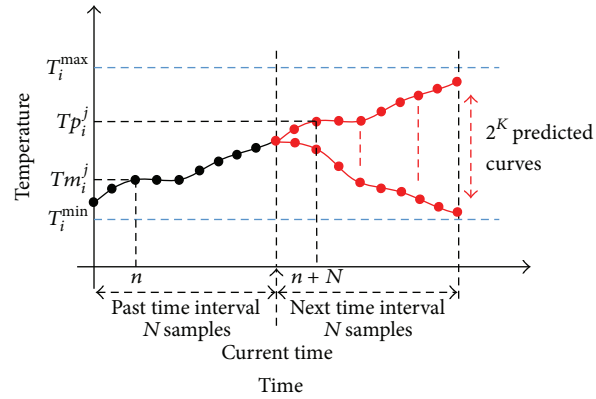


Figure 4: Prediction of temperature, a fundamental step of the energy scheduler to assess comfort in the future time interval.

is turned on. These measurements are denoted by $Tm_i^j(n)$ (as illustrated by the black curve in Figure 4), where $1 \leqslant i \leqslant M$ denotes the ith node and $1 \leqslant j \leqslant 2^K$ is the jth combination of HVACs turned on or off.

(ii) The energy cost function $C(L(\mathbf{s}_j))$, which is specified by the energy provider and depends on the smart pricing tariff. According to [11], $C(L(\mathbf{s}_j))$ can be modeled as a quadratic function, that is,

$$C\left(L\left(\mathbf{s}_j\right)\right) = p_1\left(L\left(\mathbf{s}_j\right)\right)^2 + p_2 L\left(\mathbf{s}_j\right) + p_3, \quad (1)$$

where p_1, p_2, and p_3 are parameters that the provider can dynamically vary in time. Moreover, $L(\mathbf{s}_j)$ denotes the user's energy consumption for the $\mathbf{s}_j$ combination of HVAC modules turned on. In order to describe the expression for $L(\mathbf{s}_j)$, let us define $\mathbf{P} \in \mathbb{R}^{K \times 2^K}$ a matrix that contains, in its jth column, the energy consumption of each HVAC module for the jth combination of modules switched on/off. For instance, for $K = 3$, the matrix $\mathbf{P}$ is

$$\mathbf{P} = \begin{pmatrix} 0 & x_1 & 0 & 0 & x_1 & x_1 & 0 & x_1 \\ 0 & 0 & x_2 & 0 & x_2 & 0 & x_2 & x_2 \\ 0 & 0 & 0 & x_3 & 0 & x_3 & x_3 & x_3 \end{pmatrix}. \quad (2)$$

Moreover, $\mathbf{s}_j$ is a vector of all zeros except in the jth position that has value 1. Therefore, $\mathbf{Ps}_j$ selects the energy consumption related to the jth combination of HVACs switched on and $\mathbf{1}^T(\mathbf{Ps}_j)$ is the energy consumption related to that combination where $\mathbf{1}$ is a vector of ones of length 2^K. Therefore, the total energy consumption for a given time period, denoted by $L(\mathbf{s}_j)$, can be expressed as

$$L\left(\mathbf{s}_j\right) = L_0 + \mathbf{1}^T\left(\mathbf{Ps}_j\right), \quad (3)$$

where L_0 is the accumulated consumption and $\mathbf{1}^T(\mathbf{Ps}_j)$ is the consumption in the current time interval decision.

(iii) The constraints of temperature of comfort provided by the user. For the energy scheduler proposed in Section 4.1, this corresponds to the minimum and maximum allowed temperatures at the ith location of the room, which are denoted by $T_i^{\min}$ and $T_i^{\max}$ with $1 \leqslant i \leqslant M$, respectively, since we assume that the user may specify the desired comfort at M different locations. For the energy scheduler proposed in Section 4.2, the comfort is specified by the objective temperatures $T_{u,i}$, $1 \leqslant i \leqslant M$, which the user would like to attain at the different M locations of the room, though they may allow some relaxation in order to further reduce the energy consumption.

To further clarify the operation of the proposed scheme, let us shed light on the temporal behavior of the energy schedulers and the role of the temperature constraints on it. Recall that the energy scheduler works in a time interval basis. At the end of each time interval ("current time" in Figure 4), the energy scheduler must make a new decision. That is, it must decide which HVAC modules, denoted by HVAC$_k$ in Figure 3, will be active during the next time interval. In order to make this decision, the energy scheduler should predict which would be the temperature provoked by each configuration of HVACs. As there are K HVAC modules and we assume that they are either turned on or off, this corresponds to predict 2^K curves of temperature, as it is illustrated in Figure 4. These predicted temperatures are denoted by $Tp_i^j(n)$, where i and j have the same meaning as for the case of $Tm_i^j(n)$, explained above. Finally, on one hand, the DES-CC selects the configuration of HVACs that minimizes the energy consumption cost $C(L(\mathbf{s}_j))$ within the bounds of comfort, that is, $T_i^{\min} \leqslant Tp_i^j(n) \leqslant T_i^{\max}$, while the DES-CCR selects the HVAC configuration that optimizes the tradeoff between being closer to the comfort temperatures $T_{u,i}$ and saving energy. This selection is executed by the

actuators that control the HVAC modules, which are denoted by A_k in Figure 3.

4. HVAC Energy Scheduling

4.1. Dynamic Energy Scheduler with Comfort Constraints (DES-CC). Next, we present the first of the two proposed HVAC energy schedulers. To that end, this section is divided into four parts. First, we formulate the energy scheduler as a constrained optimization problem. Second, recall that, for each time interval, the energy scheduler must decide the combination of active HVACs to minimize the energy consumption cost and fulfill the constraints of temperature of comfort. To assess these constraints, the temperature provoked in the next time interval by each configuration of HVACs turned on or off should be predicted. Thereby, the second part deals with a thermal model that paves the way to predict the future temperatures. The third part specifies how to estimate the parameters of the prediction model thanks to the measurements of temperature of the past time interval. Finally, in the fourth part, we summarize the proposed DES-CC algorithm.

4.1.1. Formulation of DES-CC as a Constrained Optimization Problem. The energy scheduler works in a time interval basis. When N samples of temperature have been collected from the WSN at the M controlled locations, the energy scheduler makes a new decision with respect to the state of the HVACs. Namely, for the next time interval, the energy scheduler selects the optimal configuration of active HVACs. This configuration, on the one hand, must minimize the energy consumption cost, while, on the other hand, it must respect the comfort constraints; that is, it should lead to predicted temperatures within the bounds of comfort. According to the definitions of the system model, this optimization problem may be formulated mathematically as

$$\begin{aligned} \underset{\mathbf{s}_j \in \{0,1\}^{2^K \times 1}}{\text{minimize}} \quad & C\left(L\left(\mathbf{s}_j\right)\right) \\ \text{subject to} \quad & T_i^{\min} \leqslant Tp_i^{\min}\left(\mathbf{s}_j\right), \quad i \in [1, M] \\ & T_i^{\max} \geqslant Tp_i^{\max}\left(\mathbf{s}_j\right), \quad i \in [1, M] \\ & \mathbf{1}^T \mathbf{s}_j = 1, \end{aligned} \tag{4}$$

where $Tp_i^{\min}(\mathbf{s}_j)$ and $Tp_i^{\max}(\mathbf{s}_j)$ are the minimum and maximum predicted temperatures, respectively, at the ith location for the jth combination of HVAC modules turned on. Given the definition of $C(L(\mathbf{s}_j))$ in (1) as a quadratic function, the optimization problem (4) has the form of a quadratic programming, but that the optimization variable is boolean. Hence, it is a boolean quadratic programming problem. The problems of this class are nonconvex and, in general, they can be solved either by a fast method that finds a local solution or by a slower method that finds the global solution. In our framework, the number of HVAC modules K is low or moderate and the latter approach is preferred; for example, the branch and bound method [23] can be used. In order to proceed, we need to model the predicted temperatures $Tp_i^{\min}(\mathbf{s}_j)$ and $Tp_i^{\max}(\mathbf{s}_j)$.

4.1.2. Model to Predict the Temperatures of the Future Time Interval. Regarding the predicted temperatures $Tp_i^{\min}(\mathbf{s}_j)$ and $Tp_i^{\max}(\mathbf{s}_j)$, they can be expressed as

$$\begin{aligned} Tp_i^{\min}\left(\mathbf{s}_j\right) &= \mathring{\mathbf{q}}_i^T \mathbf{s}_j, \\ Tp_i^{\max}\left(\mathbf{s}_j\right) &= \dot{\mathbf{q}}_i^T \mathbf{s}_j, \end{aligned} \tag{5}$$

where $\mathring{\mathbf{q}}_i$ and $\dot{\mathbf{q}}_i$ are vectors that contain the minimum and maximum predicted temperatures, respectively, for each of the possible combinations of operating HVAC modules. To further clarify, let us define $Tp_i^j(n)$; $2 \leq n \leq N$ the predicted temperature at the n time instant at the ith sensor for the jth combination of HVAC modules turned on, where $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant 2^K$. Moreover, let $Tp_i^j(n_{\min_j})$ and $Tp_i^j(n_{\max_j})$ be the minimum and maximum temperatures among $Tp_i^j(n)$; $2 \leq n \leq N$. Then, $\mathring{\mathbf{q}}_i$ and $\dot{\mathbf{q}}_i$ can be expressed as

$$\begin{aligned} \mathring{\mathbf{q}}_i^T &= \left[Tp_i^1\left(n_{\min_1}\right), \ldots, Tp_i^{2^K}\left(n_{\min_{2^K}}\right)\right], \\ \dot{\mathbf{q}}_i^T &= \left[Tp_i^1\left(n_{\max_1}\right), \ldots, Tp_i^{2^K}\left(n_{\max_{2^K}}\right)\right]. \end{aligned} \tag{6}$$

In order to proceed, a model for the predicted temperatures is necessary. Intuitively, the current temperature is correlated with the past temperature and a given combination of HVACs turned on causes a change in temperature. Moreover, the temperature dynamics are rather linear (at least locally), as it will be shown below. Therefore, the following model is proposed for the temperature prediction:

$$Tp_i^j(n) = a_i^j Tp_i^j(n-1) + \gamma_i^j, \quad 2 \leq n \leq N, \tag{7}$$

where a_i^j and γ_i^j model the relation with the past temperature and the change of temperature provoked by the jth combination of HVACs turned on, respectively. Observe that, in this expression, a_i^j and γ_i^j are unknown and must be estimated from the past measurements.

4.1.3. Estimation of the Prediction Model Parameters. In order to estimate a_i^j and γ_i^j in (7), we assume that the past measurements follow a model like (7), corrupted by noise:

$$Tm_i^j(n) = a_i^j Tm_i^j(n-1) + \gamma_i^j + w_i^j(n), \quad 2 \leq n \leq N. \tag{8}$$

Note that the evolution of the temperature is considered to be linear in (7) and (8). This is a valid assumption at least for short periods, as the real experiments that we will present in Section 5 will highlight.

For the estimation of a_i^j and γ_i^j, two situations are considered. In the first one, all the HVAC modules are switched off and, as a consequence, only a_i^j must be estimated. To that end, least squares (LS) estimator is considered as no probabilistic assumptions regarding the data are needed. This

Process: For *each time interval*, do
(1) **Input**:
  (a) The measurements of temperature of the previous interval.
  (b) Value of p_1, p_2 and p_3 in the energy cost function (1).
  (c) The user's temperature of comfort constraints at each location, that is, $T_i^{\min}, T_i^{\max}, i = 1, \ldots, M$ in (4).
(2) **Estimation Step**
  Estimate a_i^j and γ_i^j in the prediction model (7) using (9) to (12) and the temperature measurements from the past time interval.
(3) **Prediction Step**
  (a) Substitute the estimation of a_i^j and γ_i^j into the prediction model (7).
  (b) Iterate this model to populate the vectors of minimum and maximum predicted temperatures (6).
(4) **Optimization Step**
  (a) Substitute the predicted temperatures (5) and the quadratic cost function (1) into the optimization problem (4)
  (b) Solve (4) using Branch and bound method [23].
(5) **Output**
  The configuration of HVACs turned on/off that optimizes (4).

Algorithm 1: Dynamic energy scheduler with comfort constraints (DES-CC).

estimator minimizes the LS error criterion, though it is not optimal in general [24]. Given (8), the LS estimation of a_i^j, denoted by $\widehat{a}_i^j$ is given by

$$\widehat{a}_i^j = \mathbf{x}^{\#}\mathbf{y}, \tag{9}$$

where the symbol # denotes the pseudoinverse operator, which is defined as $\mathbf{x}^{\#} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$, and we define $\mathbf{x} = [Tm_i^j(1), \ldots, Tm_i^j(N-1)]^T$ and $\mathbf{y} = [Tm_i^j(2), \ldots, Tm_i^j(N)]^T$. The second situation is that some of the HVAC modules were switched on. In this case, an LS estimation is considered as well. Namely, let us denote by $\widehat{\gamma}_i^j \mid \check{a}_i^j$ the estimation of γ_i^j conditioned to the knowledge of a past estimation of a_i^j, denoted by $\check{a}_i^j$. Then, the LS estimation for $\widehat{\gamma}_i^j \mid \check{a}_i^j$ yields

$$\widehat{\gamma}_i^j \mid \check{a}_i^j = \mathbf{1}^{\#}\mathbf{z}, \tag{10}$$

where $\mathbf{1}$ is a vector of ones of length $N-1$ and $\mathbf{z}$ is given by

$$\begin{aligned}\mathbf{z} = &\left[Tm_i^j(2), \ldots, Tm_i^j(N)\right]^T \\ &- \check{a}_i^j\left[Tm_i^j(1), \ldots, Tm_i^j(N-1)\right]^T.\end{aligned} \tag{11}$$

Finally, given the estimation of γ_i^j in (10), we can update the estimation of a_i^j as

$$\widehat{a}_i^j \mid \widehat{\gamma}_i^j = \mathbf{x}^{\#}\widetilde{\mathbf{y}}, \tag{12}$$

where $\widetilde{\mathbf{y}} = [Tm_i^j(2) - \widehat{\gamma}_i^j, \ldots, Tm_i^j(N) - \widehat{\gamma}_i^j]^T$.

4.1.4. Summary of the DES-CC Energy Scheduler. At this point, all the terms in the optimization problem under study, that is, (4), are specified. The procedure to implement the proposed energy scheduler for each time interval is summarized in Algorithm 1,

4.2. Dynamic Energy Scheduler with Comfort Constraints Relaxation (DES-CCR). Despite its effectiveness and its obvious advantages, the proposed energy scheduling algorithm is completely focused on the temperature constraints, neglecting the price aspects of the problem. More specifically, although there could be time periods where the energy price increases, the energy scheduler switches on the same combination of HVAC modules, in order to respect the temperature constraints. However, in such cases, users might compromise their comfort preferences to decrease the energy consumption. In order to allow the user to have more flexibility in the energy consumption, a new energy scheduler will be presented in this section. This flexibility is implemented in terms of relaxing the temperature constraints to further reduce the energy consumption.

This new energy scheduler is formulated so that the user temperature constraints in (4) are skipped and they are incorporated as a penalty term in the objective function. Consequently, the new optimization problem can be written as

$$\begin{aligned}\underset{\mathbf{s}_j \in \{0,1\}^{2^K \times 1}}{\text{minimize}} \quad & \theta \frac{C\left(L\left(\mathbf{s}_j\right)\right)}{\alpha} + (1-\theta) \\ & \times \frac{\sum_{i=1}^{M}\sum_{n=2}^{N}\left\|\mathbf{q}_i^T(n)\mathbf{s}_j - T_{u,i}\right\|^2}{\beta},\end{aligned} \tag{13}$$

where $C(L(\mathbf{s}_j))$ is the energy cost function, defined in (1). The vector $\mathbf{q}_i^T(n)$ is defined as

$$\mathbf{q}_i^T(n) = \left[Tp_i^1(n), \ldots, Tp_i^{2^K}(n)\right], \tag{14}$$

and recall that $Tp_i^j(n)$ is the predicted temperature at the ith location for the jth combination of HVACs modules turned

Process: For *each time interval*, do
(1) **Input**:
(a) The measurements of temperature of the previous interval.
(b) Value of p_1, p_2 and p_3 of the energy cost function $C(L(s_j))$ defined in (1).
(c) The user's temperature of comfort at each location, that is, $T_{u,i}$, $i = 1, \ldots, M$ in (13).
(d) The parameter $\theta \in (0, 1)$ controlling the comfort relaxation in (13).
(2) **Estimation Step**
Estimate a_i^j and γ_i^j in the prediction model (7) using (9)
to (12) and the temperature measurements from the past time interval.
(3) **Prediction Step**
(a) Substitute the estimation of a_i^j and γ_i^j into the prediction model (7).
(b) For $i = 1$ to M
For $n = 2$ to N
Compute and store the vector of predicted temperatures
$\mathbf{q}_i^T(n)$ in (14) using (7).
End For n
End For i
(4) **Optimization Step**
(a) Compute the cost function in (13) using the vectors in
the step 3(b) and the inputs in (1).
(b) Solve the optimization problem in (13) using Branch and bound method [23].
(5) **Output**
The configuration of HVACs turned on/off that optimizes (13).

Algorithm 2: Dynamic energy scheduler with comfort constraints relaxation (DES-CCR).

on or off, see (7). Moreover, α and β are normalizing constants to adjust the values of the two terms in (13). Indeed, we set their value as

$$\alpha = C\left(L\left(\mathbf{s}_{2^K}\right)\right)$$
$$\beta = \max_{\mathbf{s}_j \in \{0,1\}^{K\times 1}} \sum_{i=1}^{M}\sum_{n=2}^{N} \left\| \mathbf{q}_i^T(n)\mathbf{s}_j - T_{u,i} \right\|^2, \quad (15)$$

where $C(L(\mathbf{s}_{2^K}))$ is the cost for all the HVAC modules turned on. The term $T_{u,i}$ is the desired temperature that the user would like to maintain at the ith location of the room. Clearly, our reformulation balances the two optimization problems, that is, the energy cost minimization and the user comfort maximization. Note that the user comfort is defined as an Euclidean norm, but it can eventually be redefined with another distance measurement.

Finally, $\theta \in (0, 1)$ is defined by the user according to their preferences. For example, in the extreme case, where $\theta = 0$, the demand response algorithm will not consider any price and it will directly control the HVAC modules so that the desired temperature is reached. On the contrary, when $\theta = 1$, the HVAC modules will always remain off. In this context, the users should set the θ value according to their preferences and experience. The DES-CCR energy scheduler is summarized in Algorithm 2.

5. Experimental Results

In order to emulate the complete communication in an IoT framework, we have designed and developed a custom testbed that integrates the described architecture. In our experiments, we focus on a heating system, although the proposed algorithms apply in general HVAC systems. In this section, we describe the testbed platform and the experimental scenario, we define a baseline thermostat model, and, finally, we present the experimental results of our proposed algorithms.

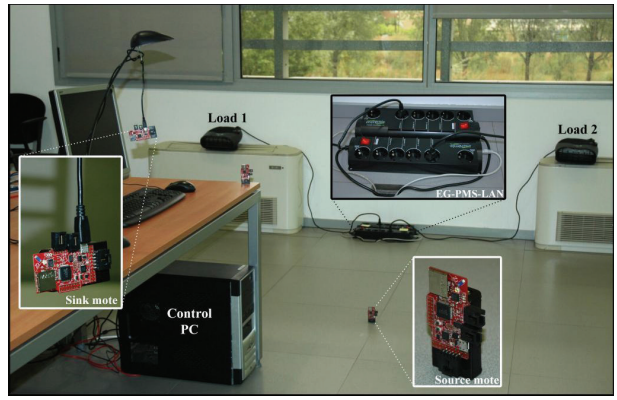


Figure 5: Z1 WSN mote.

5.1. Testbed Description and Experimental Setup. The testbed has been deployed in a 50 m^2 room within our research center facilities, as it is depicted in Figure 5. In our particular scenario, we consider three HVAC modules (i.e., $K = 3$) and two temperature sensor nodes (i.e., $M = 2$). The HVAC modules are distributed around the room, while the sensor nodes are placed in the middle of the room, monitoring the temperature and sending it to the sink mote every 30 second

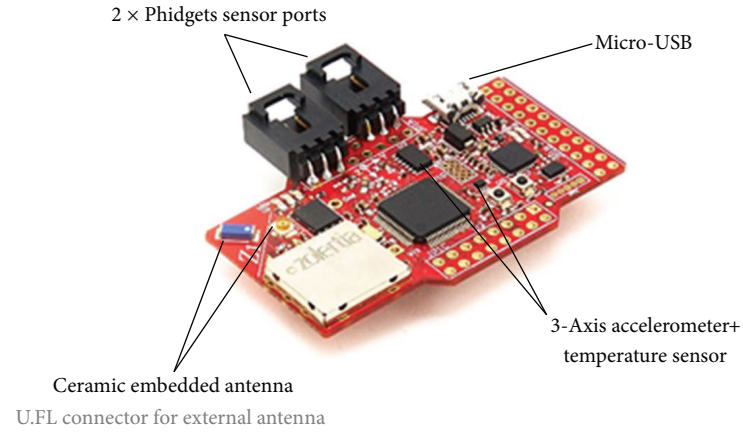


Figure 6: Overall platform detail.

(i.e., $t_m = 30$ s). In addition, the samples received from each sensor are stored in a buffer at the control center and our algorithm applies every 10 samples (i.e., $N = 10$).

Regarding the employed technology, the WSN nodes are Z1 motes by Zolertia (Figure 6). They are equipped with a second generation MSP430F2617 low power microcontroller, which features a 16-bit RISC CPU @16 MHz clock speed, a built-in clock factory calibration, an 8 KB RAM and a 92 KB flash memory. They also include the CC2420 transceiver, which is IEEE 802.15.4 compliant, operating at 2.4 GHz frequency band with a data rate of 250 kbps. The sensors support Contiki OS [25], an open-source operating system for the IoT, which connects tiny, low-cost, low-power microcontrollers to the Internet and supports IPv6 through 6LowPAN. It is worth noting that each mote can operate as either a source or a sink node. In particular, source nodes carry a TMP102 temperature sensor to monitor the target field, while the sink node receives and forwards the measured data to the gateway.

The gateway (an Ubuntu OS machine with MATLAB) implements the proposed algorithms and it is able to process the collected data. Furthermore, it connects the WSN to the Internet and acts as an application server, using Node.js and Sencha Touch. In particular, Node.js is a platform built on Chrome's JavaScript runtime for fast and scalable network applications. Figure 7 shows a screenshot of the web application built on Node.js, which enables the user to interact with the energy scheduler through Internet. Regarding Sencha Touch, it is a high-performance HTML5 mobile application framework, which enables developers to build powerful applications for various operating systems, including iOS and Android. The actuators are programmable sockets which can be controlled remotely thanks to their IP addresses. These special sockets are a set of programmable local area network surge protectors (EG-PMS-LAN) by Energenie which are connected via Ethernet to the gateway. Finally, the HVAC modules are domestic heaters with a maximum power consumption of 2000 W.

5.2. Baseline Model. To evaluate the proposed algorithms and highlight the potential energy and cost gains that can be achieved, we adopt the traditional thermostat model as the baseline reference scenario. In this model, the aim is to maintain the average temperature of the room between a certain temperature range (i.e., $[T_{\min}, T_{\max}]$), predefined by the user. To that end, when the sensed temperature is above $T_{\max}$ at the end of a time interval, all the heaters are switched off, while the heaters are switched on when the temperature falls below the $T_{\min}$ threshold.

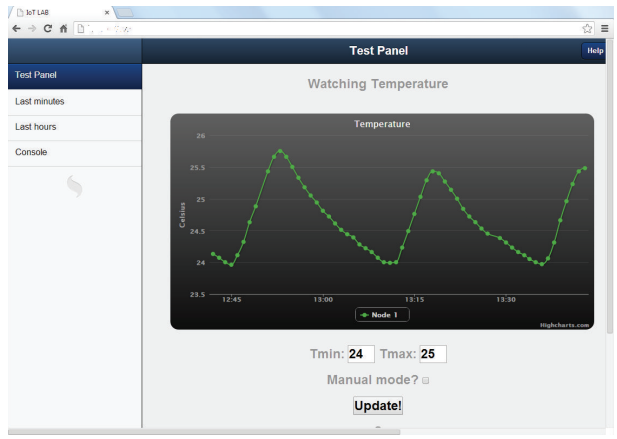


Figure 7: Google Chrome screenshot of the web application.

Figure 8 illustrates the average measured temperature inside the room, where the heaters are controlled by the thermostat. In this particular case, we consider $T_{\min} = 21^{\circ}$C and $T_{\max} = 23^{\circ}$C. As it can be seen in the figure, the thermostat algorithm is able to maintain the average temperature of the room between the desired margins during 16 hours. However, it is worth noting that, despite its proper behavior, the particular model is not cost efficient, as all HVAC modules work simultaneously, consuming a total power consumption of 6000 W. In the following sections, we evaluate our proposed methods, demonstrating that they can reduce the electrical cost with respect to the baseline approach.

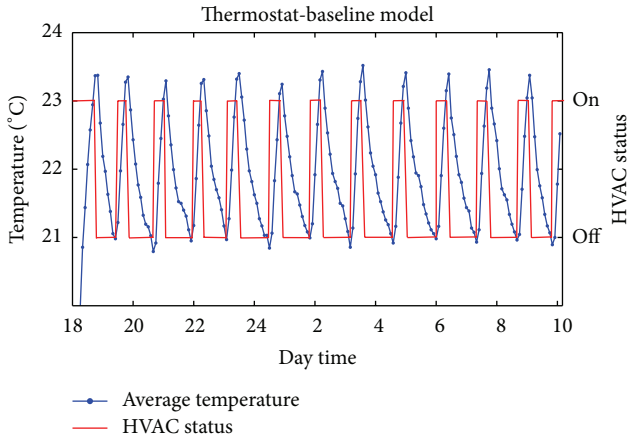


Figure 8: Experimental evaluation of the thermostat model.

5.3. Experimental Evaluation of the DES-CC in (4). Several real experiments have been carried out to assess the performance of the DES-CC algorithm, proposed in (4). In this case, the pricing parameters in (1) are $p_1 = 0.003$ and $p_2 = p_3 = 0$ euros, which are possible values according to [11]. The temperature bounds in (4) have been set to $T_1^{\min} = T_2^{\min} = 21^{\circ}$C and $T_1^{\max} = T_2^{\max} = 23^{\circ}$C, respectively.

Figure 9 plots the variation of both the measured and the estimated temperature (using (8)–(12)) in the room during our experiments. As it can be noticed, the error between the estimated and the real temperature is negligible, something that proves the accuracy of the proposed estimation model. In addition, the DES-CC guarantees the proper operation of the system, as the temperature varies between the desired range of 21°C and 23°C most of the time, with very few exceptions due to prediction errors. In the same figure, it can be also seen that the temperature remains closer to the lower part of the permitted range (i.e., 21°C), since the outcome of the proposed method provides a combination of switched on heaters that minimizes the energy consumption, satisfying a minimum acceptable temperature. Indeed, compared to the temperature variation in the baseline scenario (Figure 8), DES-CC maintains the temperature more stable and in the lower part of the allowable region, intuitively implying lower cost.

Figure 10 depicts the financial operation cost gains that can be achieved by DES-CC compared to the baseline thermostat approach. As it can be observed, the proposed energy scheduler significantly reduces the energy cost, leading to a total save of 7.19 euros/month.

5.4. Experimental Evaluation of the DES-CCR in (13). A set of experiments have been carried out for the evaluation of the DES-CCR in (13). Let us recall that DES-CCR relaxes the temperature constraints by including the constraints as a penalized term in the objective function. As a result, compared to DES-CC, this method is more flexible with respect to real time pricing tariffs. More specifically, DES-CC seeks a combination that minimizes the energy cost with respect to a minimum allowable temperature. Consequently,

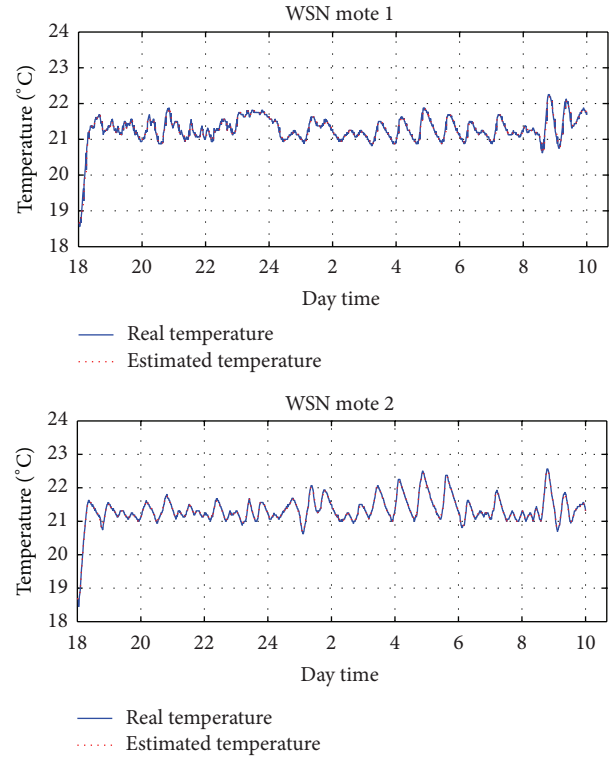


Figure 9: Real and estimated temperature using DES-CC.

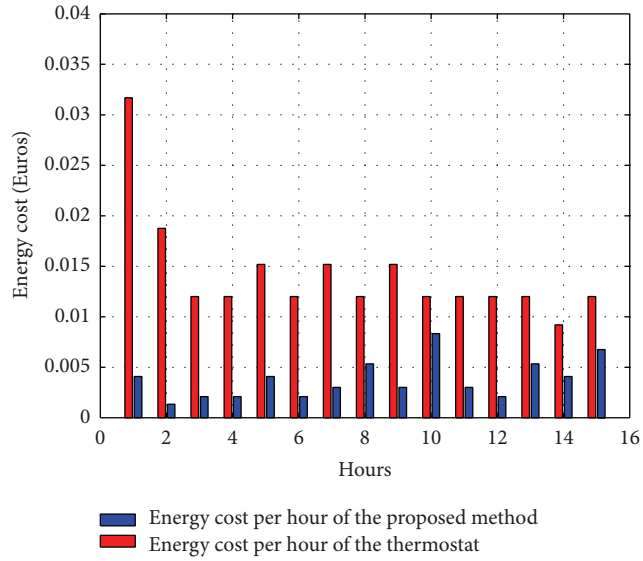


Figure 10: Energy consumption cost comparison between thermostat and DES-CC.

although the energy cost may change during time, the heater combination selected by DES-CC is the same, due to the strict temperature constraint. On the other hand, DES-CCR allows the user to further reduce the energy consumption at the cost of being outside the range of temperature of comfort. In this case, to highlight the flexibility of DES-CCR, we have set a

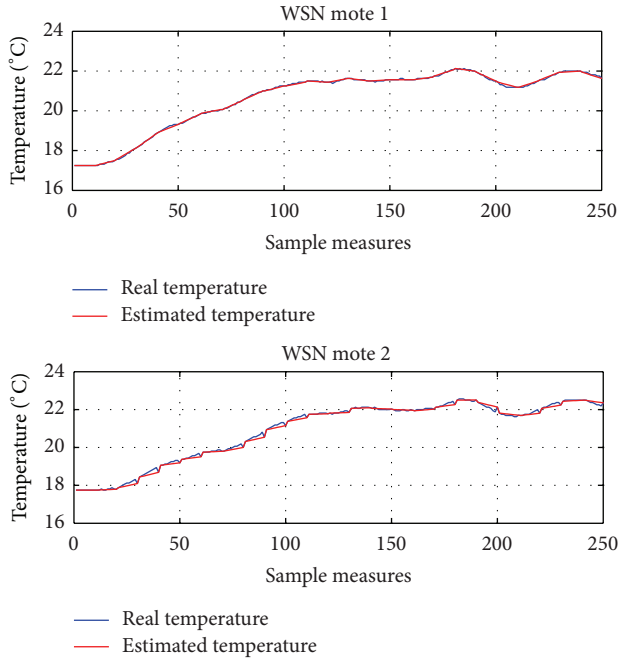


Figure 11: Real and estimated temperature using DES-CCR (θ = 0.2).

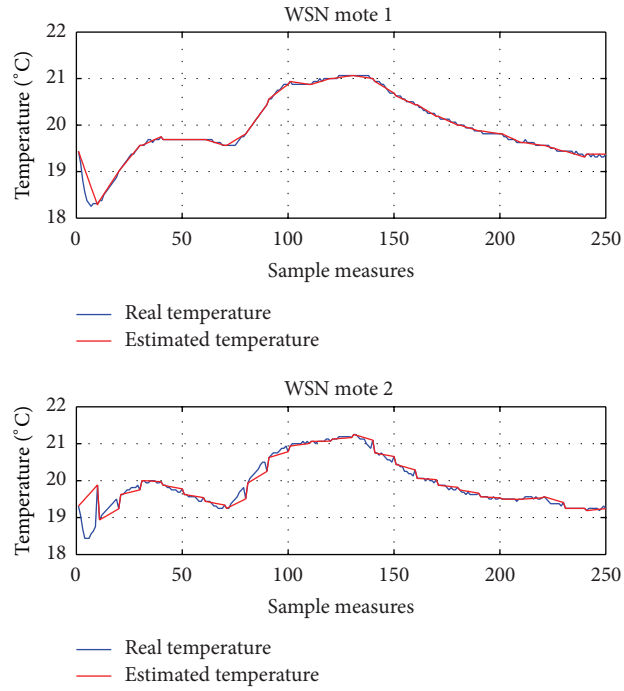


Figure 12: Real and estimated temperature using DES-CCR (θ = 0.5).

periodically variable value of p_1, which alternates between $p_1 = 0.009$ and $p_1 = 0.003$ euros every thirty minutes. Moreover, the desired temperature has been set to $T_{u,i} = 22^{\circ}$C.

Figures 11 and 12 depict the temperature variation in two different cases, where the users give low ($\theta = 0.2$) and high ($\theta = 0.5$) priority, respectively, to reduce of the energy consumption. In particular, in Figure 11 ($\theta = 0.2$), the achieved temperature is very close to the desired $T_{u,i}$. On the other hand, in Figure 12, we assume $\theta = 0.5$, which is a more adapted value to the pricing policy, as it corresponds to a user that permits a relaxation of the difference between the real and the desired temperature to reduce the energy cost. This fact implies higher energy consumption in low cost zones and lower energy consumption in high cost periods, sacrificing though the user's comfort. Therefore, the experiments confirm that the real temperature is close to $T_{u,i}$ when the energy cost is lower (i.e., between samples 60 and 120), while there is a noticeable temperature drop, which corresponds to lower energy consumption.

6. Concluding Remarks

This paper has dealt with the energy consumption management of HVACs, for a given smart pricing tariff and users' comfort constraints. Moreover, the integration within the IoT framework has been studied. To that end, we developed a real testbed consisting of (i) heaters, (ii) sensor nodes that measure the temperature, and (iii) a gateway, which provides connection to the Internet and includes a web application that permits the interaction with the user through Internet. Moreover, the gateway implements the algorithms that control the energy consumption. Regarding the proposed methods, first, we devised an energy scheduler that optimizes the energy cost in a time interval basis, for a given energy price tariff and for a given set of temperature of comfort constraints that are associated with different locations inside a room. Then, we proposed a more flexible energy scheduler, which relaxes the temperature constraints to further reduce the energy consumption. Namely, a new objective function has been considered, which consists of a convex combination of the energy cost and a penalty term that reflects the comfort. This permits to consider both the case where the user is very concerned with the comfort and the case where he allows relaxing the comfort constraint to further reduce the energy consumption. Experimental evaluations have been carried out in an isolated room, validating our proposals and highlighting their potential benefits.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work has been funded by the Energy-to-Smart Grid (E2SG) project http://www.e2sg-project.eu/ within the ENIAC joint undertaking framework with Grant agreement number 296131.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] E. Fleisch, "What is the internet of things? An economic perspective," Tech. Rep., Auto-ID Labs, 2010.

[3] National Intelligence Council, "Disruptive civil technologies—six technologies with potential impacts on us interests out to 2025," Conference Report CR 2008-07, 2008.

[4] M. R. Palattella, N. Accettura, X. Vilajosana et al., "Standardized protocol stack for the internet of (important) things," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1389–1406, 2013.

[5] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, "From today's INTRAnet of things to a future INTERnet of things: a wireless- and mobility-related view," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 44–51, 2010.

[6] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, "A survey on facilities for experimental internet of things research," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 58–67, 2011.

[7] S. Tozlu, M. Senel, W. Mao, and A. Keshavarzian, "Wi-Fi enabled sensors for internet of things: a practical approach," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 134–143, 2012.

[8] N. Bui, A. P. Castellani, P. Casari, and M. Zorzi, "The internet of energy: a web-enabled smart grid system," *IEEE Network*, vol. 26, no. 4, pp. 39–45, 2012.

[9] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—the new and improved power grid: a survey," *IEEE Communications Surveys and Tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[10] G. Lu, D. De, and W. Song, "SmartGridLab: a laboratory-based smart grid testbed," in *Proceedings of the 1st IEEE International Conference on Smart Grid Communications (SmartGridComm '10)*, pp. 143–148, Gaithersburg, Md, USA, October 2010.

[11] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 320–331, 2010.

[12] Z. Zhu, S. Lambotharan, W. H. Chin, and Z. Fan, "Overview of demand management in smart grid and enabling wireless communication technologies," *IEEE Wireless Communications*, vol. 19, no. 3, pp. 48–56, 2012.

[13] H. T. Nguyen, D. Nguyen, and L. B. Le, "Home energy management with generic thermal dynamics and user temperature preference," in *Proceedings of the IEEE International Conference on Smart Grid Communications (SmartGridComm '13)*, pp. 552–557, Vancouver, Canada, October 2013.

[14] Z. Zhu, J. Tang, S. Lambotharan, W. H. Chin, and Z. Fan, "An integer linear programming based optimization for home demand-side management in smart grid," in *Proceedings of the IEEE PES Innovative Smart Grid Technologies (ISGT '12)*, pp. 1–5, Washington, DC, USA, January 2012.

[15] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Optimal residential load control with price prediction in real-time electricity pricing environments," *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 120–133, 2010.

[16] K. M. Tsui and S. C. Chan, "Demand response optimization for smart home scheduling under real-time pricing," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1812–1821, 2012.

[17] G. Wood and M. Newborough, "Dynamic energy-consumption indicators for domestic appliances: environment, behaviour and design," *Energy and Buildings*, vol. 35, no. 8, pp. 821–841, 2003.

[18] M. Avci, M. Erkoc, and S. S. Asfour, "Residential HVAC load control strategy in real-time electricity pricing environment," in *Proceedings of the IEEE Energytech*, pp. 1–6, Cleveland, Ohio, USA, May 2012.

[19] Z. Q. Luo, W. K. Ma, A. C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[20] K. F. Fong, V. I. Hanby, and T. T. Chow, "HVAC system optimization for energy management by evolutionary programming," *Energy and Buildings*, vol. 38, no. 3, pp. 220–231, 2006.

[21] D. L. Ha, F. F. de Lamotte, and Q.-H. Huynh, "Real-time dynamic multilevel optimization for demand-side load management," in *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM '07)*, pp. 945–949, Singapore, December 2007.

[22] S. Noh, J. Yun, and K. Kim, "An efficient building air conditioning system control under real-time pricing," in *Proceedings of the International Conference on Advanced Power System Automation and Protection (APAP '11)*, pp. 1283–1286, Beijing, China, October 2011.

[23] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, Wiley-Interscience, New York, NY, USA, 1988.

[24] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, New York, NY, USA, 1993.

[25] "Contiki: the open source OS for the internet of things," http://www.contiki-os.org/.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 286963, 8 pages
http://dx.doi.org/10.1155/2014/286963

Research Article

Parallelized Dilate Algorithm for Remote Sensing Image

Suli Zhang,[1] Haoran Hu,[2] and Xin Pan[1]

[1] *School of Computer Project & Technology, Changchun Institute of Technology, Changchun 130012, China*
[2] *School of Computer & Information, Anqing Teachers College, Anqing 246011, China*

Correspondence should be addressed to Haoran Hu; haoranhu123@126.com

Received 10 February 2014; Accepted 27 March 2014; Published 11 May 2014

Academic Editors: Z. Zhou and X. Zhu

As an important algorithm, dilate algorithm can give us more connective view of a remote sensing image which has broken lines or objects. However, with the technological progress of satellite sensor, the resolution of remote sensing image has been increasing and its data quantities become very large. This would lead to the decrease of algorithm running speed or cannot obtain a result in limited memory or time. To solve this problem, our research proposed a parallelized dilate algorithm for remote sensing Image based on MPI and MP. Experiments show that our method runs faster than traditional single-process algorithm.

1. Introduction

Land use/cover information has been identified as one of the crucial data components for many aspects of global change studies and environmental applications. The development of remote sensing technology has increasingly facilitated the acquisition of such information [1]. As an important algorithm, dilate algorithm can give us more connective view of a remote sensing image which has broken lines or objects. However, with the technological progress of satellite sensor, the resolution of remote sensing image has been increasing and its data quantities become very big. This would lead to the decrease of algorithm running speed or cannot obtain a result in limited memory or time.

Paralleled program can split a big computing task into subcomputing tasks and make full use of the advantage of multicore and multicomputer to improve the computing speed [2]. To accelerate the process speed of remote sensing images algorithm, many methods had been proposed: parallel k-means or EM cluster method for remote sensing image [3, 4]. Wang utilized loud computing to a rapid processing of remote sensing images [5]. Parallel classification method has been proposed to archive a faster remote sensing images training speed [6, 7]. Parallel program can be further divided into multiprocesses parallel and multithreads parallel. Message passing interface (MPI) is a library specification for message passing, proposed as a standard by a broadly based committee of vendors, implementers, and users [8]; we can realize multiprocesses. Multiprocessing (MP) is the use of two or more central processing units (CPUs) within a single computer system [9].

In this research, we introduce MPICH2 and OpenMP technology and propose a parallelized dilate algorithm for remote sensing image (PDARSI); through PDARSI a big dilate task can be split into a lot of subtasks; each subtask can run on corresponding computer or core. Experiments show that our method runs obviously faster than traditional single-process algorithm.

2. Preliminary Knowledge

2.1. Dilate Algorithms. There are two sets A and B in Z; a complement set of A is as follows:

$$A^C = \{\omega \mid \omega \notin A\}. \quad (1)$$

Based on this formula the difference of A and B represented by $A - B$ can be defined as

$$A - B = \{\omega \mid \omega \in A, \omega \notin B\} = A \cap B^C. \quad (2)$$

The reflection of B represented as $\widehat{B}$ can be defined as

$$\widehat{B} = \{\omega \mid \omega = -b, b \in B\}. \quad (3)$$

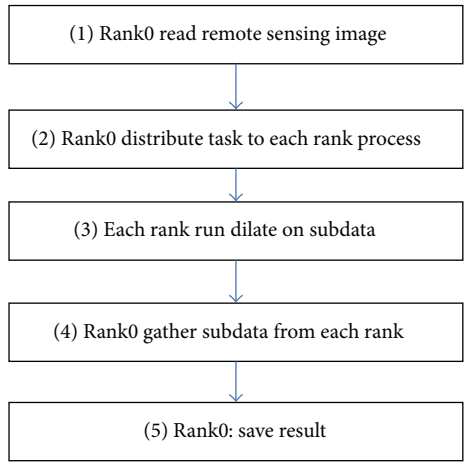


Figure 1: Five stages of algorithms.

The set A move to point $z = (z1, z2)$'s location represented by $(A)_z$ can be defined as

$$(A)_Z = \{c \mid c = a + z, a \in A\}. \quad (4)$$

The binary dilation of A by B, denoted by $A \oplus B$, is defined as the set operation:

$$A \oplus B = \left\{z \mid \left(\widehat{B}\right)_z \cap A \neq \phi\right\}. \quad (5)$$

Here $\widehat{B}$ is the reflection of the structuring element B. In other words, it is the set of pixel locations z, where the reflected structuring element overlaps with foreground pixels in A when translated to z. Note that some people use a definition of dilation in which the structuring element is not reflected [10]. In the general form of gray-scale dilation, the structuring element has a height. The gray-scale dilation of $A(x, y)$ by $B(x, y)$ is defined as

$$(A \oplus B)(x, y) = \max\left\{A\left(x - x', y - y'\right) + B\left(x', y'\right) \mid \left(x', y'\right) \in D_B\right\}, \quad (6)$$

where D_B is the domain of the structuring element B and $A(x, y)$ is assumed to be $-\infty$ outside the domain of the image. To create a structuring element with nonzero height values, use the syntax strel (nhood, height), where height gives the height values and nhood corresponds to the structuring element domain, D_B [11].

2.2. MPI and OpenMP. Message passing interface (MPI) is a standardized and portable message-passing system designed by a group of researchers from academia and industry to function on a wide variety of parallel computers. OpenMP is a portable, scalable model that gives shared-memory parallel programmers a simple and flexible interface for developing parallel applications for platforms ranging from the desktop to the supercomputer. We can use MPI in the cluster computers to realize multiprocess parallelization, and each process adopts OpenMP to realize multithreading parallelization.

3. Parallelized Dilate Algorithm for Remote Sensing Image

The generic process of parallelized dilate algorithm for remote sensing image (PDARSI) is shown in Figure 1.

Firstly, in main function, MPI interface start and initial all the processes by:

MPI::Init(argc, argv); //Initial all the process;

size = MPI::COMM_WORLD.Get_size(); //Get number of processes;

rank = MPI::COMM_WORLD.Get_rank(); //Get current process's rank number.

And then the algorithm is divided into five steps: (1) Rank0 read the entire remote sensing data, and data in accordance with the number of processes is divided into multiple subdata; (2) Rank0 send data, and the data is distributed to each process; (3) each rank processes its own data to obtain the corresponding results; (4) Rank0 collect all the data, and data integration as a result; (5) Rank0 write the result to disk. Finally, in main function, call

MPI::Finalize(); //stop all the process.

All the process was destroyed and the algorithm was ended.

Stage 1. Reading stage: in order to solve the problem of image data read, task assignment, and data transmission, PDARSI adopt a Plines class to store the remote sensing data; Plines class has the following functions: (1) storing remote sensing image in units of row; (2) the storage part of the data; (3) redundant storage of data boundary information; (4) supporting serialization; and (5) supporting the reconstruction of the data; the process of Rank0's reading and splitting can be described in Algorithm 1.

Through this procedure, Rank0 can split all the data into subdata corresponding to each Rand, and the PDARSI proceed to Step 2. In Step 2, Rank0 send all the subdata to Rank0 to Rankn by

```
Input: S (Remote Sensing image), size (the number of processes),
       window (the windows size of dilate algoritm)
Output: A array of Plines objects
Begin
  PLines object pall = read all the data of S;
  n = rows of S;
  rown = floor (n/size);
  PLines[]    array = new Plines[size];
  for i = 0 to size − 1 loop
       array[i].DataStartPosition = i ∗ rown;
    array[i].DataEndPosition = (i + 1) ∗ rown;
    array[i].UpperBuffer = window/2;
    array[i].BottomBuffer = window/2;
    if i == 0 then array[i].UpperBuffer == 0; end if;
       if i == size − 1 then
          array[i].BottomBuffer = 0;
          array[i].BottomBuffer = n − i ∗ rown
       endif
  end loop
  return array;
End
```

Algorithm 1

Table 1: Run speed in multiprocess and multithread.

The number of processes	The number of threads									
	1	2	3	4	5	6	7	8	9	10
1	79.03	41.36	29.16	22.35	18.27	15.38	13.63	11.84	13	15.8
2	57.02	37.27	31.63	28.19	26.1	24.73	23.87	23.24	23.5	23.29
3	39.74	26.11	21.54	19.15	17.82	16.83	16.05	16.42	16.21	16.3
4	37.46	28.08	24.64	22.85	21.78	21.08	20.55	20.81	20.8	20.68
5	30.88	22.7	20.36	18.47	17.58	17.1	17.29	20.42	17.2	17.14
6	31.69	24.73	22.38	21.14	20.42	19.93	19.89	19.98	19.9	19.88
7	27.67	21.49	19.71	20.92	17.71	18.42	20.71	17.75	17.62	19.74
8	28.39	23.56	23.22	20.61	19.93	22.9	19.96	19.69	22.66	19.45
9	28.59	23.7	21.98	18.31	18.24	18.62	20.69	18.02	18.02	18.56
10	26.43	22.35	21.08	20.01	20.25	22.67	19.67	19.55	19.41	22.15

MPI::COMM_WORLD.Send(&linehead,sizeof(Line Head),MPI::BYTE,i,0); //send Plines head

MPI::COMM_WORLD.Send(lines[i].GetSendBuffer Head(),(int)linehead.sendsize,MPI::BYTE,i,0); //send PLines data.

In Step 3, Plines object which own by each rank was further split into Plines array and each object of array run dilate algorithm and obtain result in a thread. In Step 4, Rank0 collect all the results from every rank process and integrate them as a result. In Step 5, Rank0 save the result to a disk. The Plines objects send and receive figure can be seen in Figure 2.

Through PDARSI remote sensing image can dilate parallel in multiprocess and multithread.

4. Experiments

This research chooses Landsat-5 TM image and extracts a band for test image; the image size is 5230 × 4736 and 23.6 M (see Figure 3).

To test the efficiency of PDARSI algorithm, we adopt a HPC cluster which contains Intel i5 2300 computer as head node; it controls two compute nodes which have AMD FX8350 8-core CPU. Each computer of cluster utilizes Fedora 16 64-bit Linux operating system and MPICH2 as MPI management interface and OpenMP as multithread library. In order to test the effectiveness of parallel algorithms, we adopt the number of processes from 1 to 10 and the number of threads from 1 to 10, totally 100 times test. The dilate operator use 15 × 15, in each pixel of image algorithm would compute 15 × 15 = 225 times, Table 1 is test result.

It can be seen from Table 1, when there are one processes and one threads, the algorithm is equivalent to the traditional

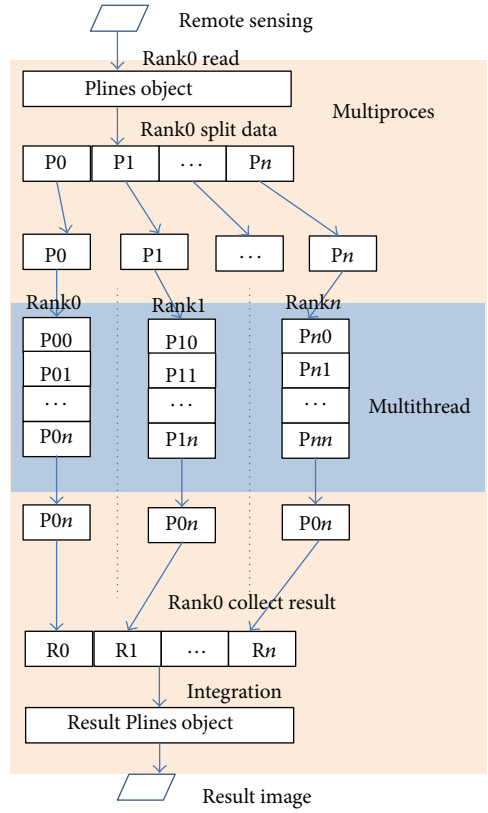


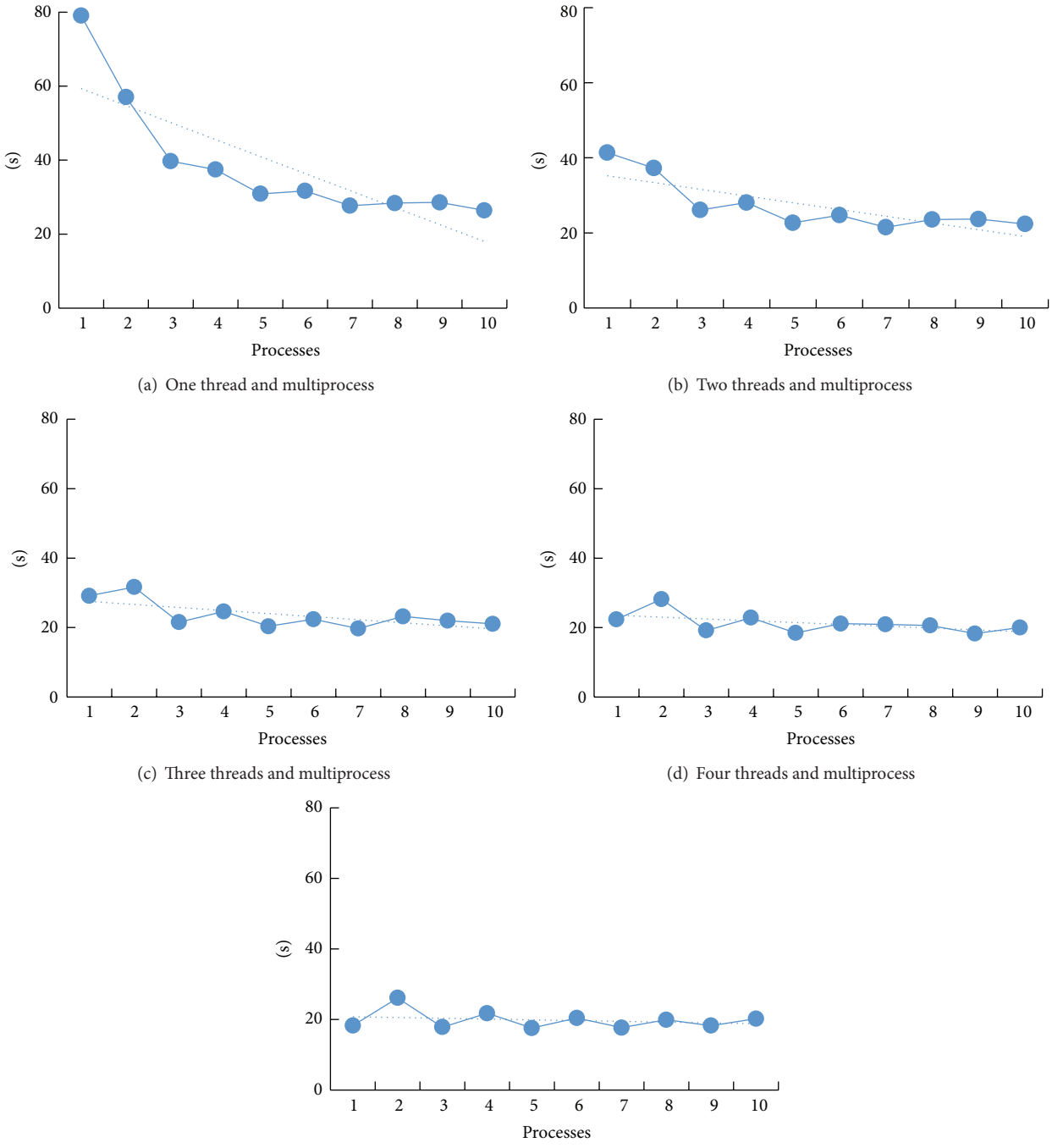
Figure 2: Plines object and its status at different stage.

Figure 3: Test remote sensing image.

single-process algorithm, the running time is slowest 79.03 seconds, with the number of processes or the number of threads increase the running speed become faster. when 10 processes and one thread the algorithm running time is 22.67 seconds, only 0.27 times of the single-process run time. When 10 threads and one process the running time is 15.8 seconds, only 0.2 times of the single-process run time. Both MP and MPI can play an important role in accelerating the program running speed.

Figure 4 is algorithm speed and its relation to the number of processes and 1–5 threads.

As can be seen from Figure 4(a), the elapsed time of algorithm declined along with the increasing number of processes, but the trend became slower when the processes number exceeds 4. Figures 4(b) and 4(c) show that the algorithm's speed increases more slowly when the thread process number is bigger than 1; this means that threads can bring more increase than the processes. From Figures 4(d) and 4(e), the number of threads in the initial stage more than 4, number of processes' increase may actually reduce the operating speed, which is due to the improvement of the process of the speed has less influence than the speed of data



(a) One thread and multiprocess

(b) Two threads and multiprocess

(c) Three threads and multiprocess

(d) Four threads and multiprocess

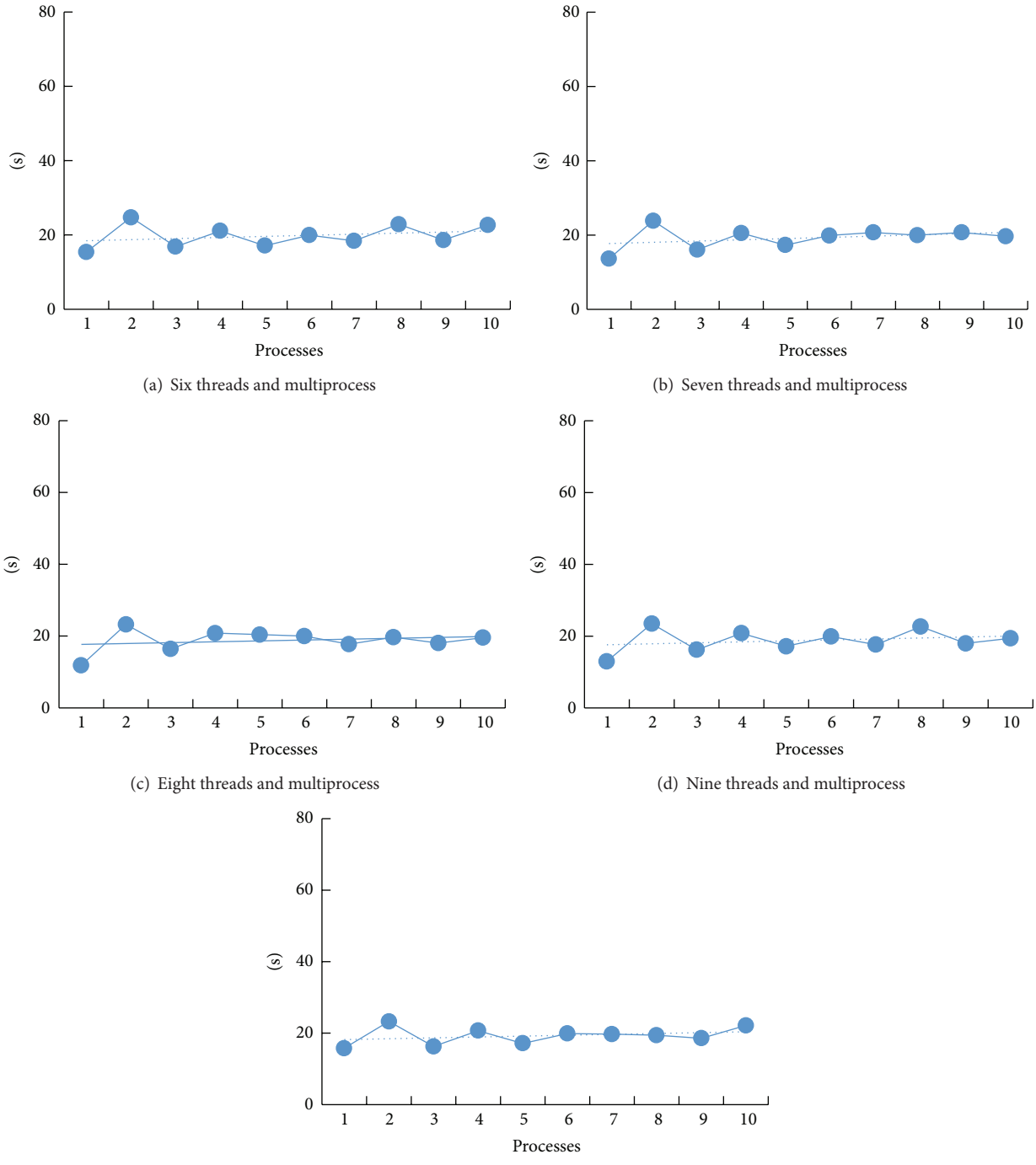
(e) Five threads and multiprocess

Figure 4: Algorithm speed and its relation to the number of processes and 1–5 threads.

transmission between processes. Figure 5 is algorithm speed and its relation to the number of processes and 6–10 threads.

In Figure 5, there are less differences among the five figures, due to HPC cluster computer hardware limitations (the compute nodes containing a total of 16 cores) and the time-consuming communication among processes. The algorithm's speed is not linear with the number of processes and threads, and when the speed limitation is reached, the increase of the number of processes or threads will not increase running speed or even decrease the speed. The algorithm will archive better result when more powerful HPC cluster is utilized.

The relation between threads and processes can be seen from Figure 6. Multithread method does not require data transmission so is can bring more obvious increase in algorithm speed, since algorithms require the transmission of data between two computers in multiprocess stage, when process number is even algorithm need transmitted half of



(a) Six threads and multiprocess

(b) Seven threads and multiprocess

(c) Eight threads and multiprocess

(d) Nine threads and multiprocess

(e) Ten threads and multiprocess

Figure 5: Algorithm speed and its relation to the number of processes and 6–10 threads.

the data from Rank0 node to another node, so the speed is more faster in processed number is odd.

The results of the PDARSI algorithm and the traditional single-process algorithm can be seen in Figure 7.

PDARSI algorithm splits a remote sensing image into subdata, and each subdata has *UpperBuffer* and *BottomBuffer* to ensure a pixel which at subdata border can access neighbor pixels on the other subdata; this mechanism guarantees that the dilation algorithm can obtain right result even the whole calculate task are partitioned into processes. When the dilation algorithm calculation in each process is completed, Rank0 collect the results from processes and integrate all the results into a result image. As can be seen from Figure 7, PDARSI does not change the results of the original algorithm and result images are exactly the same; this proves that our proposed algorithm can accelerate running speed and does not alter the results of the original algorithm.

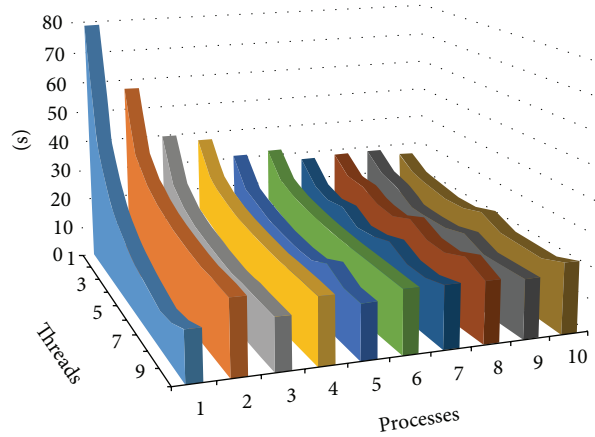
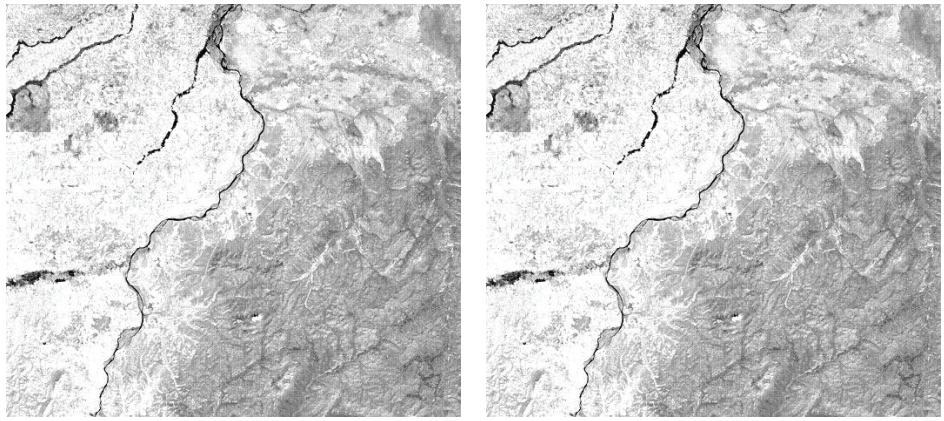


Figure 6: The relation between threads and processes.



(a) Traditional single-process method's result

(b) PDARSI algorithm's result in 10 numbers of processes

Figure 7: Result comparison.

5. Conclusions

This research uses MPICH2 and OpenMP to design a parallelized dilate algorithm; it can take full advantage of HPC cluster computing resources and achieve the purpose of rapid processing of remote sensing image. Through PDARSI a big dilate task can be split into a lot of subtasks; each subtask can run on corresponding computer or core. Experiments show that our method runs obviously faster than traditional single-process algorithm.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the National Natural Science Foundation Youth Fund of China (41101384).

References

[1] O. Yun and J. Ma, "Land cover classification based on tolerant rough set," *International Journal of Remote Sensing*, vol. 27, no. 14, pp. 3041–3047, 2006.

[2] M. Cosnard and D. Trystram, *Parallel Algorithms and Architecture*, International Thomson Computer Press, Boston, Mass, USA, 1995.

[3] J. Li, Y. Qin, and H. Ren, "Research on parallel unsupervised classification performance of remote sensing image based on MPI," *Optik*, vol. 123, pp. 1985–1987, 2012.

[4] A. Sarkar and U. Maulik, "Parallel point symmetry based clustering for gene microarray data," in *Proceedings of the7th International Conference on Advances in Pattern Recognition (ICAPR '09)*, pp. 351–354, February 2009.

[5] P. Wang, J. Wang, Y. Chen, and G. Ni, "Rapid processing of remote sensing images based on cloud computing," *Future Generation Computer Systems*, vol. 29, pp. 1963–1968, 2013.

[6] U. Maulik and A. Sarkar, "Efficient parallel algorithm for pixel classification in remote sensing imagery," *GeoInformatica*, vol. 16, no. 2, pp. 391–407, 2012.

[7] X. Pan and S. Zhang, "Ensemble remote sensing classifier based on rough set theory and genetic algorithm," in *Proceedings of the 18th International Conference on Geoinformatics*, pp. 1–5, June 2010.

[8] P. Peter, *An Introduction to Parallel Programming*, Morgan Kaufmann Publish, Burlingtom, Vt, USA, 2011.

[9] C. Barbara, J. Gabriele, and P. Ruud, *Using OpenMP: Portable Shared Memory Parallel Programming*, The MIT Press, Cambridge, UK, 2007.

[10] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, vol. 1, Addison-Wesley, Reading, Mass, USA, 1992.

[11] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*, Gatesmark Publishing, 2009.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 671964, 9 pages
http://dx.doi.org/10.1155/2014/671964

Research Article

A Color Gamut Description Algorithm for Liquid Crystal Displays in *CIELAB* Space

Bangyong Sun,[1,2] Han Liu,[2] Wenli Li,[1] and Shisheng Zhou[1]

[1] *School of Printing and Packing, Xi'an University of Technology, Xi'an 710048, China*
[2] *School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China*

Correspondence should be addressed to Bangyong Sun; sun_bang_yong@163.com

Received 18 December 2013; Accepted 22 January 2014; Published 24 April 2014

Academic Editors: X. Meng, Z. Zhou, and X. Zhu

Because the accuracy of gamut boundary description is significant for gamut mapping process, a gamut boundary calculating method for *LCD* monitors is proposed in this paper. Within most of the previous gamut boundary calculation algorithms, the gamut boundary is calculated in *CIELAB* space directly, and part of inside-gamut points are mistaken for the boundary points. While, in the new proposed algorithm, the points on the surface of *RGB* cube are selected as the boundary points, and then converted and described in *CIELAB* color space. Thus, in our algorithm, the true gamut boundary points are found and a more accurate gamut boundary is described. In experiment, a Toshiba *LCD* monitor's 3D *CIELAB* gamut for evaluation is firstly described which has regular-shaped outer surface, and then two 2D gamut boundaries (*CIE-*a^*b^* boundary and *CIE-*C^*L^* boundary) are calculated which are often used in gamut mapping process. When our algorithm is compared with several famous gamut calculating algorithms, the gamut volumes are very close, which indicates that our algorithm's accuracy is precise and acceptable.

1. Introduction

Different color devices generally have different color gamuts because of their specific coloring parameters, such as the coloring principle (additive and subtractive), colorant (pigment or dyers), substrates, and light source (D50, D65, or others). Thus, a color image often looks different when it is output by different color devices. For example, a brightly displayed photo on a monitor commonly loses some color information when printed on the paper. For the purpose of matching color image's visual effect between different devices, gamut mapping is often performed in digital imaging processing systems. Gamut mapping can be defined as the color rendering process which rearranges the colors from source device gamut to destination device gamut [1]. Because source and destination gamuts mismatch each other in some regions, there exist some colors which are inside the source gamut but outside of the destination gamut, and they must be clipped or compressed into the destination gamut.

Now there are dozens of gamut mapping algorithms (or GMAs) developed by different agencies[1]. For example, the *LCLIP* algorithm clips out-of-gamut colors onto the gamut boundary and leaves the chroma of in-gamut colors unchanged. The *CUSP* algorithm compresses colors towards a focus point, where the lightness and chroma are changed simultaneously. While for some universal GMAs, such as *CARISMA* or *GCUSP*, the mapping process is very complex but accurate. Because there are too many GMAs developed for the various color images, four specific GMAs are assigned within ICC workflow for convenience, which are absolute colorimetric, relative colorimetric, perceptual, and saturation [2].

Actually, these GMAs are sufficient to cope with all types of color images. However, the color inconsistency still occurs frequently when color images are transmitted from monitors to printers. As a matter of fact, one major reason which results in the gamut mapping errors comes from the gamut boundary calculation process, so it is significant to obtain the accurate gamut boundary before gamut mapping.

Gamut mapping is commonly performed in 2D coordinate, such as *CIE-*C^*L^* or *CIE-*a^*b^* coordinates [3]. The 2D gamut boundary is a cross-section from 3D *CIELAB* gamut,

the errors within both the 3D gamut calculation process and the 2D gamut boundary calculation process will reduce the gamut mapping precision greatly.

Because there are some problems within the present gamut calculation algorithms (analyzed in Section 2), a 2D gamut boundary calculating algorithm is developed for gamut mapping process in this paper. In experiment an *LCD* monitor's gamut boundary is calculated and compared with other algorithms, and the experiment result shows that our algorithm is precise enough. In addition, it should be noted that the proposed algorithm is not only suitable for *RGB* monitors, but also for *CMY* (*CMYK*) printers or other color devices.

2. Different Types of Gamuts Employed in Gamut Mapping Process

In color reproduction systems, color gamut refers to the subset of colors which can be accurately represented in a given circumstance, such as within a given color image or by a certain color device. While the gamut boundary means the outer surface of the 3D gamut, or the outer contour line of the 2D gamut. Color gamut is often described in *CIELAB* space; this is mainly because *CIELAB* color spaces are independent of devices, and it is more perceptually uniform than *CIEXYZ* space.

Generally, there are three types of color gamuts described in *CIELAB* space, such as 3D *CIELAB* gamut, 2D *CIE-a^*b^** gamut, and 2D *CIE-C^*L^** gamut [4]. The first gamut is often used for viewing the overall gamut of color devices or images but rarely used in gamut mapping process directly. The two other gamuts are described in 2D coordinates which are often applied in gamut mapping process, while the main task of this paper is to calculate the 2D gamut's boundary for gamut mapping algorithms.

The *CIE-a^*b^** gamut is a cross-section of 3D gamut along the constant lightness plane, and its boundary is the 3D gamut's outer intersection line with the same lightness. The *CIE-C^*L^** gamut is a cross-section of 3D gamut along the constant hue-angle plane, and the boundary is the intersection line at the specified hue angle. Because hue information is generally more important than lightness, the third gamut boundary is most widely used during gamut mapping.

It is significant to obtain accurate description of the 3D gamut, because the 2D gamuts are calculated from it. Several analytical models are used to predict the 3D gamut with relatively few sample colors, such as the *Kubelka-Munk* function [5], and *Neugebauer* equations [6]. Moreover, Herzog considered the device's 3D gamut as a distorted hexahedron [7] and used a deformation degree function to simulate the actual device gamut. Huang and Zhao used *Zernike* polynomials to determine the boundary [8], and Wang and Xu used a two-step workflow to describe CRT monitor gamut boundaries [9]. On the whole, all these prediction algorithms are based on a specific analytical model and can get accurate gamut with few sample data. However, these methods cannot be applied to all types of color devices, because the analytical expression is highly restricted by the device status such as coloring principle, substrate, inks, and ambient light sources.

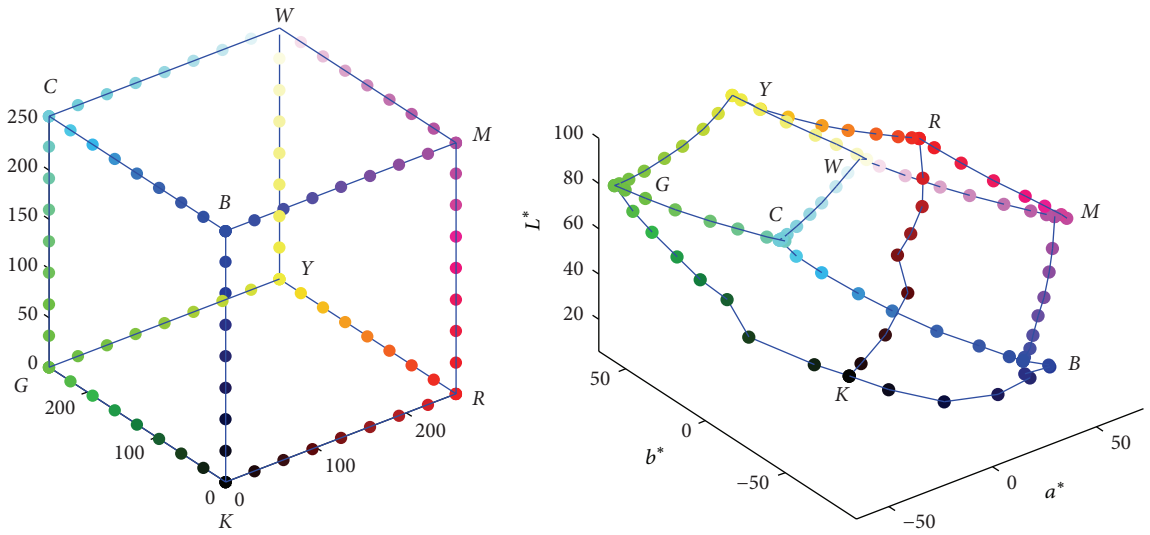
If a sufficient number of sample colors are supplied, the 3D gamut can be described with the empirical algorithms. These algorithms do not focus on analyzing the devices' coloring principle or the relationship between device color space and the corresponding *CIELAB* values; they straightly simulate the 3D gamut with the measured *CIELAB* values. Thus these methods can be used for all types of color devices, even for color images gamut calculation. Within these empirical algorithms, Balasubramanian and Dalal used a modified convex hull method to compute gamut, in a way of adjusting the concave gamut surfaces into convex [10]. Cholewo and Love used alpha-shape to describe the gamut surface for an inkjet printer [11], and the gamut was controlled by an alpha parameter which was interactively selected with a visualization system. Because it is difficult to determine the optimal parameter values, Morovic proposed the segment maxima gamut boundary description (SMGBD) algorithm to compute media and device gamut [4]. However, sometimes there are no boundary points falling into certain segments, it must be created by interpolation method, which pulls errors into the gamut calculations.

In general, the analytical algorithms can be used neither for all types of color devices, nor for the color images, while, for the empirical algorithms, the calculated boundary points are not very precise, because some inside-gamut boundary points may be mistaken for boundary points. In the paper, a new gamut description algorithm is proposed for calculating liquid crystal display (*LCD*) monitors' 3D and 2D gamut. Similar to Herzog's algorithm, the 3D *CIELAB* gamut is deemed as a deformed hexahedron, and the relationship between device and *CIELAB* gamut is analyzed and established, but it should be noted that the calculation process is far less complicated. For the experimented *LCD* monitor, the gamut boundary points are firstly found out in *RGB* space, and these obtained points are the true boundary points which are on the surface of *RGB* cube. And then the *RGB* boundary points are converted into *CIELAB* values, while the 3D or 2D gamuts are described based on these boundary points.

3. Three-Dimensional *CIELAB* Gamut Description

During the gamut description for *LCD* monitors, the device color of *RGB* signals are controlled by users to form the sample data, and the corresponding *CIELAB* values are obtained by measuring the screen when the *RGB* colors are displayed. In general, the calculation of monitor gamut in *CIELAB* space can be divided into three steps, which are *RGB* sample data generating, *CIELAB* data measuring, and 3D or 2D gamut calculation.

3.1. Generating the Sample Data. In order to obtain the monitor's *CIELAB* gamut boundary, the colors around the device gamut boundary should be selected to generate the sample data. Thus, for color monitors, the colorant gamut boundary is actually the outer surface of *RGB* cube, and the



(a) *RGB* cube (b) Corresponding *CIELAB* gamut

Figure 1: Gamut boundary formed by 12 edges of *RGB* cube.

sample colors should be collected from it. With respect to the points on the top of *RGB* cube surfaces, there is at least one or two signals maintaining the maximal or minimal values.

If the R, G, and B values are ranging within 0~255 or normalized to 0~1, a certain amount of sample colors can be obtained by dividing the *RGB* cube surfaces' two colorants into m and n sections as below:

$$O_{\text{sample}} = \left\{u + v \mid u, v \in [R, G, B],\right.$$
$$[u] = \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\right\}, \tag{1}$$
$$\left.[v] = \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}\right\}.$$

When the *RGB* sample signals are sent to the displaying driver, they will be displayed on the monitor, and the spectral transmittance is measured by using a spectrophotometer. Then the *CIEXYZ* Tristimulus can be calculated with the following:

$$X = K \int_{380}^{780} S(\lambda)\, \overline{x}(\lambda)\, \tau(\lambda)\, d\lambda,$$
$$Y = K \int_{380}^{780} S(\lambda)\, \overline{y}(\lambda)\, \tau(\lambda)\, d\lambda,$$
$$Z = K \int_{380}^{780} S(\lambda)\, \overline{z}(\lambda)\, \tau(\lambda)\, d\lambda, \tag{2}$$
$$K = \frac{100}{\int_{380}^{780} S(\lambda)\, \overline{y}(\lambda)\, R(\lambda)\, d\lambda},$$

where $S(\lambda)$ is relative spectral power distribution of the illuminant, $\overline{x}(\lambda)$, $\overline{y}(\lambda)$, and $\overline{z}(\lambda)$ are the color-matching functions for *CIE* 2° standard observer (1931), and $\tau(\lambda)$ is the spectral transmittance of color patch, while the *CIELAB* values can be converted from the *CIEXYZ* values using the following:

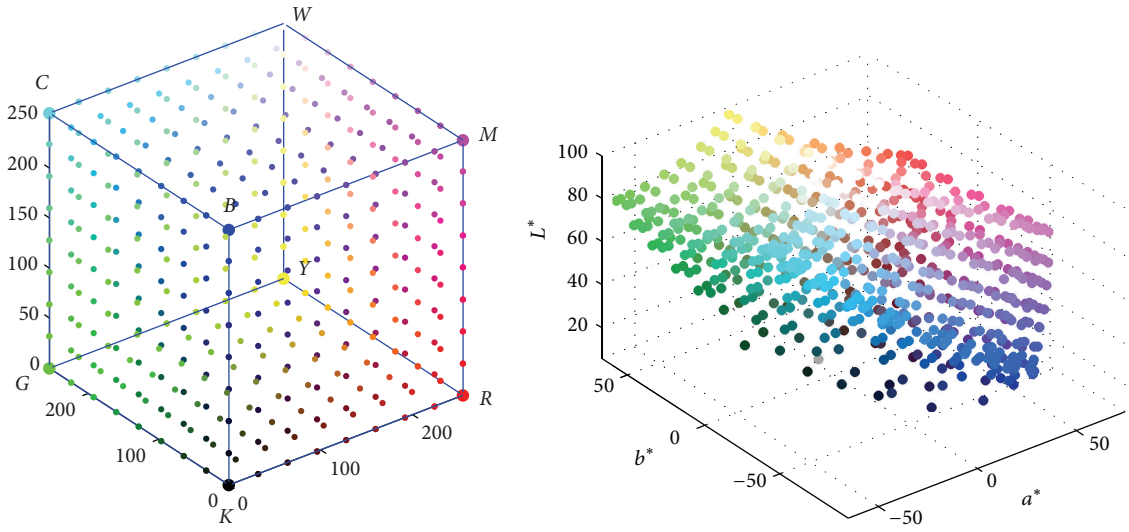
$$L^* = 116 f\left(\frac{Y}{Y_n}\right) - 16,$$
$$a^* = 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right], \tag{3}$$
$$b^* = 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right],$$

where X_n, Y_n, and Z_n are the *CIEXYZ* tristimulus values of the reference white point, and the function f is defined as follows:

$$f(t) = \begin{cases} t^{1/3} & \text{if } t > \left(\frac{29}{6}\right)^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 t + \frac{4}{29} & \text{otherwise.} \end{cases} \tag{4}$$

3.2. Calculation of 3D CIELAB Gamut. When a certain amount of sample colors are obtained, there will be many scattered points distributed in *RGB* or *CIELAB* space. The outer surface of *RGB* cube is the device gamut boundary with regular shape, and it will be severely deformed when converted into *CIELAB* space. For example, if the monitor's *RGB* cube is uniformly sampled by setting $m = n = 9$, in other words every colorant evenly changes in the range of [0 32 64 96 128 160 192 224 255], there will be 9 scattered points on each of the *RGB* cube edges.

The deformation between *RGB* cube and 3D *CIELAB* gamut can be analyzed by comparing the distribution of those 12 edges in these two color spaces. Figure 1(a) is the device *RGB* gamut boundary formed by its 12 edges, and its corresponding *CIELAB* gamut boundary points is shown



(a) Scattered *RGB* boundary

(b) Scattered *CIELAB* boundary

Figure 2: Scattered samples in colorant and colorimetric space.

in Figure 1(b). It is obvious that the colorimetric *CIELAB* gamut looks like a deformed hexahedron in [7], whereas the deformation degree varies in different locations or for different devices.

To get the complete 3D gamut description, the six faces of the *RGB* cube should be converted into *CIELAB* space. Among the sampling points above, as only a part of them located on the outer surface of gamut, the *RGB* boundary points should be firstly extracted from the six *RGB* cube faces, and they are depicted in Figure 2. It can be seen that if all these 3D scattered points are smoothly connected, the complete and closed 3D gamut can be obtained.

With those obtained boundary points in *CIELAB* space, *Delaunay* triangulation technique can be used to link them, and the final completed 3D gamut boundary is composed of a quantity of triangular facets. Since the gamut surface is partitioned into planar triangles, the volume can be calculated when an interior point is assigned. For example, when the center point O (50, 0, 0) in *CIELAB* space is selected, and all the vertices within each surface triangle are connected with the center point, then the *CIELAB* gamut is divided into many tetrahedrons. The overall gamut volume is the sum of the tetrahedrons' volumes, while each tetrahedron's volume can be calculated as below:

$$V = \frac{1}{6}\begin{vmatrix} 1 & 1 & 1 & 1 \\ 50 & L_1^* & L_2^* & L_3^* \\ 0 & a_1^* & a_2^* & a_3^* \\ 0 & b_1^* & b_2^* & b_3^* \end{vmatrix}, \quad (5)$$

where $L_1^*a_1^*b_1^*$, $L_2^*a_2^*b_2^*$, and $L_3^*a_3^*b_3^*$ are three surface vertices of the tetrahedron.

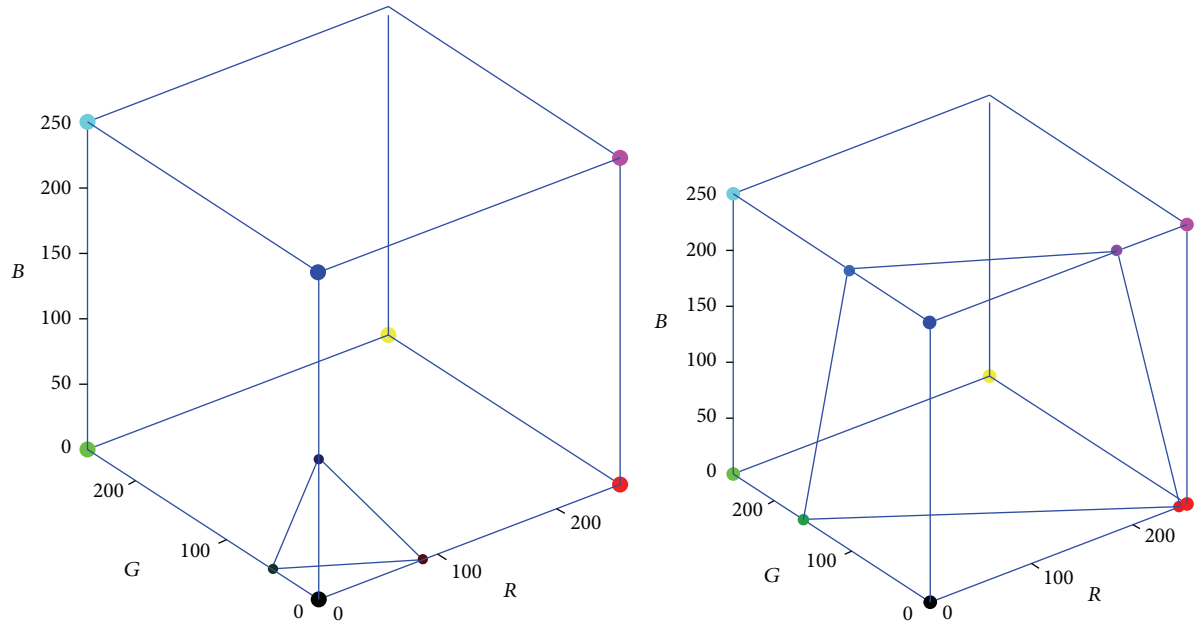
4. Two-Dimensional Gamut Boundary for Gamut Mapping Algorithms

Because gamut mapping is commonly carried out in 2D gamut boundary, such as *CIE-*a^*b^* or *CIE-*C^*L^* gamut boundary, the 2D gamut boundary should be accurately determined. The 2D gamut is a cross- section of 3D *CIELAB* gamut; thus the 2D gamut boundaries can be determined by slicing the 3D gamut along constant-lightness plane or constant-hue-angle plane.

However, because the 3D boundary points do not distribute uniformly, when the scattered 2D boundary points are interpolated from the nearby 3D boundary points, there will be lots of errors caused. We propose a new method to calculate the 2D gamut boundary, in which the 2D boundary points are firstly found within the *RGB* cube, and then their *CIELAB* values are converted which are used to describe the gamut boundary in *CIE-*a^*b^* or *CIE-*C^*L^* coordinates.

4.1. CIE-a^*b^* *Gamut Boundary Calculation.* The *RGB* cube has eight vertices, twelve edges, and six faces. For color monitors, the lightness (*CIE-*L^* values) of the vertices in the *RGB* cube varies greatly. Take the widely used ICC profile of *sRGB,* for example; the *RGB* cube's eight vertices' *CIELAB* values are listed in Table 1, and it can be seen that the vertices' lightness values are arranged in a sequence of below: $W > Y > C > G > M > R > B > K$. Actually most of the *RGB* monitors accord with this discipline.

In addition, as the lightness changes continuously on the faces of *RGB* cube, when a specific lightness value is given, a constant-lightness plane will intersect with some of the *RGB* faces, while the intersection line corresponds to the *CIE-*a^*b^* gamut boundary line. To calculate the constant-lightness gamut boundary, the scattered boundary points



(a) $L^* = 20$ (b) $L^* = 60$

Figure 3: The scattered constant-lightness gamut boundary on *RGB* cube's edges.

Table 1: *RGB* and *CIELAB* values of the vertices for *sRGB.ICC*.

Color	*RGB*	L^*	a^*	b^*
K	(0, 0, 0)	0	0	0
B	(0, 0, 255)	30	67	−128
R	(255, 0, 0)	55	79	65
M	(255, 0, 255)	60	91	−79
G	(0, 255, 0)	88	−81	71
C	(0, 255, 255)	91	−53	−34
Y	(255, 255, 0)	98	−18	83
W	(255, 255, 255)	100	−2	−19

should be firstly found, and the process can be described in the following steps.

(1) Compare the given lightness l with twelve edges' lightness, and find out the boundary points on the edges equal to the given lightness value. For example, if $L_C \le l \le L_W$, which means that the lightness value l is within the range of edge *CW*, then it can be concluded that an intersection point with lightness l must exist in the edge *CW*.

The intersection point can be found by comparing the edge's sample colors. If two closet sample colors are expressed as x and y, as their *RGB* and *CIELAB* values are known, the intersection point p's *R*values can be interpolated as below (the same for *G* and *B* values):

$$R_p = R_x + \frac{L_p^* - L_x^*}{L_y^* - L_x^*}\left(R_y - R_x\right). \tag{6}$$

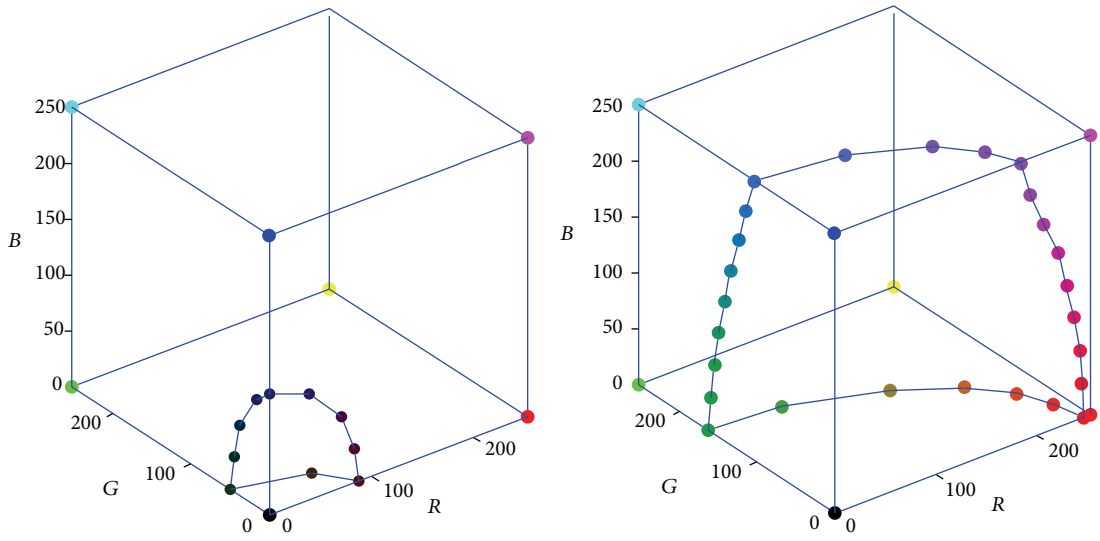
(2) For the constant-lightness lines on the *RGB* cube surface, there are generally three or four edges which contain the intersection points among all the twelve edges. As shown in Figure 3, when the edge intersection points are connected, the constant-lightness lines may form a triangle (*CIE*-$L^* = 20$) or quadrangle (*CIE*-$L^* = 60$).

(3) Because the constant-lightness gamut boundary is a closed and smooth curve, several boundary points are calculated from the corresponding cube faces, for the purpose of obtaining the full description of the gamut boundary. The face intersection points can also be found by comparing the sample colors of the *RGB* cube face, which is similar to the process of calculating edge intersection points. Thus, when all the face boundary points are connected, the two constant-lightness gamut boundaries of Figure 3 are described with more details in Figure 4.

(4) After the scattered gamut boundary points in *RGB* space are calculated, the *CIE*-a^*b^* gamut boundary can be obtained by converting the boundary points from *RGB* space to *CIELAB* space. There are actually several color conversion models, such as 3D interpolation [12–14], polynomial regression [15, 16], and neural network [17, 18]. In the paper, the polynomial regression method is used because of the model's precision and the quantity of sample colors, and the utilized polynomials are expressed as below:

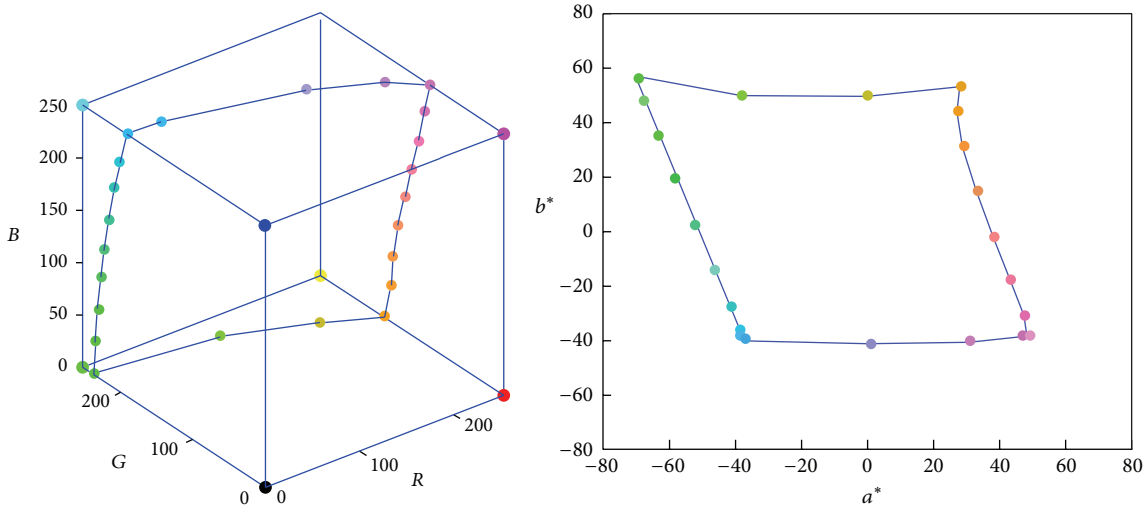
$$\begin{aligned} P(R,G,B) = {} & c_0 + c_1R + c_2G + c_3B + c_4RG + c_5GB \\ & + c_6RB + c_7R^2 + c_8G^2 + c_9B^2 + c_{10}RGB \\ & + c_{11}R^2G + c_{12}R^2B + c_{13}G^2R + c_{14}G^2B + c_{15}B^2R \\ & + c_{16}B^2G + c_{17}R^3 + c_{18}G^3 + c_{19}B^3, \end{aligned} \tag{7}$$

where c_i $(i = 0, 1, \ldots, 19)$ are the coefficients which can be determined by the least square method [19, 20]. Take one



(a) $L^* = 20$ (b) $L^* = 60$

Figure 4: The scattered constant-lightness gamut boundary on *RGB* cube's faces.



(a) 2D gamut on *RGB* cube ($L^* = 80$) (b) The gamut on *CIE*-a^*b^* plane ($L^* = 80$)

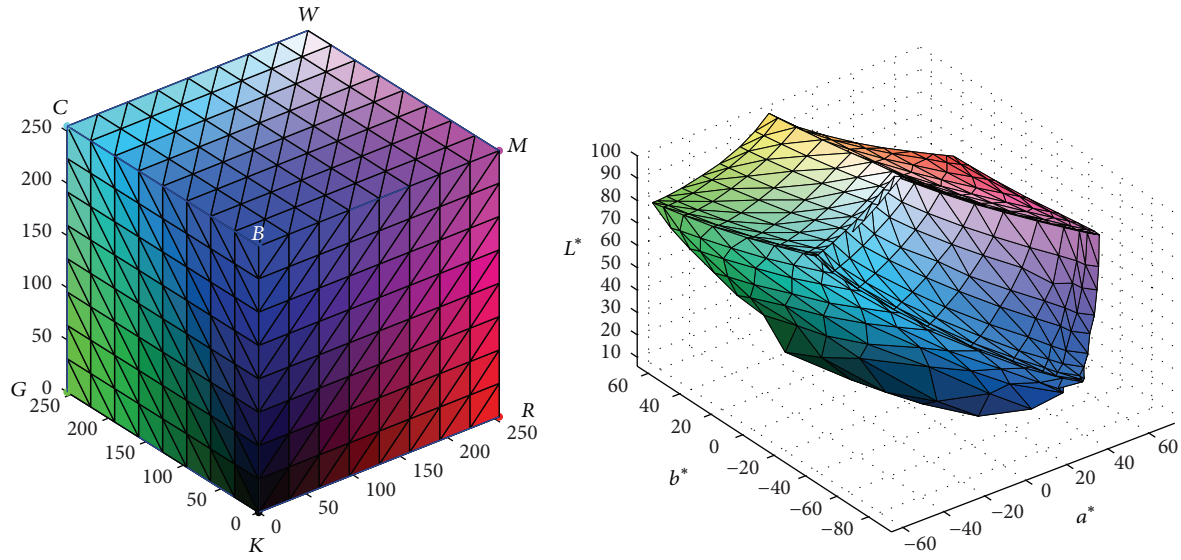
Figure 5: The constant-lightness gamut on *RGB* cube and *CIELAB* space.

constant-lightness boundary, for example; it is depicted in *RGB* and *CIELAB* spaces respectively, as shown in Figure 5.

4.2. CIE-C^*L^* Gamut Boundary Calculation. The *CIE*-C^*L^* gamut boundary can be seen as a cross-section of the constant-hue-angle plane with 3D *CIELAB* gamut, and it is widely used in gamut mapping process. As most of the 3D gamut is calculated by the scattered sample data with interpolation method, it is hard to get the line boundary point directly using the continuous 3D gamut. Similar to the calculation of constant-lightness gamut, the constant-hue-angle boundary points are firstly calculated in *RGB* cube and then converted into *CIELAB* color space and described in the *CIE*-C^*L^* coordinate, where the *CIE*-C^* is calculated from *CIE*-a^* and *CIE*-b^* values:

$$C^* = \sqrt{b^{*2} + b^{*2}}. \tag{8}$$

During description of the 2D *CIE*-C^*L^* gamut, the scattered boundary points are firstly calculated and then connected with Bezier curves or straight lines. Although the boundary looks smoother when Bezier curves are used, the computing efficiency is inferior to the straight lines, which is very important for the gamut mapping process. Thus, the *CIE*-C^*L^* gamut boundary is connected using straight lines in this paper.



(a) 3D gamut in *RGB* space

(b) 3D gamut in *CIELAB* space

Figure 6: The continual 3D gamut in *CIELAB* of Toshiba monitor.

5. Experiments

In experiment, the gamut boundaries of a Toshiba *LCD* monitor are calculated. Firstly, the *RGB* device color space is uniformly divided, and the red, green, and blue signals all range within [0 32 64 96 128 160 192 224 255]; thus there are $9^3 = 729$ sample colors totally. And then the sample *RGB* colors are displayed on the *LCD* monitor; the corresponding *CIELAB* values are measured by using the spectrophotometer *X-Rite DTP*94. At last, the monitor's 3D *CIELAB* gamut and two other 2D gamut boundaries are described using the algorithm proposed in the paper.

As shown in Figure 6, the 3D gamut is described in *CIELAB* color space, and it looks like a deformed hexahedron as Herzog proposed. Compared with the initial *RGB* cube, the *CIELAB* gamut also has eight vertices and six continual curved faces, although most of sample colors' relative positions have changed. From the described *CIELAB* gamut, the overall displaying capacity of the monitor can be evaluated, which will help to select the right gamut mapping algorithms.

Bakke et al. proposed a method for evaluating different gamut boundary calculating algorithms [21], and the two testing algorithms are compared by using a parameter of gamut volume mismatching rate, which is expressed in the following:

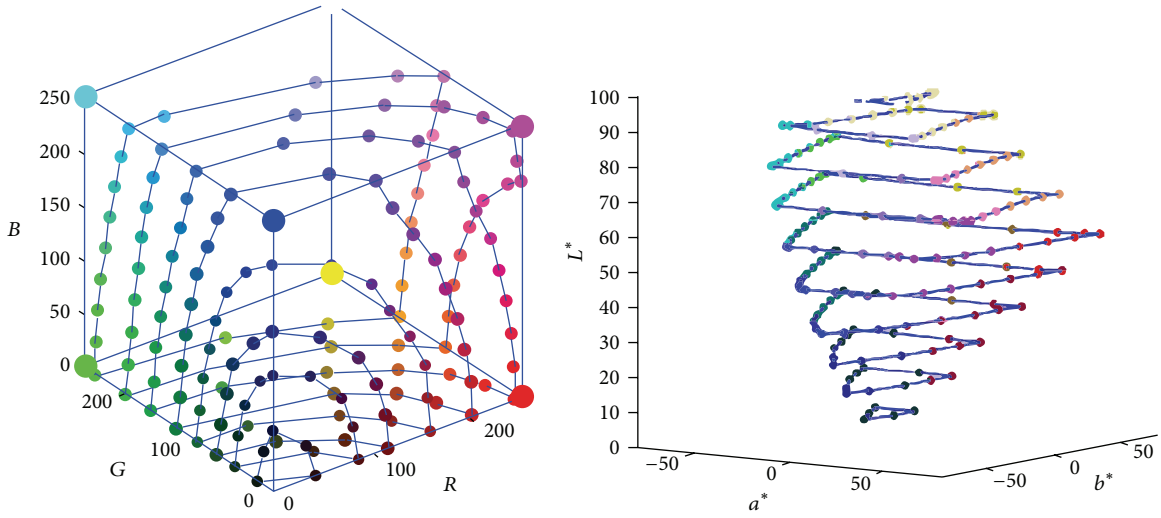
$$r_i = \frac{V\left(G_i/G_{\text{ref}}\right) + V\left(G_{\text{ref}}/G_i\right)}{V\left(G_{\text{ref}}\right)}, \tag{9}$$

where r_i represents the gamut volume mismatching rate, $V(G_{\text{ref}})$ is the volume of reference gamut, and $V(G_i/G_{\text{ref}})$ is the volume of colors which are inside of gamut G_i but outside of gamut G_{ref}. In experiment, the gamuts determined by our algorithm are selected as reference gamut, and three famous gamut description algorithms, SMGBD, convex hull, and alpha shape, are employed as a contrast. Consequently, the mismatching rates are 3.2%, 2.4%, and 3.7%, respectively, which indicate that our algorithm's accuracy is very close to the three successful algorithms.

Additionally, two kinds of 2D gamut mapping boundaries, *CIE-*a^*b^* and *CIE-*C^*L^* boundaries, are calculated using the method described in Sections 4.1 and 4.2. For the testing monitor, several constant-lightness lines with the lightness from 10 to 95 are found in *RGB* cube and then described in *CIELAB* space, as shown in Figure 7. Similarly, two *CIE-*C^*L^* boundaries are listed in Figure 8, and the hue angles are 30, 210, and 330 respectively.

6. Conclusions

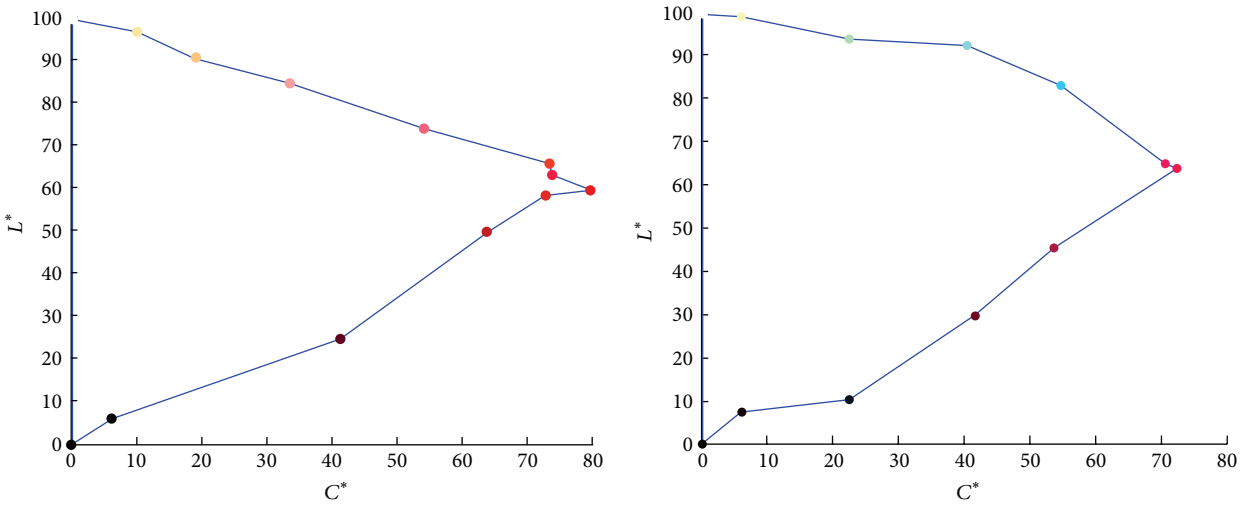
The gamut of monitor is generally bigger than printers; thus when a displayed image is printed, gamut mapping should be performed in advance. Gamut mapping is the technique of replacing nonprintable colors by printable ones, conserving the appearance of an image. There are two major reasons which influence the gamut mapping precision, selection of gamut mapping algorithms and the calculation of gamut boundaries. Because many successful GMAs are developed for various images and color devices, it is very significant to improve the accuracy of gamut boundaries. In the paper, a new gamut boundary algorithm for color devices is proposed, and an *LCD* monitor is tested with having the 3D and 2D gamut boundaries described. In experiment, the described 3D *CIELAB* gamut boundary has regular-shaped outer surface, and the 2D *CIE-*a^*b^* and *CIE-*C^*L^* boundaries both have smooth boundary lines. For the purpose of precision evaluation, our algorithm is compared with other famous gamut description algorithms, and the result shows that the gamut volume difference is very little, which indicates that this algorithm is acceptable. Besides, it should be noted that, although the color monitor's gamut



(a) Different constant-lightness gamut boundaries on *RGB* cube

(b) Corresponding gamut boundaries in *CIELAB* space

FIGURE 7: The 2D gamut boundary of different lightness.



(a) $h = 30$

(b) $h = 210$

FIGURE 8: The 2D gamut with different hue angle.

is described in experiment, the proposed gamut boundary calculating algorithm can also be applied to other color devices, such as cameras, scanners, or printers.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by Research Foundation of Department of Education of Shaanxi Province (no. 11JK0541), Doctor Foundation of Xi'an university of Technology (104-211302), "13115" Creative Foundation of Science and Technology, Shaanxi Province of China. This research was funded by a Grant (no. 104-211302) from the Research Doctor Foundation of Xi'an University of Technology.

References

[1] J. Morovic, *To develop a universal gamut mapping algorithm [Ph.D. thesis]*, University of Derby, Derby, UK, 1998.

[2] P. Green, J. Holm, and W. Li, "Recent developments in ICC color management," *Color Research and Application*, vol. 33, no. 6, pp. 444–448, 2008.

[3] T.-W. Huang and M. Ou-Yang, "Gamut boundary description for one dependent primary color," *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, vol. 26, no. 10, pp. 2163–2171, 2009.

[4] J. Morovič and M. R. Luo, "Calculating medium and image gamut boundaries for gamut mapping," *Color Research and Application*, vol. 25, no. 6, pp. 394–401, 2000.

[5] P. G. Engeldrum, "Computing color gamuts of ink-jet printing systems," *SID Proceedings*, vol. 27, no. 1, pp. 25–30, 1985.

[6] M. Mahy, "Calculation of color gamuts based on the Neugebauer model," *Color Research and Application*, vol. 22, no. 6, pp. 365–374, 1997.

[7] P. G. Herzog, "Analytical color gamut representations," *Journal of Imaging Science and Technology*, vol. 40, no. 6, pp. 516–521, 1996.

[8] Q. Huang and D. Zhao, "Gamut mapping based on gamut boundaries expressed with Zernike polynomials," *Optical Technique*, vol. 29, no. 2, pp. 168–171.

[9] Y. Wang and H. Xu, "Colorimetric characterization of liquid crystal display using an improved two-stage model," *Chinese Optics Letters*, vol. 4, no. 7, pp. 432–434, 2006.

[10] R. Balasubramanian and E. Dalal, "A Method for quantifying the color gamut of an output device," in *Color Imaging: Device-Independent Color, Color Hard Copy, and Graphic Arts*, vol. 3018 of *Proceedings of SPIE*, pp. 110–116, 1997.

[11] T. J. Cholewo and S. Love, "Gamut boundary determination using alpha-shapes," in *Proceedings of the 7th Color Imaging Conference*, pp. 200–204, Scottsdale, Ariz, USA, November 1999.

[12] P. Ibrahim and A. Yucel, "Multiscale gradients-based color filter array interpolation," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 157–165, 2013.

[13] S. Srivastava, E. J. Delp, T. H. Ha, and J. P. Allebach, "Color management using optimal three-dimensional look-up tables," *Journal of Imaging Science and Technology*, vol. 54, no. 3, pp. 30–40, 2010.

[14] R. Henry Kang, *Color Technology for Electronic Imaging Devices*, SPIE Press, Washington, DC, USA, 1997.

[15] Z. Li, H. Zhang, and H. Fu, "Red long-lasting phosphorescence based on color conversion process," *Optical Materials*, vol. 35, no. 3, pp. 451–455, 2013.

[16] S. Bianco, F. Gasparini, R. Schettini, and L. Vanneschi, "Polynomial modeling and optimization for colorimetric characterization of scanners," *Journal of Electronic Imaging*, vol. 17, no. 4, Article ID 043002, pp. 43–52, 2008.

[17] I. Ioannis, G. Alexander, and G. Barry, "Deriving ocean color products using neural networks," *Remote Sensing of Environment*, vol. 134, pp. 78–99, 2013.

[18] X. Li, "Scanner color management model based on improved back-propagation neural network," *Chinese Optics Letters*, vol. 6, no. 3, pp. 231–234, 2008.

[19] S. Kim and M. Kojima, "Solving polynomial least squares problems via semidefinite programming relaxations," *Journal of Global Optimization*, vol. 46, no. 1, pp. 1–23, 2010.

[20] F. Vogt, F. Gritti, and G. Guiochon, "Polynomial multivariate least-squares regression for modeling nonlinear data applied to in-depth characterization of chromatographic resolution," *Journal of Chemometrics*, vol. 25, no. 11, pp. 575–585, 2011.

[21] A. M. Bakke, I. Farup, and J. Y. Hardeberg, "Evaluation of algorithms for the determination of color gamut boundaries," *Journal of Imaging Science and Technology*, vol. 54, no. 5, Article ID 050502, pp. 50–62, 2010.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 129483, 10 pages
http://dx.doi.org/10.1155/2014/129483

Research Article

A Multistrategy Optimization Improved Artificial Bee Colony Algorithm

Wen Liu[1,2]

[1] *The School of Computer Science and Technology, Dalian University of Technology, Dalian, China*
[2] *Department of Electrical Engineering, Xinjiang Institute of Engineering, Tianjin Road, No. 176, Urumqi 830011, China*

Correspondence should be addressed to Wen Liu; liuwen_lw@126.com

Received 24 January 2014; Accepted 27 February 2014; Published 3 April 2014

Academic Editors: X. Meng, Z. Zhou, and X. Zhu

Being prone to the shortcomings of premature and slow convergence rate of artificial bee colony algorithm, an improved algorithm was proposed. Chaotic reverse learning strategies were used to initialize swarm in order to improve the global search ability of the algorithm and keep the diversity of the algorithm; the similarity degree of individuals of the population was used to characterize the diversity of population; population diversity measure was set as an indicator to dynamically and adaptively adjust the nectar position; the premature and local convergence were avoided effectively; dual population search mechanism was introduced to the search stage of algorithm; the parallel search of dual population considerably improved the convergence rate. Through simulation experiments of 10 standard testing functions and compared with other algorithms, the results showed that the improved algorithm had faster convergence rate and the capacity of jumping out of local optimum faster.

1. Introduction

The artificial bee colony algorithm is a new heuristic optimization algorithm proposed in recent years by Karaboga [1]. References [2, 3] pointed out that by comparing the performance of optimization of differential evolution algorithm [4] and the particle swarm algorithm [5], ABC algorithm obtained more favorable test results and is one of the most outstanding function optimization methods, which has become a hot topic at the forefront of domestic and foreign heuristic algorithm researches.

However, similar to other intelligent algorithms, the standard ABC algorithm also had disadvantages of easily prematurely falling into local optima and slow convergence rate in later stage. In this regard, a number of scholars made corresponding improvement. Reference [6] proposed an artificial bee colony algorithm solving the problem of minimum spanning tree; [7] proposed an improved algorithm where combining particle swarm algorithm, to some extent, accelerated the local convergence rate of algorithm; and [8] proposed an improved algorithm combining differential evolution algorithm and artificial bee colony algorithm, which effectively improved the searching accuracy and the convergence rate of the algorithm to some extent. However, improved algorithms above did not effectively improve the convergence rate and avoid premature convergence problem at the same time.

In order to overcome premature to improve the convergence speed and optimal accuracy, this paper proposes a new improved artificial bee colony algorithm. First, a chaos reverse learning strategy was proposed and introduced into the initialization phase of artificial bee colony algorithm, making the initial population uniformly distributed in the search space, in order to improve the quality of population solution, thus speeding up the global convergence rate of the algorithm. Secondly, according to the idea of separate optimization and survival of the fittest, two populations was filter optimized by using dual population structure, thus the optimization process was accelerated in the case of maintaining the population diversity. The introduction of the dynamic adaptive idea to the algorithm and the improvement of nectar update formula based on the comparison of population diversity measurement values, the problem of algorithm falling into a local optimum was solved. The simulation results of

optimization of 10 standard test functions that were widely used showed that, comparing with the existing two artificial bee colony algorithms, the proposed algorithm had better optimization accuracy, convergence rate, and robustness.

2. Algorithm Description

2.1. Artificial Bee Colony Algorithm. Artificial bee colony algorithm uses simulating the mechanism of bees collecting nectar to achieve the optimization processing function. In artificial bee colony algorithm, the bees are divided into three categories, that is, employed bees, onlookers, and scouts [9]. The main task of employed bees and onlookers is to search and mine nectar, and scouts are used to search and compare nectar in order to avoid few nectar species. The location and quantity of the nectar are the solution of function optimization problem and corresponding function value. The process of searching optimal nectar is as follows: employed bees find nectar and memorize and search for new nectar in the vicinity of each nectar; at the same time employed bees release information that is proportional to the mass of marked nectar to attract onlookers. Onlookers select the appropriate marked nectar under some mechanism and search for new nectar source in the vicinity and compare with selected nectar. Select excellent quality nectar as final marked nectar, looking for the best nectar in repeated cycles. If during the process of collecting nectar, after several searches, nectar is unchanged, then corresponding employed bees are changed into scouts and they randomly search for new nectar [9–11]. The function optimization problem can be expressed as follows:

$$\min \text{ fit} = \text{fit}(\mathbf{x}), \quad \mathbf{x} = (x_1, x_2, \ldots, x_n) \in S, \ S = [x_{i,L}, x_{i,H}], \tag{1}$$

where fit represents the objective function, $\mathbf{x}$ is an n-dimensional variable, and $[x_{i,L}, x_{i,H}]$ is the corresponding upper and lower bounds of the ith dimensional variable. Set the number of nectar, employed bees, and onlookers to be N in the ABC algorithm. The specific steps of ABC algorithm are as follows.

Step 1. $2N$ nectar positions are generated randomly by the following formula:

$$V_{ij} = x_{j,L} + \text{rand} \times \left(x_{j,H} - x_{j,L}\right), \tag{2}$$

where V_{ij} is the corresponding search position of ith bee in jth dimension; $x_{j,L}$, $x_{j,H}$ are upper and lower bounds of the jth dimensional variables; select N positions with low fitness values as the position of nectar.

Step 2. Employed bees search and update nectar in the vicinity of nectar according to the following:

$$V_{ij} = x_{ij} + r_{ij} \times \left(x_{ij} - x_{kj}\right), \tag{3}$$

where V_{ij} is the position of new nectar, x_{ij} is the jth dimensional position of nectar i, x_{kj} is jth dimensional position of randomly selected nectar k, and $k \neq i$, r_{ij} is a random number of $[-1, 1]$.

Step 3. Comparing the pros and cons of before and after nectar, replace the previous nectar, if after searching nectar is superior to previous nectar.

Step 4. According to the way of roulette and nectar information released by employed bees, onlookers select nectar, the selection probability of onlookers is as follows:

$$P_i = \frac{\text{fit}(x_i)}{\sum_{i=1}^{N} \text{fit}(x_i)}. \tag{4}$$

Step 5. Onlookers search for new nectar in accordance with (3), and compared with the nectar quantity searched by employed bees, Set the N position of more nectar as the position of employed bees; the rest is position of onlookers.

Step 6. If some nectar is unchanged after limit cycles, the nectar is given up, corresponding employed bees are turned into onlookers, and new nectar is randomly generated according to (2).

Step 7. Record location of best nectar source and return to Step 2 until the termination condition is met.

2.2. Particle Swarm Algorithm. Mathematical description of the particle swarm intelligence algorithm [5] is as follows. Suppose in a D-dimensional target space, N particles with potential problem solutions composed a group, where ith particle is represented as a D-dimensional vector, $X_i = [X_{i1}, X_{i2}, \ldots, X_{iD}]^T$ ($i = 1, 2, \ldots, N$); position of the ith particle in D-dimensional search space is X_i; flight speed is V_i; P_i is the personal best position searched by ith particle so far; and remember P_g is the global optimal position searched by particle swarm so far; in each iteration, particles update speed and position in accordance with

$$\begin{aligned} V_i^{t+1} &= wV_i^t + c_1 r_1 \left(P_i^t - X_i^t\right) + c_2 r_2 \left(P_g^t - X_i^t\right), \\ X_i^{t+1} &= X_i^t + V_i^{t+1}, \end{aligned} \tag{5}$$

where, $i = 1, 2, 3, \ldots, N$, t is the number of iterations; w is the inertia coefficient; c_1, c_2 are learning factors and suitable c_1, c_2 can speed up convergence and not easily fall into local optimum; and r_1, r_2 are random numbers in $[0, 1]$. Particles find P_g which is the global optimal solution via constantly learning and updating [5, 12–14]. Particle swarm algorithm is applied to the positioning phase of the proposed algorithm; the main steps are as follows.

Step 1. Determine parameters: the number of particles N, the inertia factor w, and the number of iterations t.

Step 2. Randomly generate a population of N particles.

Step 3. Update velocity and position of particle using (5).

Step 4. Global optimal solution P_g is obtained by comparing and calculating the fitness function values; solution is the coordinates of unknown node.

Step 5. Determine whether the condition of the loop termination was met, if it was, record coordinates of unknown node, otherwise return to Step 3.

3. Improved Artificial Bee Colony Algorithm

3.1. Chaos Reverse Learning Strategies. Population initialization is particularly important in intelligent algorithm, because initialization quality directly affects the algorithm global convergence speed and the corresponding solution quality. Under normal circumstances, due to the lack of a priori information, random initialization is often used to generate the initial solution of algorithm. Reference [15] proposed a chaotic initialization method in the process of researching particle swarm algorithm, while [16] proposed initialized method of reverse learning. On this basis, this paper proposed a chaotic reverse learning strategy by combining these two initialization methods, and the strategy was used to initialize ABC algorithm; concrete steps are as follows.

(1) Set maximum chaotic iteration step $K \geq 400$ and the population size $2N$. The N-S charts of chaotic phase and reverse learning phase are given in Figures 1 and 2.

(2) Select $2N$ best fitness value particles as the initial bee swarm from $\{V(2N) \cup \text{Opl\_}V(2N)\}$.

3.2. Dynamic Self-Adaptive Nectar Update Strategy. In the process of searching nectar source, employed bees often choose nectar source with more nectar quantity, but when many employed bees select the same nectar, this information amount of nectar will increase in vain, which causes too many employed bees to concentrate on one nectar source, causing blockage or stagnation. When solving the optimization problem, this will manifest premature and local convergence [17–20]. In order to solve this problem, a new dynamic self-adaptive nectar update strategy was proposed. This strategy introduced the concept of population diversity measurement and it was used to the redefinition of nectar update formula, in order to improve the algorithm search capabilities.

Reference [21] pointed out that the difference between the average particle distance and particle fitness was commonly used to indicate the population diversity. On the basis of analyzing disadvantages of this approach, the similarity degree of the individuals in population was used to characterize the population diversity, which was introduced to the updated nectar formula that is formula (3). Let individual number of ABC algorithm be $2N$, and the jth individual of ith generation bee colony $\mathbf{x}_i$ is $\mathbf{x}_{i,j} = (x_{i,j(1)}, x_{i,j(2)}, \ldots, x_{i,j(n)})$, where n is the number of nectar solution dimensions; the jth individual successive dynasties nectar optimum position is $\mathbf{y}_j = (y_{j(1)}, y_{j(2)}, \ldots, y_{j(n)})$. Combine individual nectar position and successive dynasties optimum position together, referred to as $\mathbf{Z}_j = (x_{i,j(1)}, x_{i,j(2)}, \ldots, x_{i,j(n)})$; all individuals $\mathbf{Z}_j$ in bee colony can be composed of a matrix $\mathbf{Z}$ of $2N \times 2n$ order, normalization process $\mathbf{Z}$, and matrix $\mathbf{Z}$ of $2N \times 2n$ order can be obtained as

$$\mathbf{Z}' = \frac{\mathbf{Z}_{uv} - \min_{1 \leq g \leq 2N, 1 \leq l \leq 2n} \mathbf{Z}_{gl}}{\max_{1 \leq g \leq 2N, 1 \leq l \leq 2n} \mathbf{Z}_{gl} - \min_{1 \leq g \leq 2N, 1 \leq l \leq 2n} \mathbf{Z}_{gl}}, \quad (6)$$

where $1 \leq u \leqq 2N$, $1 \leq v \leq 2n$, and each row vector can be seen as a fuzzy set, expressed as membership degrees of the nectar current location and successive position of each component searched by jth employed bee or onlooker; any similarity degree of two $\mathbf{Z}'_u$, $\mathbf{Z}'_v$ can be expressed by nearness; that is,

$$L(u, v) = 1 - \frac{1}{2n} \sum_{t=1}^{2n} \left| \mathbf{Z}'_{u,t} - \mathbf{Z}'_{v,t} \right|. \quad (7)$$

Bees Diversity Measurement F can be expressed by population average nearness

$$F = \frac{2 \sum_{u=1}^{2N-1} \sum_{v=u+1}^{2N} L(u, v)}{2N(2N-1)}. \quad (8)$$

$F \leq 1$ is obtained by $0 \leq L(u, v) \leq 1$; if individuals in bee colony are identical, then the diversity is the worst; F is the maximum value 1.

In the update nectar formula of ABC algorithm, since r_{ij} is randomly generated numerical value in $[-1, 1]$, the relationship between nectar source and the diversity of employed bees and bee colony is ignored. Therefore, let r_{ij} therefore, let adjustment formula of r_{ij} be:

$$r_{ij}(k) = \frac{1}{(a - bF(k-1))}, \quad (9)$$

where $r_{ij}(k)$ is the update coefficient of kth generation bee colony, $F(k-1)$ is diversity measurement of $(k-1)$th bee colony, and a, b are constants. Update formula of improved nectar is

$$V_{ij}(k) = x_{ij} + r_{ij}(k) \times \left(x_{ij} - x_{kj} \right). \quad (10)$$

3.3. Dual Population Search Strategy. Since the update methods of ABC and PSO Algorithm individual, as well as different optimization strategies, the effect of optimization also varies. In order to improve the population diversity of ABC algorithm and accelerate the speed of algorithm searching for optimal solution, inspired by reference [8, 22], taking into consideration the advantages of particle swarm algorithm which has a simple structure, easy to implement, few parameters, and fast algorithm convergence rate in early stage, the advantages of this algorithm and ABC algorithm are combined to propose a dual population search strategy. Main ideas of the strategy are to randomly divide the population into two groups, each group using different optimization strategies to find optimal solution. Better solution is selected as the algorithm optimal solution after comparison. The specific process is as follows.

Step 1. Initialize population behaviors using chaos reverse learning strategy mentioned above.

Step 2. Initialized population is randomly divided into two groups; one group uses improved ABC algorithm mentioned above, nectar update using formula (10); another group uses particle swarm algorithm, individual update using formula (5).

Step 3. Two kinds of populations are searching for optimal solution in accordance with their respective search strategy under algorithm termination condition. And based on the idea of survival of the fittest, respective proceeds in accordance with the respective optimal solutions are compared, and position of the better solution is recorded.

Diversity of the population is ensured by mixing two populations and two populations parallel searching; at the same time, algorithm convergence rate is improved to a large extent; the algorithm has a higher convergence rate in reasonable computational complexity.

4. Convergence Analysis

IMABC algorithm in this paper determines convergence according to methods given in the literature [23–25].

4.1. Convergence Criteria. If the result of the iteration of optimization problem $\{A, f\}$ is x_k, then the next iteration is $x_{k+1} = Q(x_k, \eta)$, of which A is solution space, f is fitness function, and η is the solution which the algorithm has found. The function

$$R_{\delta,Z^+} = \begin{cases} \{x \in A \mid f(x) < \beta + \delta\}, & \beta < |\infty|,\ \delta > 0; \\ \{x \in A \mid f(x) < -Z^+\}, & \beta = -\infty; \end{cases} \tag{11}$$

is defined as the optimal area; if the algorithm finds a point in R_{δ,Z^+}, then the algorithm can be considered to find the optimal algorithm or approximate optimal solution.

Condition 1. $f(Q(x,\eta)) \le f(x)$; if $\eta \in A$, then $f(Q(x,\eta)) \le f(\eta)$. If the algorithm satisfies this condition, it can be stated that fitness is nonincremental.

Condition 2. For any Borel subset B of A, if in the set B the Lebesgue measure $v[B] > 0$, then $\prod_{k=0}^{\infty}(1 - u_k[B]) = 0$. If the algorithm satisfies the condition, it can be stated that after bee colony unlimitedly searches optimization, the probability of global optimum that cannot be found is 0.

Theorem 1. *Set the function f as measurable; A is a measurable subset of R^n, algorithm Q satisfies Conditions 1 and 2, and $\{x_k\}_{k=0}^{\infty}$ is the solution sequence generated by algorithm Q; there $\lim_{k\to\infty} P(x_k \in R_{\delta,Z^+}) = 1$.*

4.2. Algorithm Convergence Analysis

Lemma 2. *IMABC algorithm meets Condition 1.*

Proof. The algorithm uses chaotic reverse learning strategies to initialize population, double-population search is conducted in each iteration, and the optimal value is saved; that is,

$$H\left(P_{g,t}V_{i,t}x_i\right) = \begin{cases} P_{g,t}x_i, & f\left(V_{i,t}\right) \le f\left(x_i\right), \\ V_{i,t}x_i, & f\left(V_{i,t}\right) > f\left(x_i\right). \end{cases} \tag{12}$$

Condition 1 is met. □

Definition 3. Assuming optimal solution is g_{best}; optimal solution set is defined as $G = \{s = (X) \mid f(X) = f(g_{\text{best}}), s \in S\}$.

Theorem 4 (see [13, 23–25]). *In the algorithm, for bee colony state sequence $\{s(t); t \ge 0\}$, set G as a closed set in state space S.*

Proof. Set $\forall s_i \in G$, $\forall s_j \notin G$; for any transfer step length l, $l \ge 1$, the probability $P^l_{s_i,s_j}$ of bee colony state transferred from s_i to s_j by l steps can be obtained by bee colony algorithm

$$P^l_{s_i,s_j} = \sum_{s_{r_1}\in s} \cdots \sum_{s_{r_{l-1}}\in s} P\left(T_s\left(s_i\right) = s_{r_1}\right) \times P\left(T_s\left(s_{r_1}\right) = s_{r_2}\right) \cdots P\left(T_s\left(s_{r_{l-1}}\right) = s_j\right), \tag{13}$$

where $P(T_s(s_{r_{c-1}}) = s_{r_c})$ is the probability of bee colony state transferred from s_{c-1} to s_c, $1 \le c \le l$, and the probability is determined by transition probability of each bee; that is, $P(T_s(s_{r_{c-1}}) = s_{r_c}) = \prod_{m=1}^{SN} p(T_s(X_{im}) = X_{jm})$. SN is the number of bees. $P(T_s(s_{r_{c-1}}) = s_{r_c})$ exists in each expression in (13), since $s_{r_{c-1}} \in G$, $s_{r_c} \notin G$, $f(X_c) > f(X_{c-1}) = f(g_{\text{best}}) = \inf(f(a))$, $a \in A$; at least there exists $P(T_s(s_{r_{c-1}}) = s_{r_c}) = 0$, $P^l_{s_i,s_j} = 0$, so set G as a closed set in state space S. □

Theorem 5 (see [23–25]). *Bee colony state space S does not have a nonempty closed set M, making $M \cap G = \phi$.*

Proof. Assume that there exists a nonempty closed set M and $M \cap G = \phi$. Set $s_i = (g_{\text{best}}, g_{\text{best}}, \ldots, g_{\text{best}}) \in G$, $\forall s_j = (x_{j1}, x_{j2}, \ldots, x_{jd}) \in M$; since $f(X_{jc}) > f(g_{\text{best}})$, after a finite number of iterations, the probability of scouter transferred from X_i to X_j is $p_{sc}(T_s(X_i) = X_j) > 0$. Therefore when the step size is large enough, in expansion $P^l_{s_i,s_j}$ there must be a certain product expression greater than 0; that is $P(T_s(s_{r_{c-i}}) = s_{r_{c+i+1}}) > 0$. From (13) in Theorem 4, $P^l_{s_i,s_j} > 0$; M is not a closed set, which is contradicted with the question conditions. Therefore, there is no closed set in the state space S except G. □

Theorem 6 (see [24]). *Assume Markov Chain has a nonempty closed set E and there is no other nonempty closed set O; let $E \cap O = \phi$. Therefore, when $j \in E$, $\lim_{n\to\infty} P(X_n = j) = \pi_j$ and when $j \notin E$, $\lim_{n\to\infty} P(X_n = j) = 0$ [2].*

Theorem 7. *When bee colony is unlimitedly iterated and optimized, all state sequences are present in the optimal state set*

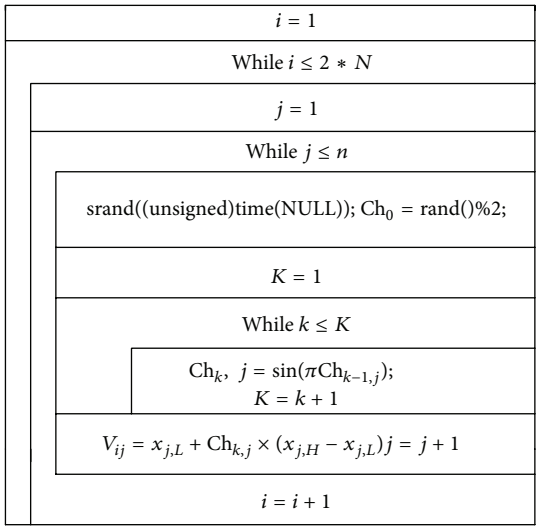


Figure 1: The N-S chart of chaotic phase.

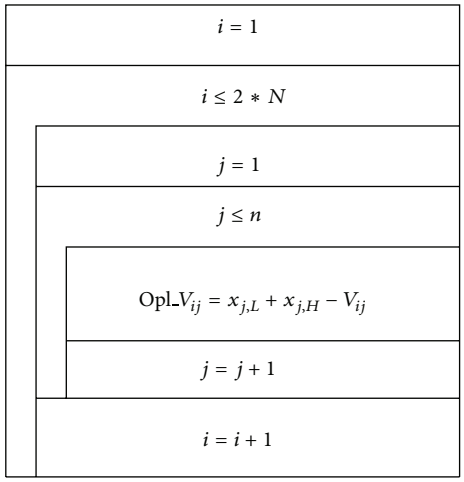


Figure 2: The N-S chart of reverse learning phase.

G. Theorem 7 is established which can be drawn from Theorems 4–6.

Lemma 8. *Algorithm satisfies Condition 2.*

Proof. By Theorem 7, after bee colony is unlimitedly optimized, the probability of no global optimum is 0; then there is $\prod_{k=0}^{\infty}(1 - u_k[B]) = 0$. □

Theorem 9. *IMABC algorithm converges to global optimum.*

As the bee colony algorithm satisfies Conditions 1 and 2, the algorithm can be obtained and converged to the global optimum by Theorem 1. The algorithm in this paper uses two optimization algorithms and the optimal solutions obtained from two populations are recorded; the two populations are independent. As long as the global optimal solution probability of one of the optimization algorithms converges to 1, the global optimal solution probability of the entire algorithm also converges to 1. Therefore, this algorithm is a global convergence algorithm.

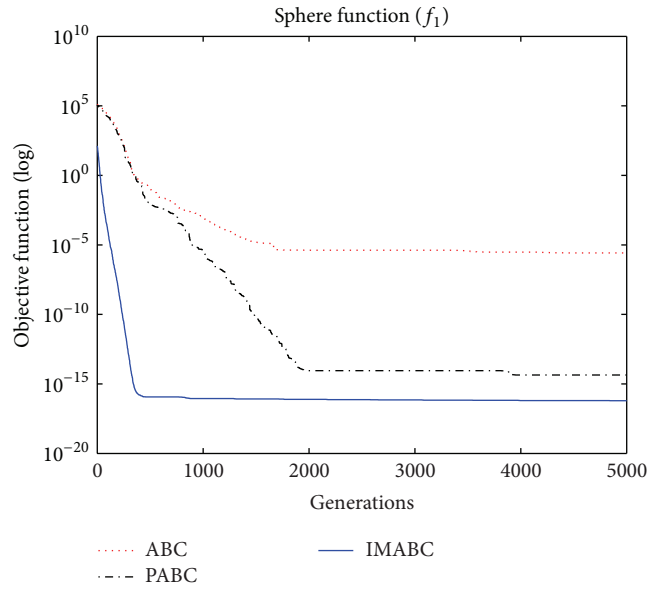


Figure 3: f_1 function convergence performance comparison.

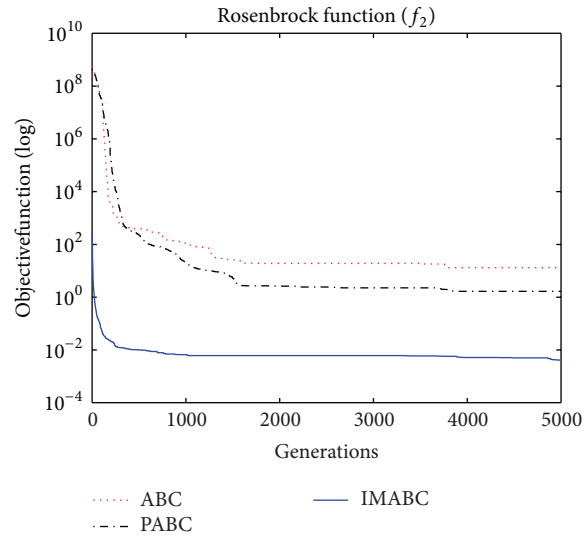


Figure 4: f_2 function convergence performance comparison.

5. Simulation Experiment and Results Analysis

In order to verify the validity of above analysis and the improved algorithm performance, comparison experiments were done on this improved algorithm (abbreviated as IMABC), traditional ABC algorithm, and improved integration algorithm (abbreviated as PABC) combined by ABC algorithm proposed by [7] and PSO algorithm. In the simulation experiment, 10 test functions [26–29] were selected; $f_1 \sim f_8$ are high-dimensional functions, where f_1 and f_2 are unimodal functions; $f_3 \sim f_8$ are multimodal functions; and

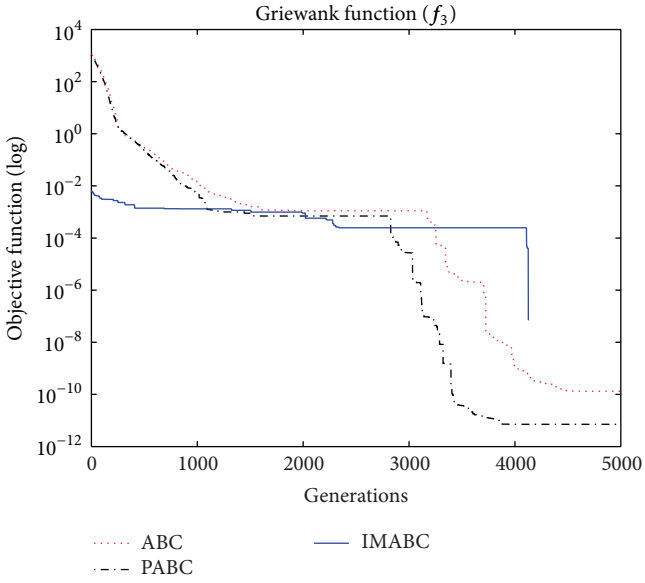


Figure 5: f_3 function convergence performance comparison.

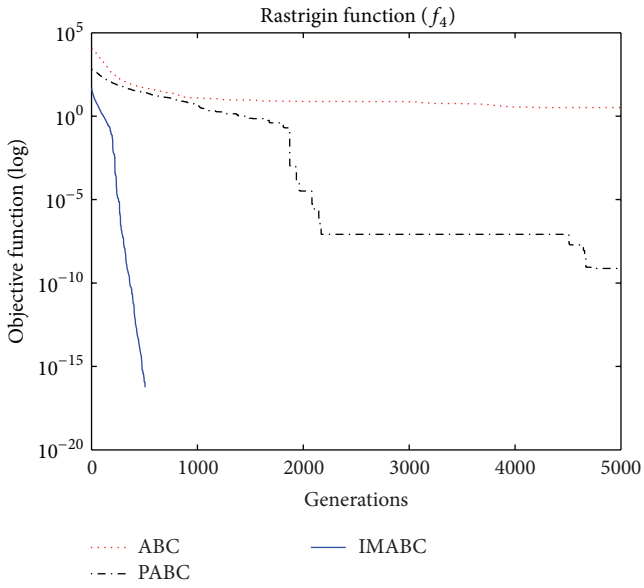


Figure 6: f_4 function convergence performance comparison.

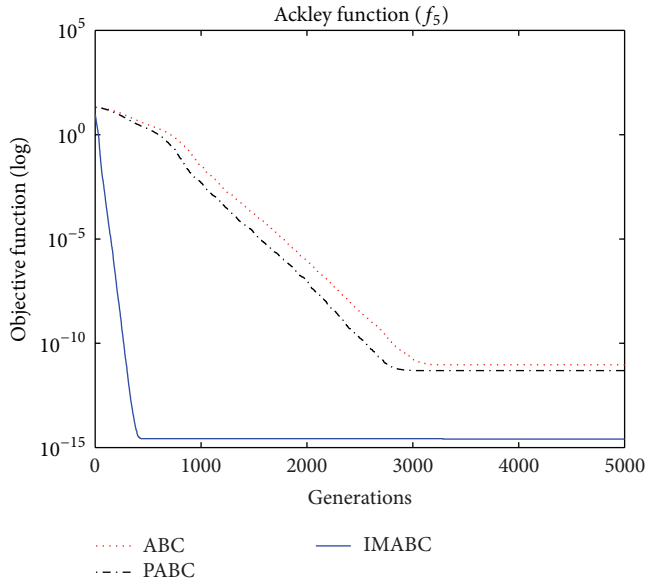


Figure 7: f_5 function convergence performance comparison.

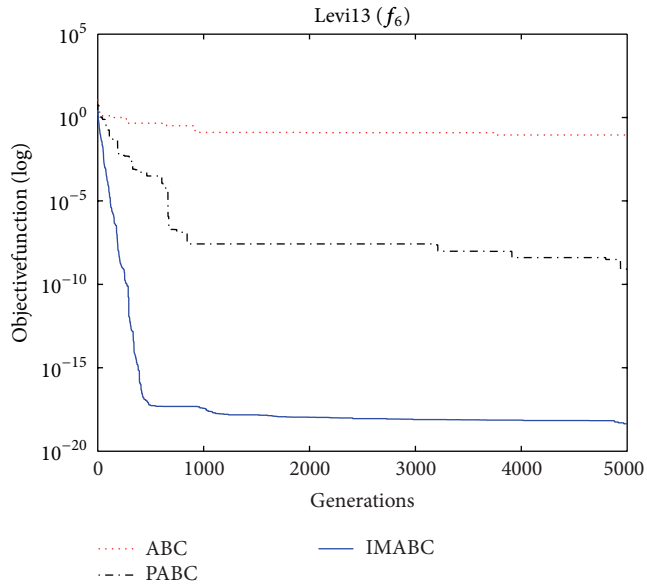


Figure 8: f_6 function convergence performance comparison.

f_9 and f_{10} are two-dimensional functions. Table 1 lists names, dimensions, definitions, ranges, and theory global optimal solutions of these test functions. In the experiments, population sizes of the three algorithms are all 40, K is 300, limit is 40, and the corresponding maximum number of iterations is 5000, where parameter settings of dual population strategy particle swarm algorithm are seen in [27]; that is, $c_1 = c_2 = 2$, $w = 0.4$. For each test function, every algorithm is randomly run 30 times to find the best value, the worst value, average, and standard deviation. The best and worst values reflect the solution quality; average tells the accuracy that algorithm can achieve under a given number of function evaluations, reflecting the algorithm convergence rate; variance reflects the stability and robustness of the algorithm. Results are shown in Table 2.

As can be seen from the data comparison in Table 2, among most standard test functions, whether it is the solution quality or algorithm convergence accuracy and stability, IMABC algorithm has been greatly more improved than PABC algorithm and ABC algorithm. In functions f_1 and f_9, although compared to PABC algorithm IMABC algorithm's minimum, the worst value, average, and variance all slightly increase, significant improvement is achieved when compared to the standard ABC algorithm. In functions f_2, f_5, f_6, and f_8, the improved algorithm is significantly better than the standard ABC algorithms and PABC algorithm in various

Table 1: Dimension, search space, and optimal value of test functions.

Function	Mathematical representation	Dimension (D)	Range of search (S)	Theoretical optimum $f_{\min}$
Sphere (f_1)	$f_1(x) = \sum_{i=1}^{D} x_i^2$	50	$[-100, 100]^D$	0
Rosenbrock (f_2)	$f_2(x) = \sum_{i=1}^{D} \left[100\left(x_{i+1} - x_i^2\right)^2 + \left(x_i - 1\right)^2\right]$	50	$[-30, 30]^D$	0
Griewank (f_3)	$f_3(x) = \frac{1}{4000}\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	50	$[-600, 600]^D$	0
Rastrigin (f_4)	$f_4(x) = \left(\sum_{i=1}^{D} x_i^2 - 10\cos(2\pi x_i) + 10\right)$	50	$[-5.12, 5.12]^D$	0
Ackley (f_5)	$f_5(x) = -20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right) - \exp\left(-0.2\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)\right) + 20 + e$	50	$[-30, 30]^D$	0
Levi11 (f_6)	$f_6(x) = \sin(3\pi x_1)^2 + (x_1 - 1)^2(1 + \sin(3\pi x_1)^2) + (x_2 - 1)^2(1 + \sin(2\pi x_2)^2)$	50	$[-10, 10]^D$	0
Schwefel's problem (f_7)	$f_7(x) = \sum_{i=1}^{D} \lvert x_i\rvert + \prod_{i=1}^{D} \lvert x_i\rvert$	50	$[-500, 500]^D$	−837.9658
Alpine (f_8)	$f(x) = \sum_{i=1}^{D} \lvert x_i \sin(x_i) + 0.1x_i\rvert$	50	$[-100, 100]^D$	0
Cross (f_9)	$f_{10}(x) = \left\lvert \sin(x_1)\sin(x_2)e^{\lvert 100 - \sqrt{x_1^2 + x_2^2}/\pi\rvert} + 1\right\rvert$	2	$[-10, 10]^D$	0
Schaffer's F6 (f_{10})	$f_9(x) = 0.5 + \frac{\sin^2\left(\sqrt{x_1^2 + x_2^2}\right) - 0.5}{\left(1 + 0.001(x_1^2 + x_2^2)\right)^2}$	2	$[-100, 100]^D$	0

Table 2: Results comparison of 10 test functions of 3 algorithms.

Function	Algorithm	Best	Worst	Mean	Std.
f_1	ABC	8.85038e − 015	7.93808e − 005	2.65944e − 006	1.44905e − 005
	PABC	2.50765e − 015	4.22993e − 014	9.22438e − 015	8.65892e − 015
	IMABC	3.48561e − 017	8.9191e − 017	6.2547e − 017	1.33411e − 017
f_2	ABC	5.90141	44.6327	18.4201	10.1225
	PABC	0.185485	7.19489	1.54921	1.72969
	IMABC	0.000204645	0.00921194	0.00381647	0.00285877
f_3	ABC	4.44089e − 015	3.56306e − 009	1.32808e − 010	6.48804e − 010
	PABC	4.996e − 015	1.44799e − 010	7.14078e − 012	2.67773e − 011
	IMABC	0	0	0	0
f_4	ABC	1.20019e − 008	11.6005	2.66884	2.94451
	PABC	7.42517e − 013	1.32277e − 005	5.01149e − 007	2.40813e − 006
	IMABC	0	0	0	0
f_5	ABC	6.59917e − 013	1.04211e − 010	9.28022e − 012	1.86668e − 011
	PABC	4.21885e − 013	2.06368e − 011	4.90198e − 012	4.5731e − 012
	IMABC	8.88178e − 016	2.66454e − 015	2.54611e − 015	6.48634e − 016
f_6	ABC	0.000957132	0.259845	0.0545307	0.0597714
	PABC	1.18297e − 013	3.51652e − 009	5.77724e − 010	9.57809e − 010
	IMABC	6.08383e − 021	1.4873e − 018	3.51197e − 019	2.97612e − 019
f_7	ABC	−837.767	−826.26	−833.611	3.18261
	PABC	−730.356	−730.356	−730.356	3.46891e − 013
	IMABC	−837.966	−837.966	−837.966	0
f_8	ABC	2.37327	6.15694	4.09479	1.21718
	PABC	0.0443285	0.419608	0.200788	0.142808
	IMABC	2.99468e − 016	5.48062e − 016	4.18311e − 016	8.39077e − 017
f_9	ABC	4.85217e − 005	4.95895e − 005	4.87645e − 005	3.21009e − 007
	PABC	4.84822e − 005	4.86088e − 005	4.85033e − 005	3.7978e − 008
	IMABC	4.84822e − 005	4.84822e − 005	4.84822e − 005	4.03232e − 018
f_{10}	ABC	0.000235347	0.0851231	0.0240707	0.0235249
	PABC	8.98791e − 011	8.1512e − 005	6.98519e − 006	1.63762e − 005
	IMABC	0	0	0	0

test results; especially in functions f_3, f_4, f_7, and f_{10}, IMABC algorithm not only has good test results but also can converge to optimal solution, showing good searching performance.

In order to compare algorithm optimization effect more visually, IMABC algorithm, PABC algorithm, and ABC algorithm are compared. Corresponding test function convergence curves are given in Figures 3 to 12. According to the figures, because of the use of a new initialization method and dual population parallel search strategy, as well as the dynamic self-adaptation of nectar location update, the improved Artificial Bee Colony algorithm can jump out of local optimal solution and gradually converge to the global optimal solution when processing multimodal functions, and has a faster convergence rate when in processing unimodal functions and low-dimensional functions. While processing unimodal functions and low-dimensional functions, this improved algorithm has a faster convergence rate. It can be seen from Figures 3, 4, 7, and 8, since standard test function has a high complexity in 50th-dimension, three algorithms are all unable to converge to the optimal solution, but this improved algorithm, compared to ABC algorithm and PABC algorithm, has a faster convergence rate and significantly superior convergence accuracy; it can be seen from Figures 5, 6, 9, and 12 that, compared to the other two algorithms, IMABC algorithm has higher convergence accuracy and can converge to a global optimal solution faster and stabilize. As can be seen from Figure 10, this proposed algorithm gradually approaches function optimal solution with the increase of iterations, although it cannot converge; the extent of approaching and search accuracy are significantly better than the other two algorithms. As can be seen from Figure 11, IMABC algorithm and the other two algorithms are all approaching function optimal solution with the increase of iterations. When the number of iterations is more than 60 times, this algorithm has little difference with PABC algorithm. But when the number of iterations is less than 60 times, it can be seen that IMABC algorithm has higher convergence rate. Therefore, it can be concluded that

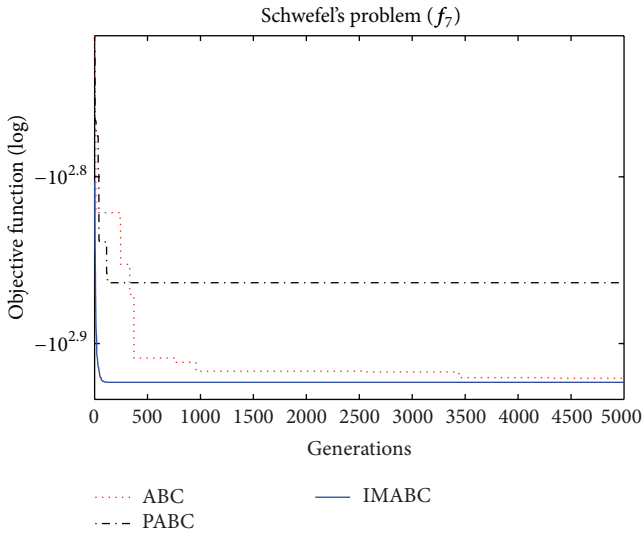


Figure 9: f_7 function convergence performance comparison.

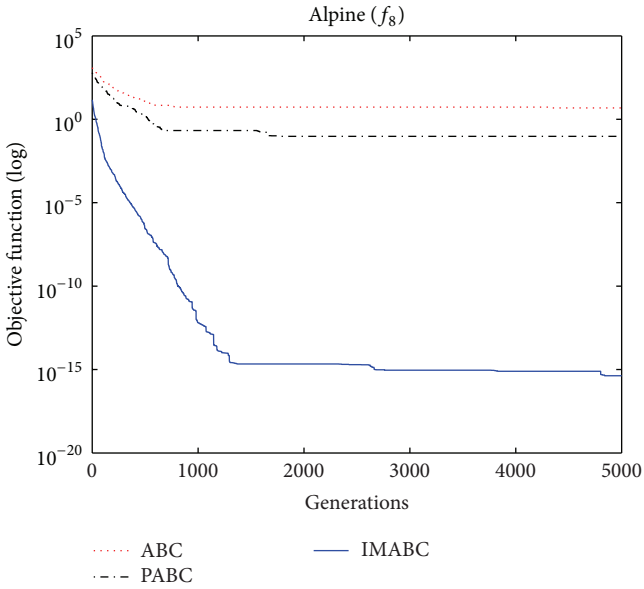


Figure 10: f_8 function convergence performance comparison.

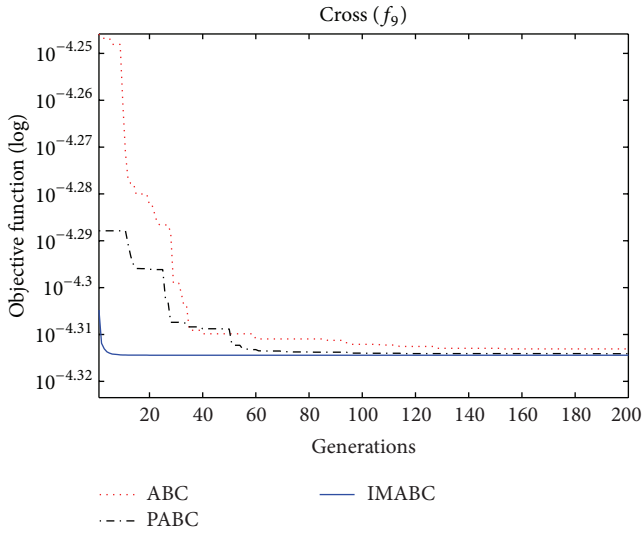


Figure 11: f_9 function convergence performance comparison.

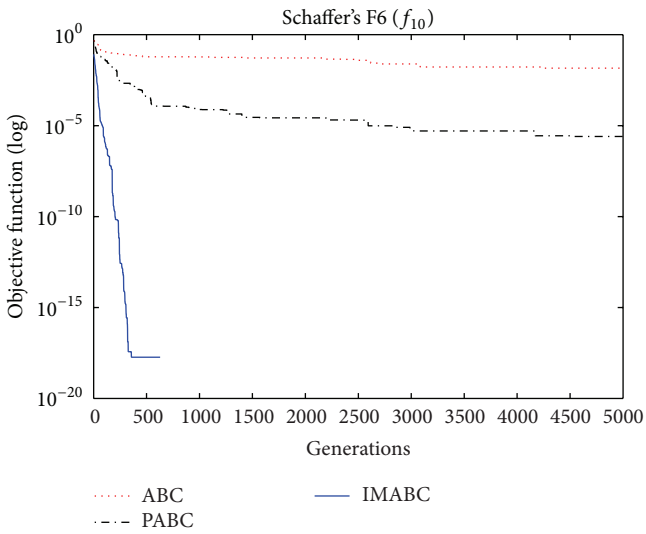


Figure 12: f_{10} function convergence performance comparison.

the overall optimization performance of this proposed IMABC algorithm is superior to the standard ABC algorithm and the PABC improved algorithm proposed in [7].

6. Conclusion

In order to avoid falling into local optimum resulting from premature and improve convergence rate of ABC algorithm, an improved artificial bee colony algorithm based on multipolicy optimization was proposed. In order to improve the global search ability and keep the algorithm diversity, improved algorithm proposed a chaotic reverse learning initialization method on the basis of existing research results; in order to avoid the algorithm falling into a local optimum, improved algorithm introduced dual population search mechanism into search phase of the algorithm. Advantages of particle swarm algorithm and standard ABC algorithm were merged; meanwhile, algorithm convergence rate was increased. In addition, in order to improve the algorithm population diversity and global search capability, the concept of population similarity degree was introduced into the improved algorithm, and an indicator of population diversity measure was proposed to dynamic self-adaptive adjustment of the nectar location. Experimental results of 10 standard test functions optimization showed that this proposed algorithm improved more greatly than standard ABC algorithm and PABC in optimization efficiency, optimization performance, and robustness.

Furthermore, this improved algorithm also has certain limitations: though optimization performance is improved,

the algorithm complexity is increased to a certain extent. How to ensure algorithm jumping out of local optima and having high convergence rate, at the same time, possessing low algorithm complexity, will be the next step in research.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the Scientific Research Program of the Higher Education Institution of XinJiang (no. XJEDU2010S48).

References

[1] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Tech. Rep., Erciyes University, Engineering Faculty, Computer Engineering Department, Kayseri, Turkey, 2005.

[2] D. Karaboga and B. Basturk, "Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems," in *Foundations of Fuzzy Logic and Soft Computing*, vol. 4529 of *Lecture Notes in Computer Science*, pp. 789–798, Springer, Berlin, Germany, 2007.

[3] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Applied Soft Computing*, vol. 8, no. 1, pp. 687–697, 2008.

[4] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[5] A. H. Anderson, "A comparison of two privacy policy languages: EPAL and XACML," in *Proceedings of the 3rd ACM Workshop on Secure Web Services (SWS '06)*, pp. 53–60, ACM, 2006.

[6] A. Singh, "An artificial bee colony algorithm for the leaf-constrained minimum spanning tree problem," *Applied Soft Computing Journal*, vol. 9, no. 2, pp. 625–631, 2009.

[7] G. P. Zhu and S. Kwong, "Gbest-guided artificial bee colony algorithm for numerical function optimization," *Applied Mathematics and Computation*, vol. 217, no. 7, pp. 3166–3173, 2010.

[8] A. Abraham, R. K. Jatoth, and A. Rajasekhar, "Hybrid differential artificial bee colony algorithm," *Journal of Computational and Theoretical Nanoscience*, vol. 9, no. 2, pp. 1–9, 2012.

[9] N. Karaboga, "A new design method based on artificial bee colony algorithm for digital IIR filters," *Journal of the Franklin Institute*, vol. 346, no. 4, pp. 328–348, 2009.

[10] L. Wang, G. Zhou, and Y. Xu, "An artificial bee colony algorithm for solving hybrid flow-shop scheduling problem with unrelated parallel machines," *Control Theory & Applications*, vol. 29, no. 12, pp. 1551–1556, 2012.

[11] J.-Q. Li, Q.-K. Pan, and K.-Z. Gao, "Pareto-based discrete artificial bee colony algorithm for multi-objective flexible job shop scheduling problems," *International Journal of Advanced Manufacturing Technology*, vol. 55, no. 9–12, pp. 1159–1169, 2011.

[12] Q. Li, J. Gong, and J.-F. Tang, "Multi-objective particle swarm optimization algorithm for cross-training programming," *Control Theory & Applications*, vol. 30, no. 1, pp. 18–22, 2013.

[13] Z.-H. Zhan, J. Zhang, Y. Li, and H. S.-H. Chung, "Adaptive particle swarm optimization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 6, pp. 1362–1381, 2009.

[14] C. W. Jiang and B. Etorre, "A hybrid method of chaotic particle swarm optimization and linear interior for reactive power optimisation," *Mathematics and Computers in Simulation*, vol. 68, no. 1, pp. 57–65, 2005.

[15] B. Liu, L. Wang, Y.-H. Jin, F. Tang, and D.-X. Huang, "Improved particle swarm optimization combined with chaos," *Chaos, Solitons and Fractals*, vol. 25, no. 5, pp. 1261–1271, 2005.

[16] R. S. Rahnamayan, H. R. Tizhoosh, and M. M. A. Salama, "Opposition-based differential evolution," *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 1, pp. 64–79, 2008.

[17] G. P. Zhu and S. Kwong, "Gbest-guided artificial bee colony algorithm for numerical function optimization," *Applied Mathematics and Computation*, vol. 217, no. 7, pp. 3166–3173, 2010.

[18] Y. Liu and L. Ma, "Bees algorithm for function optimization," *Control and Decision*, vol. 27, no. 6, pp. 886–889, 2012.

[19] W. F. Gao and S. Y. Liu, "Improved artificial bee colony algorithm for global optimization," *Information Processing Letters*, vol. 111, no. 17, pp. 871–882, 2011.

[20] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 398–417, 2009.

[21] D.-X. Zhang, Z.-H. Guan, and X.-Z. Liu, "Adaptive particle swarm optimization algorithm with dynamically changing inertia weight," *Control and Decision*, vol. 23, no. 11, pp. 1253–1257, 2008.

[22] B. Alatas, "Chaotic bee colony algorithms for global numerical optimization," *Expert Systems with Applications*, vol. 37, no. 8, pp. 5682–5687, 2010.

[23] J. C. Zeng, J. Jie, and Z. H. Cui, *Particle Swarm Algorithm*, Science Press, Beijing, China, 2004.

[24] J.-P. Luo, X. Li, and M.-R. Chen, "The Markov model of shuffled frog leaping algorithm and its convergence analysis," *Acta Electronica Sinica*, vol. 38, no. 12, pp. 2875–2880, 2010.

[25] H. Zhang, H. Wang, and Z. Hu, "Analysis of particle swarm optimization algorithm global convergence method," *Computer Engineering and Applications*, vol. 47, no. 34, pp. 61–63, 2011.

[26] X. Yao, Y. Liu, and G. M. Lin, "Evolutionary programming made faster," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 2, pp. 82–102, 1999.

[27] J. Kennedy and R. Mendes, "Population structure and particles warm performance," in *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 1671–1676, Honolulu, Hawaii, USA, 2002.

[28] F. Kang, J. J. Li, and Z. Y. Ma, "Rosenbrock artificial bee colony algorithm for accurate global optimization of numerical functions," *Information Sciences*, vol. 181, no. 16, pp. 3508–3531, 2011.

[29] Q.-K. Pan, M. F. Tasgetiren, P. N. Suganthan, and T. J. Chua, "A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem," *Information Sciences*, vol. 181, no. 12, pp. 2455–2468, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 695168, 9 pages
http://dx.doi.org/10.1155/2014/695168

Research Article

A Novel Key-Frame Extraction Approach for Both Video Summary and Video Index

Shaoshuai Lei, Gang Xie, and Gaowei Yan

Taiyuan University of Technology, No. 79 West Yingze Avenue, Taiyuan, Shanxi 030024, China

Correspondence should be addressed to Gang Xie; xiegang@tyut.edu.cn

Received 23 December 2013; Accepted 5 February 2014; Published 16 March 2014

Academic Editors: Z. Zhou and X. Zhu

Existing key-frame extraction methods are basically video summary oriented; yet the index task of key-frames is ignored. This paper presents a novel key-frame extraction approach which can be available for both video summary and video index. First a dynamic distance separability algorithm is advanced to divide a shot into subshots based on semantic structure, and then appropriate key-frames are extracted in each subshot by SVD decomposition. Finally, three evaluation indicators are proposed to evaluate the performance of the new approach. Experimental results show that the proposed approach achieves good semantic structure for semantics-based video index and meanwhile produces video summary consistent with human perception.

1. Introduction

In the last few years, the prompt increasing of video data needs efficient techniques for browsing and index of this data [1]. However, the substantially different nature of video data is not suited for conventional data management techniques. Therefore, much research work has been done about the key-frame extraction which can convert video processing to image processing. Key-frames, also called representative frames, are defined as the most informative frames that capture the major elements in a video in terms of content. Key-frames can generate summaries of the videos to provide browsing capabilities to the users [2, 3]. Apart from browsing, key-frames can also help users quickly locate a semantically relevant position in a video, namely, generating an index for a video.

1.1. Related Work. Early key-frame extraction approaches can be classified into two categories: based on interframe difference and based on clustering. In the approaches based on interframe difference, a new key-frame is extracted only if the interframe difference overtakes a certain threshold [4–6]. Clustering-based approaches try to group frames with similar low-level features and select the frame closest to each cluster centre as a key-frame [7–10]. These approaches may not grasp the interesting events and objects for viewers or they cannot find visually salient key-frames. Therefore, the semantically relevant approaches are advanced, and the representative categories include based on motion and based on visual attention approaches.

The approaches based on motion think that motion is an intrinsic attribute of video and human eyes are very sensitive to motion; thus they take into account motion events and camera operations in key-frame extraction. In [11], Liu et al. apply a triangle model of perceived motion energy (PME) to model a motion event and determine the frame with the maximum motion energy as a key-frame. Ma et al. [12] assume that the change of motion states attracts more attention than motion itself. They define the frames with the most significant acceleration (MSA) as key-frames. Some researchers [13, 14] believe that video content will change after each camera operation, such as pan, zoom, and tilt; therefore they determine key-frames by detecting camera movements.

The approaches based on visual attention attempt to find the semantically relevant key-frames by simulating human visual perception mechanism. The approaches usually combine several representative feature maps (values) into a single saliency map (value) which can be used as an indication of the attention level. Lai and Yi [15] first compute dynamic and static attention values of each frame based on motion, color, and texture features, then the two attention values are fused to build an attention curve of a video, and finally

the key-frames are extracted at the crests of this attention curve. In the work of [16], spatial attention value is computed based on the foreground of an image, and temporal attention value is obtained based on the changes in pixel values across neighboring frames to highlight the important areas of interframe motion. The static and dynamic visual attention values are fused nonlinearly into an attention curve for key-frame extraction. Aiming at sports videos, [17] uses prior knowledge to extend the visual attention model in which spatial, temporal, facial, and contextual attention features are fused.

1.2. Methodology. All semantically relevant methods attempt to find the key-frames by recognizing video semantic content; yet automatic understanding of semantic content is unachievable for contemporary computers, and there are many unsolved problems, especially the following two problems.

(1) All existing methods focus on video summary yet ignore the index task of key-frames. A new key-frame extraction approach should be found so as to take account of both tasks. It is the future direction of development and remains an important challenge in which establishing semantic structure for a video is the essential part.

(2) Current methods only extract the frame at each peak point which easily leads to content jumps in video summary. Therefore, some intermediate frames, having continuity and similarity in video content, are needed to help viewers to infer the original video content.

To address these problems, this paper proposes a new key-frame extraction method, and the basic concept can be described as follows.

(1) This paper divides a shot into several clips (hereafter called subshot) in chronological order according to the overall discrepancies between video frames themselves. Each subshot consists of similar content frames, and there are great visual differences between subshots. Since similar video content expresses the same semantic element, subshot segmentation also means semantic structure division which is the basis of video index.

(2) After subshot segmentation, proper key-frames from the same subshot can ensure visual continuity. If each frame is represented as an m-dimensional vector, the subshot including n frames can be expressed as a $m \times n$ matrix $\mathbf{A}$ and the key-frame extraction can be seen as subset selection. As singular values can reflect the rank of a matrix, this paper computes the approximate rank of matrix $\mathbf{A}$ by singular values to determine the number of key-frames then uses the distance of adjacent frames to determine the specific locations of key-frames.

The algorithm can be separated into four steps: (1) extract an HSV color feature vector for each frame; (2) divide the shot into subshots using a dynamic distance separability algorithm; (3) calculate the number k of key-frames by SVD decomposition; and (4) extract the k frames with the largest visual differences as key-frames in each subshot.

The remainder of this paper is organized as follows. The subshot segmentation method is described in Section 2, and the key-frame extraction from subshots is described in Section 3. Experimental results and evaluations of the new approach are, respectively, presented in Sections 4 and 5. Finally, conclusions are stated in Section 6.

2. Subshot Segmentation

2.1. Feature Extraction. Compared with other color spaces, HSV color space is the closest to the characteristics of human vision [18]. Because the human eyes are most sensitive to hue component, the hue H is divided into 7 parts, the saturation S into 2 parts, and the brightness V into 2 parts, and the quantization is shown in (1) through (3). When S is small enough ($s < 0.2$), the perceptual color is a black area; therefore the range can be neglected. Similarly, when V is small enough ($v < 0.2$), it is neglected as a gray area:

$$H = \begin{cases} 0, & \text{if } h \in (330, 360] \cup (0, 22] \\ 1, & \text{if } h \in (22, 45] \\ 2, & \text{if } h \in (45, 70] \\ 3, & \text{if } h \in (70, 155] \\ 4, & \text{if } h \in (155, 186] \\ 5, & \text{if } h \in (186, 278] \\ 6, & \text{if } h \in (278, 330], \end{cases} \tag{1}$$

$$S = \begin{cases} 0, & \text{if } s \in (0.2, 0.65] \\ 1, & \text{if } s \in (0.65, 1.0], \end{cases} \tag{2}$$

$$V = \begin{cases} 0, & \text{if } v \in (0.2, 0.7] \\ 1, & \text{if } v \in (0.7, 1.0]. \end{cases} \tag{3}$$

The three color components are synthesized one-dimensional vector $\mathbf{x}$ by (4), in which Q_s and Q_v represent the quantization level of the components S and V, $Q_s = 2$, $Q_v = 2$. Thus the range of values of $\mathbf{x}$ is from 0 to 27, and this means that each frame can be represented by a column vector $\mathbf{x}$, as shown in

$$\mathbf{x} = HQ_sQ_v + SQ_v + V, \tag{4}$$

$$\mathbf{x} = [c_1, c_2, \ldots, c_{28}]^T. \tag{5}$$

2.2. Subshot Segmentation. The frames within a subshot, showing similar video content, can be considered the same class, and different subshots can be viewed as different classes. According to distance separability criterion, the greater the between-class distance and the smaller the within-class distance, the higher the separability of two classes. Applying this criterion to subshot segmentation, that is to find the border frames which make that the greatest between-class distance between two subshots on the sides of a border frame

and the smallest within-class distance among each subshot. This paper extends this criterion to a dynamic distance separability algorithm for subshot segmentation, and the process can be described as follows: a sliding window of length $2L + 1$ is established in the frame sequences and the preceding L frames of the sliding window are selected as sample set ω_1, while the latter L frames are selected as sample set ω_2. The sliding window is moved back frame by frame, and, at each position, the within-class distance and between-class distances of the two sample sets are calculated. When the ratio of between-class distance and within-class distance reaches a local-maximum, the middle frame of the sliding window, where the video content changes dramatically, is selected as the border frame of two subshots.

This approach uses the dynamic distance separability to achieve subshot segmentation. This method uses the overall differences among frames to track video content changes, rather than some certain factors such as objects, motions, or other physical characteristics, which assures the accuracy and robustness. In addition, similar video content carries identical semantic element; therefore subshot segmentation based on video content is equivalent to subshot segmentation based on semantic structure.

2.2.1. Dynamic Distance Separability Algorithm

(1) Establish a sliding window of length $2L + 1$ and select the preceding L frames as sample set ω_1 and the latter L frames as sample set ω_2.

(2) Calculate the mean vector $\overline{\mathbf{m}}_i$ of sample set ω_i ($i = 1, 2$) according to (6). $\overline{\mathbf{m}}_1$ represents the mean vector of sample set ω_1 (the preceding L frames), and $\overline{\mathbf{m}}_2$ represents the mean vector of sample set ω_2 (the latter L frames), where L is the number of frames in each sample set and $\mathbf{x}$ is the feature vector of each frame as given in (5):

$$\overline{\mathbf{m}}_i = \frac{1}{L}\sum_{\mathbf{x}\in\omega_i}\mathbf{x}, \quad i = 1, 2. \tag{6}$$

(3) There are various definitions of distance separability criteria. In practice, the most widely used criterion is based on the within-class dispersion matrix and the between-class dispersion matrix. Equations (7) and (8) are, respectively, used to represent the within-class dispersion matrix $\mathbf{S}_{\omega_i}$ and the between-class dispersion matrix $\mathbf{S}_b$. The within-class dispersion matrix expresses the dispersion of each sample around the mean vector, and the between-class dispersion matrix expresses the distance distribution between two sample sets:

$$\mathbf{S}_{\omega_i} = \frac{1}{L}\sum_{\mathbf{x}\in\omega_i}\left(\mathbf{x} - \overline{\mathbf{m}}_i\right)\left(\mathbf{x} - \overline{\mathbf{m}}_i\right)^T, \quad i = 1, 2, \tag{7}$$

$$\mathbf{S}_b = \left(\overline{\mathbf{m}}_1 - \overline{\mathbf{m}}_2\right)\left(\overline{\mathbf{m}}_1 - \overline{\mathbf{m}}_2\right)^T. \tag{8}$$

(4) The greater the between-class dispersion, the smaller the within-class dispersion and the better the class separability. As shown in (9), we use the trace of matrix as the class separability criterion [19]. When F value reaches the maximum, the middle frame of sliding window coincidently lies on the border of two subshots:

$$F = \frac{\operatorname{trace}\left(\mathbf{S}_b\right)}{\operatorname{trace}\left(\mathbf{S}_{\omega_1}\right) + \operatorname{trace}\left(\mathbf{S}_{\omega_2}\right)}. \tag{9}$$

2.2.2. Calculation of F Value Curve. As the sliding window is moved backward frame by frame, the F value is calculated according to (9), and all the F values constitute a curve. When the sliding window is in the same subshot, the F values keep basically constant and even approximate zero in the ideal situation; when the latter L frames of sliding window step frame by frame into the next subshot, the F value increases gradually; when the latter L frames are entirely in the next subshot and the preceding L frames are still in the current subshot, the F-value achieves a local-maximum and subsequently decreases gradually until the preceding L frames also fall entirely in the next subshot. Therefore, the frames corresponding to the maximum values of F value curve can be taken as subshot segmentation boundaries. This process is illustrated by Figure 1, which depicts the F value curve of a video.

2.2.3. Subshot Segmentation. In the calculation of F-values, spikes caused by noise will occur in the curve. As shown in Figure 1, besides the two larger local-maximum points, there are also several minor local-maximum points which are not real subshot segmentation points. To remove these interferences, the real border frames are detected using

$$\lambda = \frac{F_i}{F_{\max}}, \tag{10}$$

where F_i represents the F-value at the ith largest local-maximum and $F_{\max}$ represents the largest F-value. Once the ratio exceeds threshold λ, the frame corresponding to F_i should be determined the border frame.

According to (9), if all F-values are less than 1, it means that there is almost no visual difference between the two halves of the sliding windows. Therefore, we add the following definition: if all F-values in F-curve are less than 1, the subshot segmentation is not needed.

Next, we need to determine the frame number of each local-maximum. Assume a new function $F' = f(i)$, where i is the sequence number of the frames in a shot, and $f(i)$ represents the F-value corresponding to the ith frame. The twice-difference method is used to extract the local-maxima as shown in

$$\mathrm{TD} = \operatorname{sign}\left[f(i+1) - f(i)\right] - \operatorname{sign}\left[f(i) - f(i-1)\right], \tag{11}$$

where $\operatorname{sign}(x)$ represents the sign function:

$$\operatorname{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0. \end{cases} \tag{12}$$

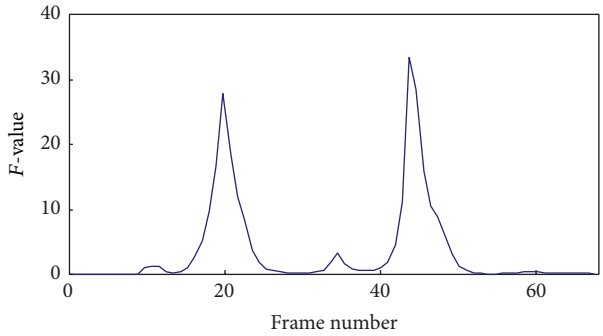


Figure 1: F-curve schematic diagram.

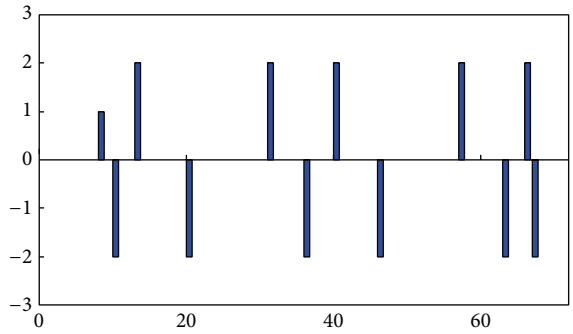


Figure 2: Local-maximum of F-curve.

At local-maximum points of the F-value curve, the twice-difference results are equal to −2; at local-minimum points, the twice-difference results are equal to 2; in other cases, the twice-difference results are equal to 0 or 1. The twice-difference results are shown in Figure 2, which unmistakably indicates the frame number corresponding to the local-maxima. Due to the existence of a sliding window, the F-values of the last L frames in a video cannot be computed. To prevent the last L frames being a separate subshot because of rapid content changes, the last L frames are classified as a subshot if the last F-value exceeds one-third of the maximum.

3. Key-Frame Extraction from Subshots

Existing algorithms consider only the spatial information of a frame, but not temporal characteristics between the frames. Therefore it is difficult to determine the number and the location of key-frames as a whole. If each frame is represented as a m-dimensional vector (in this paper, each frame is represented by a 28-dimension feature vector which has been mentioned in Section 2.1), the subshot including n frames can be expressed as a $m \times n$ matrix $\mathbf{A}$. Key-frame extraction problem can be transformed into finding maximal independent set of matrix $\mathbf{A}$, and the specific process includes the following two steps.

3.1. Calculate the Number of Key-Frames. Determine the number of key-frames; namely, determine the rank of matrix $\mathbf{A}$. We know that the number of singular values is equivalent to the rank of matrix. Video data is a nonstructured data, and there is not a simple linear relationship between video frames, so the rank of matrix $\mathbf{A}$ is usually too big. Therefore, we determine the approximate rank of matrix $\mathbf{A}$ by singular value decomposition (SVD). Concerning SVD, an important property given in Theorem 1 can be used in the determination of appropriate rank of matrix $\mathbf{A}$. The complete proof can be found in [20].

Theorem 1. *For* $\mathbf{A} \in \mathbf{R}^{m\times n}$, $q = \min(m, n)$, *if* $k < r = \operatorname{rank}(\mathbf{A})$ *and*

$$\mathbf{A}_k = \sum_{i=1}^{k} \boldsymbol{\sigma}_i \mathbf{u}_i \mathbf{v}_i^T, \tag{13}$$

then

$$\min_{\operatorname{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F = \sqrt{\sum_{i=k+1}^{q} \boldsymbol{\sigma}_i^2}. \tag{14}$$

Theorem 1 gives us significant implications. Discarding smaller singular values means removing the linearly semidependent or nonessential axes of the feature vector space. That is, the truncated SVD still reserves the most information of underlying spatiotemporal structure.

We use (15) to determine the approximate rank of matrix $\mathbf{A}$, namely, the number of key-frames to be extracted. As shown in (15), this equation remains the main information by the elimination of smaller singular values. The largest integer k that satisfies $v(k) \geq \alpha$ is selected as the appropriate rank r, where the larger the threshold α, the more the selected key-frames and the more the available video details.

For a static video, as the frames are very close in video content, they are approximate linear relation, which means that the rank of matrix $\mathbf{A}$ is very small. With the increasing complexity of video content, the nonlinear relationship between frames is enhanced; therefore singular values become more dispersed; that is to say, the rank of matrix $\mathbf{A}$ becomes larger and the selected key-frames become more. It is duly in compliance with the common sense that more key-frames should be extracted for the videos with higher complexity:

$$v(k) = \frac{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2}}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_h^2}}, \quad h = \min(m, n). \tag{15}$$

3.2. Locate Specific Key-Frames. Locate key-frames; namely, select linearly independent sub-set of matrix $\mathbf{A}$. The smallest correlation means the largest visual differences, and the video content differences between frames can be represented by inter-frame distance; therefore, this paper uses inter-frame distance to select the frames with the largest visual differences. First we calculate the histogram distance between each frame and its previous frame, as shown in (16), where $P_i(j)$

represents the gray value of jth pixel in the frame and n represents the total number of the frames within a shot; and then we extract the k frames with the largest distance as key-frames in each subshot:

$$\text{dist}(i, i+1) = \sqrt{\sum_{i=2}^{n} [P_i(j) - P_{i-1}(j)]^2}. \quad (16)$$

4. Experiments and Results

4.1. Selection of the Parameters. There are three parameters in the proposed algorithm that must be determined: the window length L, λ in (10), and α in Section 3. The selecting principle of parameter L is that when the sliding window is in the same subshot, there is little difference between within-class distance and between-class distance. Generally speaking, the faster the video content changes, the smaller the parameter L should be made. In our experiments, we find that $L = 6$ is proper for the shots with object motion and with fast camera motion and $L = 10$ is proper for other types of video shots.

The parameter λ determines the number of subshots. Increasing the value of λ will bring less subshots. Relying on our experiments, λ is specified as 0.5 which can ensure the accuracy of subshots segmentation. We have proved that the parameter $\alpha = 0.8$ is sufficient to preserve the most original information, which can satisfy the human perception very well in video summary. Users can adjust parameter values to control the quality and detail level according to actual circumstances and concrete perception.

4.2. Experimental Results. To determine the performance of the proposed method, various test videos are downloaded from the standard video library OPENVIDEO. Six extreme shots with different characteristics are selected in this section.

The first video, *hcil_2002*, is a shot with little change, in which a person is making a speech. As all F-values are smaller than 1.0, the shot does not need subshot segmentation. As shown in Figure 3, one key-frame is extracted, and it is enough to represent the original content.

The second video, *ROAD*, is a shot with fast camera movement, in which the camera coupled with the car moves forward rapidly, swerves and films roadside trees and a house, and lastly drives on a new road. As shown in Figure 4, the shot is divided into corresponding five subshots to describe the semantic progress, and the extracted key-frames do not miss the main visual information.

The third video is a shot with object movement, in which a man in white comes to a corner and waits for another man's arrival and, after a short talk, returns by his original route. According to the above semantic structure, the video shot is divided into two subshots, with the results shown in Figure 5. By the selected key-frames, viewers can correctly infer the video content.

The fourth video is a shot with both object movement and camera movement, in which a girl comes from afar, suddenly stops, then looks around, and finally runs in the opposite direction. When the girl looks around, her face is in a close-up. As shown in Figure 6, the extracted key-frames provide a good summary of the original shot.

Figure 3: Key-frames extracted from video *hcil_2002*.

Except for object and camera motion, artificial editing effects can also give rise to video content changes. The fifth video is a shot with special effects, in which many ordered plates gradually come together into a stack and then disappear suddenly. The extraction results are shown in Figure 7, from which it is apparent that the extracted key-frames can reproduce the process.

The last video is a shot with scrolling captions, in which a yacht is heading from shore out to sea, and suddenly a motorboat comes up fast from behind and gradually moves out of sight with the yacht. Besides, captions of large areas rapidly glide on the screen all the way. The extracted key-frames are displayed in Figure 8. It is obvious that the effect is not ideal.

Table 1 gives more information about the results described above. The third column shows the number of subshots labeled manually as the baseline.

5. Evaluation and Analysis

Due to the absence of well-defined objective criteria [21], some subjective evaluation schemes are mentioned to attempt to judge the perception of users towards video summary. The most common is mean opinion score (MOS) criterion [9, 16]. This criterion asks three users to rate the quality of each summary after watching the full video and the corresponding summary. Because current evaluation schemes are only for video summary, this paper advances three subjective evaluation indicators to fit our approach. Moreover, there are no benchmarking or ground truth results for key-frame determination algorithms so far; we do not perform any comparison between the proposed algorithm and others.

5.1. Evaluation Indicators

(1) Structure. This is essentially segmentation accuracy. The segmentation in this research is based on semantic meaning, which is determined by subjective criteria. Therefore, each original video was first divided up artificially based on perceived semantic structure, and then this baseline is compared with the experimental results.

Table 1: Extraction results for different videos.

Video name	Total frames	Subshots number (manually)	Subshots number (automatically)	Number of key frames	Video characteristics
hcil_2002	329	1	1	1	Little change
ROAD	84	5	5	6	Fast camera motion
vipmen	283	2	2	4	Object motion
Sassy girl	107	4	4	6	Both camera and object motion
Broken	76	3	4	4	Special effects
BOR10_013	328	3	4	6	Scrolling captions

Subshot 1 Subshot 2 Subshot 3

Subshot 4 Subshot 5

Figure 4: Key-frames extracted from video *ROAD*.

Subshot 1

Subshot 2

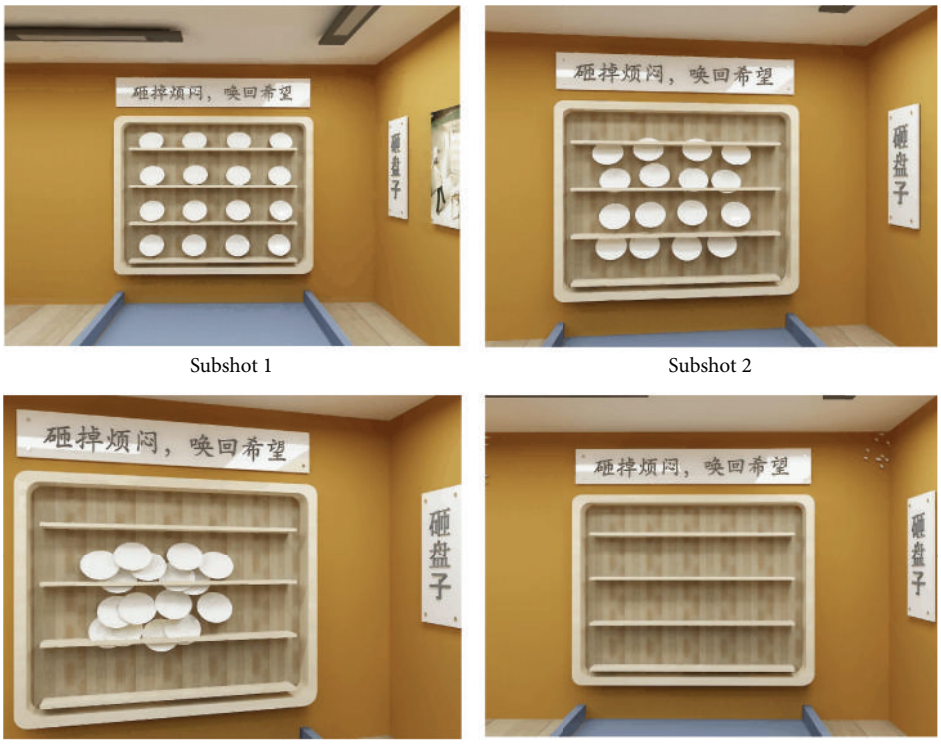
Figure 5: Key-frames extracted from *vipmen*.

Subshot 1 Subshot 2

Subshot 3

Subshot 4

Figure 6: Key-frames extracted from *Sassy girl*.

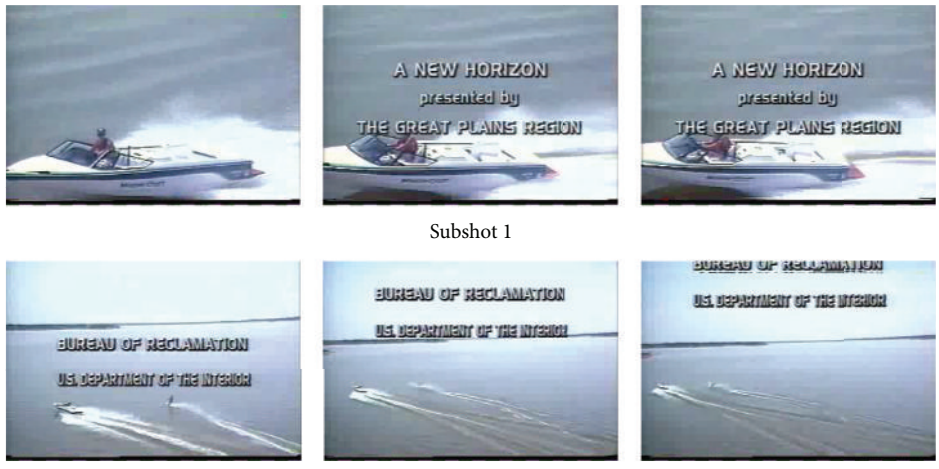


Subshot 1 Subshot 2

Subshot 3 Subshot 4

Figure 7: Key-frames extracted from *Broken*.

It can be seen that, by analysis of Table 1 and Figures 3–8, except the subshot segmentation in the fifth video and the sixth video, other segmentations are in agreement with the manual segmentations. For Figure 7, the second image and the third image belong to the same subshot which describes the process of plates coming together. However, the proposed method generates one more subshot than the manual segmentation, which is called oversegmentation. The reason for oversegmentation is that there are great visual differences in the gathering process, even though these frames possess



Subshot 1

Subshot 2 Subshot 3 Subshot 4

Figure 8: Key-frames extracted from *BOR10_013*.

Table 2: Assessment outcome for different types of videos.

Video type	Structure	Continuity	Repetition
Lecture	0.91	0.93	0.95
News	0.92	0.90	0.94
Documentary	0.90	0.92	0.91
Entertainment	0.87	0.91	0.84

the same semantic meaning. For Figure 8, the third subshot and the fourth subshot should be merged into the same one in the perspective of semantic meaning. The oversegmentation in Figure 8 is caused by the scrolling captions; yet our method mistakes it for significant video content change.

By the analysis above, subshot segmentation based on video content and subshot segmentation based on semantic meaning are not fully identical. Fortunately, in most cases, video content and semantic meaning are basically identical. Therefore, the method in this paper can carry out subshot segmentation based on semantic structure, which can detect both temporal and semantic independence between the frames.

(2) Continuity. The extracted key-frames must be as continuous as possible. A summary with many jumps is unlikely to be attractive to users. For visual continuity, some intermediate frames should be appended, even though they only play a part in connecting visual impressions and do not include important video content. In Figure 7, there is a visual discontinuity between the third and fourth frames. The reason is that the plates broke up instantly under the action of special effects; thus it is very difficult to detect the intermediate frames. To detect the intermediate frames, users can change the parameter α to get more details in the key-frame extraction of the fourth subshot.

(3) Repetition. While ensuring the presence of important information and visual continuity, the proposed algorithm attempts to eliminate redundancy and repetitive frames with the same semantic element. As shown in Figure 8, the first three key-frames and the last two frames exhibit visual and semantic redundancy. The redundancy in the first three frames is caused by the scrolling captions and that in the last two frames is caused by subshots oversegmentation. Having the mentioned above, the oversegmentation is also caused by the scrolling captions. This is an indication that our method is sensitive to scrolling captions. Beyond these, other experimental results show that this algorithm can control redundancy excellently.

5.2. Overall Evaluations. To verify the robustness of the proposed algorithm, 100 video shots are clipped from four different types of videos: lecture, news, documentary, and entertainment. We refer to the mean opinion score (MOS) criterion and recruit twenty testers to give subjective scores to the key-frame extraction results. First, every tester is given five shots, which covered the four types of videos. After viewing the extraction results and the original videos, the testers are asked to assign scores to the extraction results in terms of structure, continuity, and repetition. A scale (0.0–1.0) is used for scoring, where 0 represents great dissatisfaction and 1.0 represents great satisfaction. The scores from each tester are averaged to yield the assessment outcome shown in Table 2.

Table 2 shows that the method proposed in this paper is able to detect dependencies between subshots, eliminate repetitive frames with small alterations, and extract key-frames with maximum visual information. Therefore, the proposed method could be considered a good algorithm for both video summary and video index.

6. Conclusion

This paper is the first study to fit both video summary and video index; the new method achieves good semantic

structure, good visual continuity, and low redundancy, not only can provide a video summary which is consistent with human perception but also can provide an index for further video operations and analysis.

Note that because of the complexity and diversity of videos, the proposed algorithm cannot be proved to be capable of demonstrating good and stable performance on all videos. More experiments should be done to confirm the area of applicability of the algorithm. In addition, the deep reason of oversegmentation is overly simplified feature selection; future researches should also concentrate on composite feature selection to resist scrolling captions.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the Natural Science Foundation of Shanxi Province, China (Grant no. 2011011012-2), and Taiyuan Special Fund for Science and Technology Talents (Grant no. 120247-28). Acknowledgments are due to Cao Changqing, Duan Hao and Yang Qian for their collaboration in the realization of field experiments. The authors would also like to thank the reviewers for their time and their valuable comments.

References

[1] J. Son, H. Lee, and H. Oh, "PVR: a novel PVR scheme for content protection," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 173–177, 2011.

[2] B. T. Truong and S. Venkatesh, "Video abstraction: a systematic review and classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, no. 1, article no. 3, pp. 1–37, 2007.

[3] A. G. Money and H. Agius, "Video summarisation: a conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[4] R. M. Jiang and A. H. Sadka, "Advances in video summarization and skimming," in *Recent Advances in Multimedia Signal Processing and Communications*, M. Grgic, K. Delac, and M. Ghanbari, Eds., vol. 231, pp. 27–50, Springer, Berlin, Germany, 1st edition, 2009.

[5] W. B. Jiang, H. Jin, R. Zheng, and D. Q. Zou, "Key frame extraction based on scale invariant feature transform," in *Proceedings of 3rd International Conference on Multimedia and Ubiquitous Engineering*, pp. 45–48, QingDao, China, June 2009.

[6] N. Ejaz, T. Bin Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *Journal of Visual Communication and Image Representation*, vol. 23, pp. 1031–1040, 2012.

[7] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings of the International Conference on Image Processing (ICIP '98)*, pp. 866–870, Chicago, Ill, USA, October 1998.

[8] S. E. F. de Avila, A. P. B. Lopes, A. Da Luz Jr., and A. de Albuquerque Araújo, "VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[9] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 47–69, 2010.

[10] S. K. Kuanar, R. Panda, and A. S. Chowdhur, "Video key frame extraction through dynamic Delaunay clustering with a structural constraint," *Journal of Visual Communication and Image Representation*, vol. 24, pp. 1212–1227, 2013.

[11] T. Liu, H.-J. Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 10, pp. 1006–1013, 2003.

[12] Y. Ma, Y. Chang, and H. Yuan, "Key-frame extraction based on motion acceleration," *Optical Engineering*, vol. 47, no. 9, Article ID 090501, 2008.

[13] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequances," in *Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '09)*, pp. 25–28, London, UK, May 2009.

[14] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: from humans to computers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 2, pp. 289–301, 2009.

[15] J.-L. Lai and Y. Yi, "Key frame extraction based on visual attention model," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 114–125, 2012.

[16] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, vol. 28, pp. 34–44, 2013.

[17] H.-C. Shih, "A novel attention-based key-frame determination method," *IEEE Transactions on Broadcasting*, vol. 59, pp. 556–562, 2013.

[18] G. Paschos, "Perceptually uniform color spaces for color texture analysis: an empirical evaluation," *IEEE Transactions on Image Processing*, vol. 10, no. 6, pp. 932–937, 2001.

[19] X. Zhang, *Pattern Recognition*, Tsinghua University Press, Beijing, China, 3rd edition, 2010.

[20] X. Zhang, *Matrix Analysis and Applications*, Tsinghua University Press, Beijing, China, 2nd edition, 2004.

[21] M. Slaney, "Precision-recall is wrong for multimedia," *IEEE Multimedia*, vol. 18, no. 3, pp. 4–7, 2011.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 841608, 5 pages
http://dx.doi.org/10.1155/2014/841608

Research Article

Interlayer Simplified Depth Coding for Quality Scalability on 3D High Efficiency Video Coding

Mengmeng Zhang,[1] Hongyun Lu,[1] and Huihui Bai[2]

[1] *North China University of Technology, No. 5 Jin Yuanzhaung Road, Shijingshan District, Beijing 100144, China*
[2] *Beijing Jiaotong University, No. 3 Shangyuancun, Haidian District, Beijng 100044, China*

Correspondence should be addressed to Mengmeng Zhang; zhang_mengmengbbb@163.com

Received 23 December 2013; Accepted 22 January 2014; Published 16 March 2014

Academic Editors: X. Meng, Z. Zhou, and X. Zhu

A quality scalable extension design is proposed for the upcoming 3D video on the emerging standard for High Efficiency Video Coding (HEVC). A novel interlayer simplified depth coding (SDC) prediction tool is added to reduce the amount of bits for depth maps representation by exploiting the correlation between coding layers. To further improve the coding performance, the coded prediction quadtree and texture data from corresponding SDC-coded blocks in the base layer can be used in interlayer simplified depth coding. In the proposed design, the multiloop decoder solution is also extended into the proposed scalable scenario for texture views and depth maps, and will be achieved by the interlayer texture prediction method. The experimental results indicate that the average Bjøntegaard Delta bitrate decrease of 54.4% can be gained in interlayer simplified depth coding prediction tool on multiloop decoder solution compared with simulcast. Consequently, significant rate savings confirm that the proposed method achieves better performance.

1. Introduction

In January 2013, the Joint Collaborative Team on Video Coding has finalized a final draft about the next generation video standard, that is, High Efficiency Video Coding (HEVC) [1]. Scalable High Efficiency Video Coding (SHVC) and 3D video coding are being formulated as the extensions of HEVC [2]. Given the special performance of HEVC in delivering the target resolution and frame rates [3], 3D video coding and SHVC are also applied to a variety of consumer domains with their different application opportunities. Currently, 3D video coding and scalable video coding are entering broad and possibly sustainable mass markets. Magnificent and excellent three-dimensional scenes in 3D TV can be provided with the mature 3D video coding technology. Scalable video coding is applied to cope with the heterogeneity of networks and devices used in the video service environment. With the rapid evolution in theory and techniques, the technology of scalable video coding based on the 3D video coding may be applied to 3D TV or mobile terminal in the future. Therefore, the investigation of a scalable 3D video coding scenario is important and necessary.

Generally, scalable video coding is a highly attractive solution to the problems caused by the characteristics of modern video transmission systems. The scalable video coding method can be used to achieve the adaptation of a bitrate with features such as temporal and spatial scalabilities [4]. The quality scalability could be treated as a special case of spatial scalability with the same resolution in different layers [5]. This paper proposes a quality scalable 3D video coding that is equipped with both quality scalability and 3D visualization and modified based on the 3D video coding to reduce complexity. Different layers of several simultaneous views are coded into the bitstream for quality scalable 3D video coding.

Depth-Image-Based Rendering (DIBR) is widely used for view synthesis in 3D video coding. Thus, the texture video and the associated depth map are required to be scaled simultaneously [6]. DIBR notes that the execution of proposed scalable methods can be approached in terms of

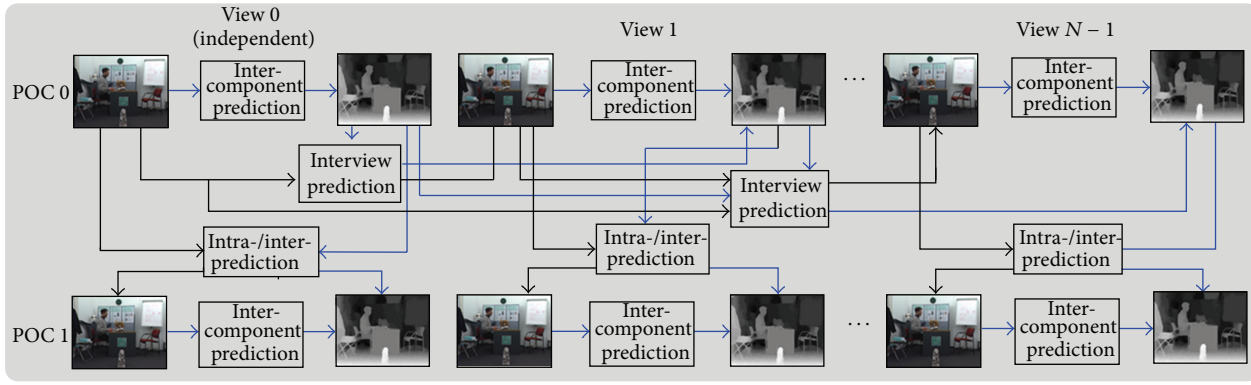


Figure 1: Basic encoder structure with interview and intercomponent prediction.

two considerations in this paper. First, an interlayer texture Prediction mechanism will be employed to eliminate the redundancy of the different layers on multiloop decoding solution that was proposed in Van Wallendael et al. [7]. The inter-layer texture prediction mechanism will be utilized in both texture videos and depth maps. This method achieves higher compression efficiency, yet it maintains backwards compatibility with multiple views coded by HEVC. Second, an extraordinary interlayer prediction tool called an interlayer simplified depth coding (interlayer SDC) is used to reduce interlayer redundancy. As depth maps exhibit unique characteristics such as piecewise smooth regions bounded by sharp edges at depth discontinuities [8], new coding tools are required to approximate these signal characteristics [9]. These coding tools include simplified depth coding (SDC) and depth modeling modes (DMM) [10]. Moreover, all the results are tested in an all intraconfigurations as every frame includes three views (view 0 is predicted in I slice, and view 1 and view 2 are predicted in P slice), and the SDC is only chosen in intraframe.

The rest of this paper is organized as follows. Section 2 introduces the 3D high efficiency video coding. The details of the proposed interlayer SDC tool are presented in Section 3. Section 4 describes the test scenarios and presents the analysis results. Section 5 concludes the paper.

2. 3D High Efficiency Video Coding

As one of the extensions of HEVC, the upcoming 3D video coding makes use of the efficient single-view coding tools used in HEVC. HEVC is the latest video coding standard developed by a joint effort between ISO/IEC and ITU-T and succeeding H.264/AVC. This design still follows a traditional hybrid coding approach [9], such as interprediction based on the motion compensated, interprediction residuals of the two-dimensional transform, and quadtree. 3D HEVC is achieved by coding each video view and associated depth map component using a 2D video coding structure that is based on the technology of HEVC. In order to provide backward compatibility with 2D video services, the independent view is coded using a fully HEVC compliant codec [11]. Except that intra-/interframe prediction is still exiting in 3D HEVC, interview prediction for views and intercomponent prediction for views and maps are added into the 3D extension. The prediction structure is depicted in Figure 1 in detail. The blue arrows denote the prediction for depth maps, and the black shows the prediction between the views.

3. Proposed Quality Scalability on 3D Video Coding

3.1. The Framework of Scalability on 3D Video Coding. The proposed quality scalable scheme employs interlayer simplified depth coding and interlayer texture prediction to remove interlayer redundancy based on the multiloop decoder structure. The multiloop decoding solution is integrated into the proposed scenario as a whole framework of the scalable 3D video coding. The interlayer texture prediction is considered as a basic prediction mode that the base layer (BL) needs to be decoded entirely before the enhancement layers are reconstructed. The transform and quantization processes of interlayer texture prediction predict that CUs are the same as an intrapredicted CU on the QP of the enhancement layer, in which discrete sine transform and discrete cosine transform are applied to the different types of TUs. Figure 2 depicts the block diagram of the proposed interlayer prediction in encoder. The color-marked parts represent the prediction signalling mechanisms for depth maps in the same view in enhancement layer. The red-marked arrows show the prediction process with transform and quantization, and the blue-marked arrows denote the prediction process without transform and quantization. All modes, including traditional prediction modes and interlayer prediction modes, could be chosen as the best prediction mode by rate distortion optimization (RDO) [12] for texture views and virtual synthesis optimization (VSO) [13] for depth maps. Details of the proposed scalable algorithms are described in the following.

3.2. Feasibility Analysis of Interlayer Simplified Depth Coding. As an alternative intracoding mode, the SDC approach is added into the intracoded block, and the prediction mode is still INTRA for an SDC-coded block with additional SDC_Flag signals in depth map coding [14]. The information of base layer SDC-blocks can be directly used as the reference

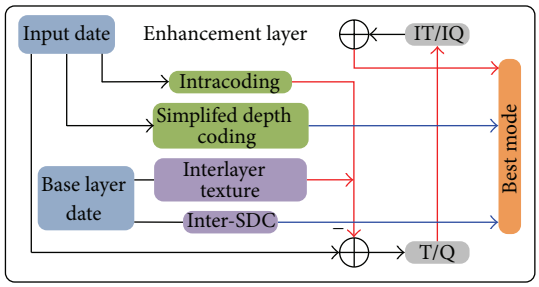


Figure 2: High-level block diagram of the encoder.

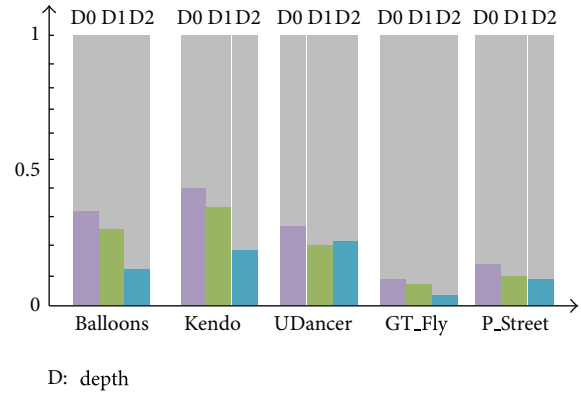


Figure 3: Percentages of SDC-coded block in 300 I-frames.

for enhancement layer because the clear textural feature leads to many distortionless SDC-blocks, and the dynamic quantization parameter (QP) does not affect the distortion of SDC-blocks as a result of no transform and quantization. The amount of available SDC-blocks becomes an issue of crucial importance as only the distortionless SDC-blocks could be the reused for interlayer SDC. Thus, more attention is paid to the amount of SDC-blocks that could be applied to interlayer SDC. A statistical experiment was then performed. The percentages of distortionless SDC-blocks in all the blocks were calculated in 300 I-frames (Figure 3). Figure 3 shows that the number of distortionless SDC-blocks occupies a large proportion in all depth blocks in the experiment.

3.3. Interlayer Simplified Depth Coding Prediction Tool. The interlayer SDC tool is sufficient in generating a good prediction signal and eliminates ringing artifacts for SDC-coded blocks. Instead of coding quantized transform coefficients, the following three parts of information are coded in SDC-coded blocks. The first part is the type of segmentation/prediction of the current block with possible values of DC, DMM, or Planar. The second part is the additional prediction information when the DMM mode is selected as the type of segmentation. The third part is the residual value for each resulting segment, which is present in the original, uncompressed depth map using a depth lookup table (DLT). Initial analysis shows that the DLT is constructed by analyzing a certain number of frames of the input sequence before coding. The residual is obtained from the difference of the prediction index and the original index according to the DLT. The structure of the algorithm is described as follows (Figure 4).

The interlayer SDC utilizes the decompressed date from the collocated distortionless SDC-coded CUs due to the characteristics of distortionless SDC-coded CUs in the base layer. This paper considers the interlayer SDC method as an extension of the intracoding mode, in which the type of segmentation/prediction, additional prediction information, and residual value of the index are obtained from the corresponding base layer CU. The prediction image of the current enhancement layer is rebuilt based on the type of segmentation/prediction and additional prediction information, which have been selected by the traditional SDC in the base layer. The reconstructed image is derived from the residual value of the index and the index prediction image. No

Table 1: Identifiers on interlayer SDC and traditional SDC.

Identifiers of the modes		Specified mode
Inter_SDC_Flag	SDC_Flag	
1	0	Interlayer SDC
0	1	Traditional SDC
0	0	Intracoded

transform and quantization processes occur in the interlayer simplified depth coding prediction method, in which only the information of the SDC in the base layer needs to be decoded for the enhancement layer. Thus, no additional data are transmitted for the enhancement layer except a flag of the interlayer SDC mode in the proposed interlayer SDC tools. When the "Inter_SDC_Flag" that is signaled in the enhancement layer is set to be true, the current block is decoded in an interlayer SDC decoder. The "SDC_Flag" is the coding identifier of the traditional SDC for 3D video coding. The "SDC_Flag" and "Inter_SDC_Flag" are used to codetermine the decoding mode in the enhancement layer. Table 1 shows the detailed implications of identifiers.

Moreover, the DLT in the base layer is used in the interlayer SDC. The RD cost of interlayer texture prediction and the interlayer SDC method is calculated for the enhancement layer CUs in addition to the RD selection procedure used in unmodified 3D video encoder. The optimal prediction mode is selected to minimize the cost function:

$$\text{Cost}_{\text{mode}} = \min \left\{ \begin{array}{c} \text{RD (Inter-layer texture prediction)} \\ \text{RD (Intra, Traditional SDC)} \\ \text{RD (Inter-layer SDC)} \end{array} \right\}, \quad (1)$$

where $\text{Cost}_{\text{mode}}$ is the minimalized rate-distortion cost of the current mode. The RD (·) is the rate-distortion cost of every mode by the RDO (VSO).

4. Results and Analysis

The scalable 3D video coding theme is implemented based on HTM 6.0 [15]. The common test configurations are defined in D1100 [16]. Two layers (one base layer and one enhancement

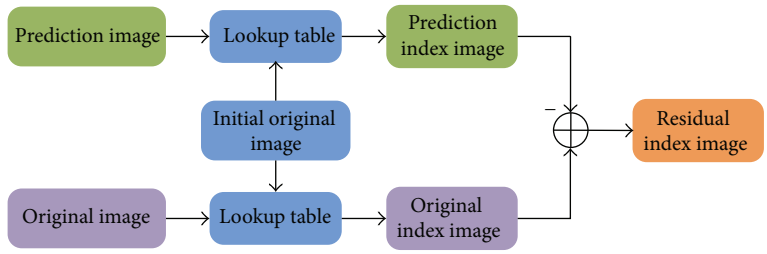


Figure 4: The Structure of the simplified depth coding algorithm.

Table 2: BD rate of inter-SDC prediction in multiloop to simulcast encoding.

Sequence	BD rate (%) of 2 synthesized views			BD rate (%) of 3 synthesized views		
	Y	*U*	*V*	*Y*	*U*	*V*
Kendo	55.7	63.6	56.5	53.5	70.6	55.6
Balloons	55.3	55.4	55.4	54.6	54.7	54.7
UDancer	63.3	60.2	59.9	58.5	59.3	58.1
GT_Fly	56.8	53.5	54.2	54.7	48.4	53.2
P_Street	44.9	47.9	38.8	50.9	51.3	47.8
Average	55.2	56.1	53.0	54.4	56.9	53.9

layer) and three views are simultaneously evaluated in all I-frames. The base layer and enhancement layer are encoded in different QPs with a spatial ratio of 1 : 1. The quantization parameters of base and enhancement texture views are Q1 and Q2. The common conditions specify the biggish Q1 (30, 35, 40, and 45), and the delta QP between two layers is 5. The QPs of depth maps change according to the QP of the associated texture view in 3D video coding. Two experimental methods of displays intuitively describe the experimental results, namely, the PSNR-Bitrate graph and the BD rate table. Comparing the results, the bitrate contains all the layers, and the PSNR is the highest enhancement layer with Q2. Moreover, all the experimental results come from the synthetic view after scaling the texture views and depth maps. Three schemes were realized in this experiment for comparison, namely, single-loop, simulcast coding, and inter-SDC prediction in multiloop solution. The simulcast solution that two layers were both coded in 3D video without the interlayer prediction will be used as the anchor to evaluate our proposed scalable scenario.

The simulcast and the scalable scenarios in interlayer SDC prediction with interlayer texture prediction are compared in Table 2. The results indicate a 55.2% Y-BD-rate decrease for 2 synthesized views, whereas a 54.4% Y-BD-rate decrease is noted for 3 synthesized views. Thus, the overall experimental results of Table 2 show that the proposed interlayer SDC is an inevitable tool for improving compressive performance. Figures 5 and 6 show the efficiency of 3 synthetic views coding, which are present in the PSNR-Bitrate graph in 1024 × 768 balloons and 1920 × 1088 Undo_Dancer sequences. The results also imply that the compression efficiency of Kendo

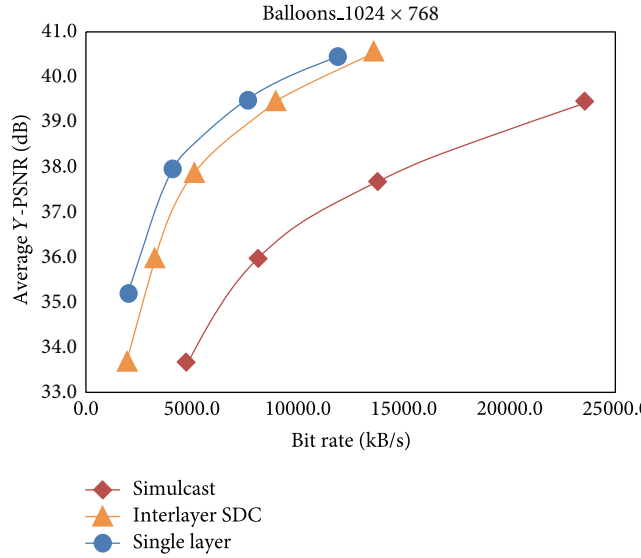


Figure 5: Performance comparison for balloons.

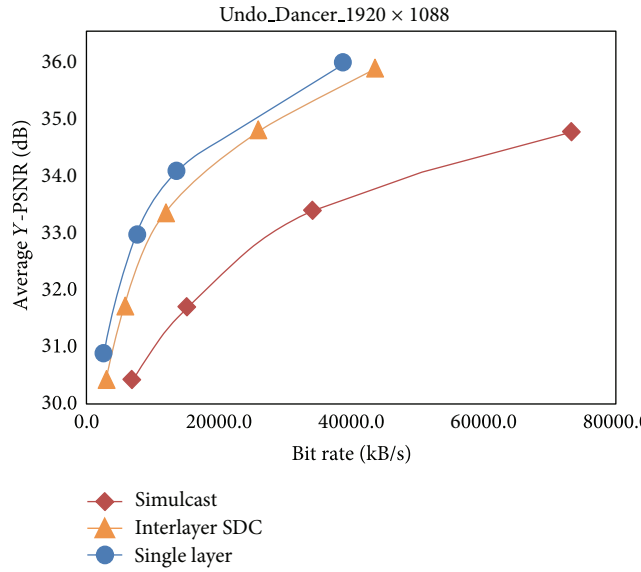


Figure 6: Performance comparison for Undo_Dancer.

and Balloons is better than that of Undo_Dancer and GT_Fly. These results prove that the coding efficiency depends on the specific sequences, and the numbers of distortionless base layer SDC-CUs are an important factor in influencing the performance of interlayer SDC.

5. Conclusion

We presented a scalable 3D video coding theme on the emerging HEVC, which supports two interlayer prediction methods on a multiloop decoder structure. The interlayer texture prediction method simultaneously exploits the interlayer correlation for texture views and depth maps. The interlayer SDC prediction tool achieves significant bitrate decrease and complexity reducing for the depth maps. Experimental results demonstrate the effectiveness of our proposed scenario. Improvement of the coding performance of scalable 3D video coding theme can be examined in future research. More interlayer prediction methods will be proposed to accomplish the scalability on 3D video coding.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the Natural National Science Foundation of China (no. 61370111, no. 61103113, and no. 61272051) and Beijing Municipal Education Commission General Program (KM201310009004).

References

[1] B. Bross, "High Efficiency Video Coding (HEVC) text specification draft 10 (for FDIS & Consent)," in *Proceedings of the 12th Meeting of Joint Collaborative Team on Video Coding (JCT-VC '13)*, JCTVC-L1003_v23, Geneva, Switzerland, January 2013.

[2] J. Chen, *SHVC Test Model 1 (SHM 1)*, JCTVC-L1007, Qualcomm, Geneva, Switzerland, 2013.

[3] G. J. Sullivan, J. Ohm, J. W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[5] Z. Shi and X. Sun, "Spatially scalable video coding for HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1813–1826, 2012.

[6] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand, "An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution," in *Proceedings of the Picture Coding Symposium (PCS '09)*, pp. 1–4, May 2009.

[7] G. van Wallendael, J. de Cock, R. Van de Walle, and M. Mrak, "Multi-loop quality scalable based on high efficiency video coding," in *Proceedings of the Picture Coding Symposium (PCS '12)*, pp. 445–448, 2012.

[8] S. Tao, Y. Chen, M. M. Hannuksela, Y. K. Wang, M. Gabbouj, and H. Li, "Joint texture and depth map video coding based on the scalable extension of H.264/AVC," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '09)*, pp. 2353–2356, May 2009.

[9] G. Tech, "3D-HEVC test model 1," in *Proceedings of the Meeting of Joint Collaborative Team on 3D Video Coding (JCT-3V '12)*, JCT3V-A1005_d0, Stockholm, Sweden, July 2012.

[10] M. Zhang, C. Zhao, J. Xu, and H. Bai, "A fast depth-map wedgelet partitioning scheme for intra prediction in 3D video codin," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '13)*, pp. 2852–2855, May 2013.

[11] K. Müller, H. Schwarz, and D. Marpe, "3D high-efficiency video coding for multi-view video and depth data," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3366–3378, 2013.

[12] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for: video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[13] B. Oh, J. Lee, D. Park et al., "3D-CE8.h results on view synthesis optimization by Samsung, HHI and LG-PKU," A0093, Stockholm, Sweden, July 2012.

[14] F. Jäger, "3D-CE6.h results on simplified depth coding with an optional depth lookup table," JCT3V-B0036, Shanghai, China, 2012.

[15] G. Tech, "3D-HEVC test model 3," JCT3V-C1005_d0, Geneva, Switzerland, 2013.

[16] D. Rusanovskyy, K. Müller, and A. Vetro, "Common test condition of 3DV core experiments," JCD3V-D1100, Incheon, Republic of Korea, 2013.

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 924090, 11 pages
http://dx.doi.org/10.1155/2014/924090

Research Article

Multiview Discriminative Geometry Preserving Projection for Image Classification

Ziqiang Wang, Xia Sun, Lijun Sun, and Yuchun Huang

School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China

Correspondence should be addressed to Ziqiang Wang; zi-qiang-wang@hotmail.com

Received 19 December 2013; Accepted 22 January 2014; Published 9 March 2014

Academic Editors: X. Meng, Z. Zhou, and X. Zhu

In many image classification applications, it is common to extract multiple visual features from different views to describe an image. Since different visual features have their own specific statistical properties and discriminative powers for image classification, the conventional solution for multiple view data is to concatenate these feature vectors as a new feature vector. However, this simple concatenation strategy not only ignores the complementary nature of different views, but also ends up with "curse of dimensionality." To address this problem, we propose a novel multiview subspace learning algorithm in this paper, named multiview discriminative geometry preserving projection (MDGPP) for feature extraction and classification. MDGPP can not only preserve the intraclass geometry and interclass discrimination information under a single view, but also explore the complementary property of different views to obtain a low-dimensional optimal consensus embedding by using an alternating-optimization-based iterative algorithm. Experimental results on face recognition and facial expression recognition demonstrate the effectiveness of the proposed algorithm.

1. Introduction

Many computer vision and pattern recognition applications involve processing data in a high-dimensional space. Directly operating on such high-dimensional data is difficult due to the so-called "curse of dimensionality." For computational time, storage, and classification performance considerations, dimensionality reduction (DR) techniques provide a means to solve this problem by generating a succinct and representative low-dimensional subspace of the original high-dimensional data space. Over the past two decades, many dimensionality reduction algorithms have been proposed and successfully applied to face recognition [1]. The most representative ones are principal component analysis (PCA) and linear discriminant analysis (LDA) [2].

PCA is an unsupervised dimensionality reduction method, which aims to project the high-dimensional data into a low-dimensional subspace spanned by the leading eigenvectors of a covariance matrix. LDA is supervised and its goal is to pursue a low-dimensional subspace by maximizing the ratio of between-class variance to within-class variance. Due to the utilization of label information, LDA usually outperforms PCA for classification tasks when sufficient labeled training data are available. While these two algorithms have attained reasonable good performance in pattern classification, they may fail to discover a highly nonlinear submanifold embedded in the high-dimensional ambient space as they seek only a compact Euclidean subspace for data representation and classification [3].

Recently, there has been considerable interest in manifold learning algorithms for dimensionality reduction and feature extraction. The basic consideration of these algorithms is that the high-dimensional data may lie on an intrinsic nonlinear low-dimensional manifold. In order to detect the underlying manifold structure, nonlinear dimensionality reduction algorithms such as ISOMAP [4], locally linear embedding (LLE) [5], and Laplacian eigenmap (LE) [6] have been proposed. All of these algorithms are defined only on the training data, and the issue of how to map new testing data remains difficult. Therefore, they cannot be applied to classification problem directly. To overcome the above so-called out-of-sample problem, He and Niyogi [7] developed the locality preserving projection (LPP), in which the linear projection function is adopted for mapping new data samples. As LPP is originally unsupervised, some recent attempts have exploited the discriminant information and derived many discriminant

manifold learning algorithms to enhance the classification performance. The representative algorithms include local discriminant embedding (LDE) [8], locality sensitive discriminant analysis (LSDA) [9], margin Fisher analysis (MFA) [10], local Fisher discriminant analysis (LFDA) [11], and discriminative geometry preserving projection (DGPP) [12]. Despite having different assumptions, all these algorithms can be unified into a general graph embedding framework (GEF) [10] with different constraints. While these algorithms have utilized both local geometry and the discriminative information for dimensionality reduction and achieved reasonably good performance in different pattern classification tasks, they assume that the data are represented in a single vector. They can be regarded as single-view-based methods and thus cannot handle data described by multiview features directly. In many practical pattern classification applications, different views (visual features) have their own specific statistical properties, and each view represents the data partially. To address this problem, the traditional solution for multiple view data is to simply concatenate vectors of different views into a new long vector and then apply dimensionality reduction algorithms directly on the concatenated vector. However, this concatenation ignores the diversity of multiple views and thus cannot explore the complementary nature and specific statistical properties of different views. Recent studies have provided convincing evidence of this fact [13–15]. Hence, it is more reasonable to assign different weights to different views (features) for feature extraction and classification. In computer vision and machine learning research, many works have shown that leveraging the complementary nature of the multiple views can better represent the data for feature extraction and classification [13–15]. Therefore, an efficient manifold learning algorithm that can cope with multiview data and place proper weights on different views is of great interest and significance.

Motivated by the above observations and reasons, we propose unifying different views under a discriminant manifold learning framework called multiview discriminative geometry preserving projection (MDGPP). Under each view, we can implement the discrimination and local geometry preservations as those used in discriminative geometry preserving projection (DGPP) [12]. Unifying different views in such a multiview discriminant manifold learning framework is meaningful, since data with different features can be appropriately integrated to further improve the classification performance. Specifically, we first implement the discrimination preservation by maximizing the average weighted pairwise distance between samples in different classes and simultaneously minimizing the average weighted pairwise distance between samples in the same class. Meanwhile, the local geometry preservation is implemented by minimizing the reconstruction error of samples in the same class. Then, we learn a low-dimensional feature subspace by utilizing both intraclass geometry and interclass discrimination information, such that the complementary nature of different views (features) can be fully exploited when classification is performed in the derived feature subspace. Experimental results on face recognition and facial expression recognition are presented to demonstrate the effectiveness of the proposed algorithm.

The remainder of the paper is organized as follows. Section 2 reviews the related works. Section 3 presents the details of the proposed MDGPP algorithm. Experimental results on face recognition are presented in Section 4, and the concluding remarks are provided in Section 5.

2. Related Works

Multiview learning is one important topic in the machine learning and pattern recognition communities. In such a setting, view weight information is introduced to measure the importance of different features in characterizing data, and different weights reflect different contribution to the learning process. The aim of multiview learning is to exploit more complementary information of different views rather than only a single view to further improve the learning performance. The traditional solution for multiview data is to concatenate all features into one vector and then conduct machine learning for such feature space. However, this solution is not optimal as these features usually have different physical properties. Simply concatenating them will ignore the complementary nature and specific statistical properties of different views, and thus causing performance degradation. In addition, this simple concatenation will end up with the curse of dimensionality problem for the subsequent learning task.

In order to perform multiview learning, much effort has been focused on multiview metric learning [14], multiview classification and retrieval [16], multiview clustering [15], and multiview semisupervised learning [17]. All these approaches demonstrated that the learning performance can be significantly enhanced if the complementary nature of different views is exploited and all views are appropriately integrated. It is very natural that multiview learning idea should also be considered in dimensionality reduction. However, most of the existing dimensionality reduction algorithms are designed only for single view data and cannot cope with multiview data directly. To address this problem, Long et al. [18] first proposed multiple view spectral embedding (MVSE) method. MVSE performs a dimensionality reduction process on each view independently, and then based on the obtained low-dimensionality representation, it constructs a common low-dimensional embedding that is close to each representation as much as possible. Although MVSE allow selecting different dimensionality reduction algorithms for each view, the original multiview data are invisible to the final learning process. Thus, MVSE cannot well explore the complementary information of different views. Xia et al. [19] proposed multiview spectral embedding (MSE) method to find low-dimensional and sufficiently smooth embedding based on the patch alignment framework [20]. However, MSE ignores the flexibility of allowing shared information between subset of different views owing to the global coordinate alignment process. To unify different views for dimensionality reduction under a probabilistics, Xie et al. [21] extended the stochastic

neighbor embedding (SNE) to its multiview version and proposed multiview stochastic neighbor embedding (MSNE). Although MSNE operates on a probabilistic framework, it is an unsupervised method and its classification abilities may be limited since the class label information is not used in the learning process. More recently, inspired by the recent advances of sparse coding technique, Han et al. [22] proposed spectral sparse multiview embedding (SSMVE) method to deal with dimensionality reduction for multiview data. Although SSMVE can impose sparsity constraint on the loading matrix of multiview dimensionality reduction, it is unsupervised and does not explicitly consider the manifold structure on which the high dimensional data possibly reside. In the next section, focusing on the manifold learning and pattern classification, we propose a novel multiview discriminative geometry preserving projection (MDGPP) for multiview dimensionality reduction, which explicitly considers the local manifold structure and discriminative information as well as the complementary characteristics of different views in high-dimensional data.

3. Multiview Discriminative Geometry Preserving Projection (MDGPP)

In this section, we propose a new manifold learning algorithm called multiview discriminative geometry preserving projection (MDGPP), which aims to find a unified low-dimensional and sufficiently smooth embedding over all views simultaneously. To better explain the algorithm details of the proposed MDGPP, we introduce some important notations used in the remainder of this paper. Capital letters such as X denote data matrices, and X_{ij} represents the (i, j) entry of X. Lower case letters such as x denote data vectors, and x_i represents the ith data element of x. Superscript (i) such as $X^{(i)}$ and $x^{(i)}$ represents data from the ith view, respectively. Based on these notations, MDGPP can be described as follows according to the DGPP framework [12].

Given a multiview data set with n data samples and each with m feature representations, that is, $X = \{X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}] \in R^{p_i \times n}\}_{i=1}^{m}$, wherein $X^{(i)}$ represents the feature matrix for the ith view, the aim of MDGPP is to find a projective matrix $W \in R^{p_i \times d}$ to map $X^{(i)}$ into a low-dimensional representation $Y^{(i)}$ through $Y^{(i)} = W^T X^{(i)}$, where d denotes the dimension of low-dimensional feature representation and satisfies $d < p_i$ $(1 \le i \le m)$. The workflow of MDGPP can be simply described as follows. First, MDGPP builds a part optimization for a sample on a single view by preserving both the intraclass geometry and interclass discrimination. Afterward, all parts of optimization from different views are unified as a whole via view weight coefficients. Then an alternating-optimization-based iterative algorithm is derived to obtain the optimal low-dimensional embedding from multiple views.

Given the ith view $X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}] \in R^{p_i \times n}$, MDGPP first makes an attempt to preserve discriminative information in the reduced low-dimensional space by maximizing the average weighted pairwise distance between samples in different classes and simultaneously minimizing the average weighted pairwise distance between samples in the same class on the ith view. Thus, we have

$$\begin{aligned}
&\arg\max_{y_j^{(i)}} \sum_{j,t=1}^{n} h_{jt}^{(i)} \left\| y_j^{(i)} - y_t^{(i)} \right\|^2 \\
&= \arg\max_{y_j^{(i)}} \sum_{j,t=1}^{n} h_{jt}^{(i)} \left\| y_j^{(i)} - y_t^{(i)} \right\|^2 \\
&= \arg\max_{y_j^{(i)}} \operatorname{Tr}\left(\sum_{j,t=1}^{n} h_{jt}^{(i)} \left(y_j^{(i)} - y_t^{(i)} \right) \left(y_j^{(i)} - y_t^{(i)} \right)^T \right) \\
&= \arg\max_{y_j^{(i)}} \operatorname{Tr}\left(\sum_{j,t=1}^{n} h_{jt}^{(i)} \left(y_j^{(i)} \left(y_j^{(i)} \right)^T - y_t^{(i)} \left(y_j^{(i)} \right)^T \right.\right. \\
&\qquad \left.\left. - y_j^{(i)} \left(y_t^{(i)} \right)^T + y_t^{(i)} \left(y_t^{(i)} \right)^T \right) \right) \\
&= \arg\max 2\operatorname{Tr}\left(Y^{(i)} D^{(i)} \left(Y^{(i)} \right)^T - Y^{(i)} H^{(i)} \left(Y^{(i)} \right)^T \right) \\
&= \arg\max 2\operatorname{Tr}\left(Y^{(i)} L^{(i)} \left(Y^{(i)} \right)^T \right) \\
&= \arg\max 2\operatorname{Tr}\left(U^T X^{(i)} L^{(i)} \left(X^{(i)} \right)^T U \right),
\end{aligned} \tag{1}$$

where $\operatorname{Tr}(\,)$ denotes the trace operation of matrix, $L^{(i)} (= D^{(i)} - H^{(i)})$ is the graph Laplacian on the ith view, $D^{(i)}$ is a diagonal matrix with its element $D_{jj}^{(i)} = \sum_t H_{jt}^{(i)}$ on the ith view, and $H^{(i)} = [h_{jt}^{(i)}]_{j,t=1}^{n}$ is the weighting matrix which encodes both the distance weighting information and the class label information on the ith view

$$h_{jt}^{(i)} = \begin{cases} c_{jt} \left(\dfrac{1}{n} - \dfrac{1}{n_l} \right), & \text{if } l_j = l_t = l \\ \dfrac{1}{n}, & \text{if } l_j \ne l_t, \end{cases} \tag{2}$$

where in l_j is the class label of sample $x_j^{(i)}$, n_l is the number of samples belonging to the lth class, and c_{jt} is set as $\exp(-\|x_j^{(i)} - x_t^{(i)}\|/\delta^2)$ according to LPP [7] for locality preservation.

Second, we try to implement the local geometry preservation by assuming that each sample $x_j^{(i)}$ can be linearly reconstructed by the samples $x_t^{(i)}$ which share the same class label with $x_j^{(i)}$ on the ith view. Thus, we can obtain the reconstruction coefficient $w_{jt}^{(i)}$ by minimizing the reconstruction error $\sum_{j=1}^{n} \|\varepsilon_j\|^2$ on the ith view; that is,

$$\begin{aligned}
&\arg\min_{w_{jt}} \sum_{j=1}^{n} \left\| \varepsilon_j \right\|^2 \\
&\qquad = \arg\min_{w_{jt}} \sum_{j=1}^{n} \left\| x_j^{(i)} - \sum_{t:l_t=l_j} w_{jt}^{(i)} x_t^{(i)} \right\|^2
\end{aligned} \tag{3}$$

under the constraint

$$\sum_{t:l_t=l_j} w_{jt}^{(i)} = 1, \qquad w_{jt}^{(i)} = 0 \quad \text{for } l_t \neq l_j. \tag{4}$$

Then, by solving (3) and (4), we have

$$w_j^{(i)} = \frac{\sum_p C_{j,p}^{-1}}{\sum_{p,q} C_{p,q}^{-1}}, \tag{5}$$

where $C_{p,q} = (x_j^{(i)} - x_p^{(i)})^T (x_j^{(i)} - x_q^{(i)})$ denotes the local Gram matrix and $l_p = l_q = l_j$.

Once obtaining the reconstruction coefficient $w_{jt}^{(i)}$ on the ith view, then MDGPP aims to reconstruct $y_j^{(i)} (= U^T x_j^{(i)})$ from $y_t^{(i)} (= U^T x_t^{(i)})$ (where $l_t = l_j$) with $w_{jt}^{(i)}$ in the projected low-dimensional space; thus we have

$$\begin{aligned}
&\arg\min_{y_j^{(i)}} \sum_{j=1}^{n} \left\| y_j^{(i)} - \sum_{t:l_t=l_j} w_{jt}^{(i)} y_t^{(i)} \right\|^2 \\
&= \arg\min_{y_j^{(i)}} \sum_{j=1}^{n} \left\| y_j^{(i)} - \sum_{t:l_t=l_j} w_{jt}^{(i)} y_t^{(i)} \right\|^2 \\
&= \arg\min_{y_j^{(i)}} \operatorname{Tr} \left(\sum_{j=1}^{n} \left(y_j^{(i)} - \sum_{t:l_t=l_j} w_{jt}^{(i)} y_t^{(i)} \right) \right. \\
&\qquad \left. \times \left(y_j^{(i)} - \sum_{t:l_t=l_j} w_{jt}^{(i)} y_t^{(i)} \right)^T \right) \\
&= \arg\min 2\operatorname{Tr} \left(Y^{(i)} I^{(i)} \left(Y^{(i)}\right)^T - Y^{(i)} W^{(i)} \left(Y^{(i)}\right)^T \right) \\
&= \arg\min 2\operatorname{Tr} \left(Y^{(i)} \left(I^{(i)} - W^{(i)}\right) \left(Y^{(i)}\right)^T \right) \\
&= \arg\min 2\operatorname{Tr} \left(U^T X^{(i)} \left(I^{(i)} - W^{(i)}\right) \left(X^{(i)}\right)^T U \right),
\end{aligned} \tag{6}$$

where $I^{(i)}$ is an identity matrix defined on the ith view, and $W^{(i)} = [w_{jt}^{(i)}]_{j,t=1}^{n}$ is the reconstruction coefficient matrix on the ith view.

As a result, by combining (1) and (6) together, the part optimization for $X^{(i)}$ is

$$\begin{aligned}
&\arg\max_{y_j^{(i)}} \left(\sum_{j,t=1}^{n} h_{jt}^{(i)} \left\| y_j^{(i)} - y_t^{(i)} \right\|^2 \right. \\
&\qquad \left. -\lambda \sum_{j=1}^{n} \left\| y_j^{(i)} - \sum_{t:\, l_t=l_j} w_{jt}^{(i)} y_t^{(i)} \right\|^2 \right) \\
&= \arg\max \operatorname{Tr} \left(U^T X^{(i)} L^{(i)} \left(X^{(i)}\right)^T U \right. \\
&\qquad \left. -\lambda U^T X^{(i)} \left(I^{(i)} - W^{(i)}\right) \left(X^{(i)}\right)^T U \right) \\
&= \arg\max \operatorname{Tr} \left(U^T \left(X^{(i)} L^{(i)} \left(X^{(i)}\right)^T \right. \right. \\
&\qquad \left. \left. -\lambda X^{(i)} \left(I^{(i)} - W^{(i)}\right) \left(X^{(i)}\right)^T \right) U \right) \\
&= \arg\max \operatorname{Tr} \left(U^T Q^{(i)} U \right),
\end{aligned} \tag{7}$$

where $Q^{(i)} = X^{(i)} L^{(i)} (X^{(i)})^T - \lambda X^{(i)} (I^{(i)} - W^{(i)}) (X^{(i)})^T$, and λ is a tradeoff coefficient which is empirically set as 1 in this experiment.

Based on the local manifold information encoded in $L^{(i)}$ and $W^{(i)}$, (7) aims at finding a sufficiently smooth low-dimensional embedding $Y^{(i)} (= U^T X^{(i)})$ by preserving the interclass discrimination and intraclass geometry on the ith view.

Because multiviews could provide complementary information in characterizing data from different viewpoints, different views certainly have different contributions to the low-dimensional feature subspace. In order to well discover the complementary information of data from different views, a nonnegative weighted set $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_m]$ is imposed on each view independently. Generally speaking, the larger σ_i is, the more the contribution of the view $X^{(i)}$ is made to obtain the low-dimensional feature subspace. Hence, by summing over all parts of optimization defined in (7), we can formulate MDGPP as the following optimization problem:

$$\arg\max_{U,\sigma} \sum_{i=1}^{m} \sigma_i \operatorname{Tr} \left(U^T Q^{(i)} U \right) \tag{8}$$

subject to

$$U^T U = I, \qquad \sum_{i=1}^{m} \sigma_i = 1, \quad \sigma_i \geq 0. \tag{9}$$

The solution to σ in (8) subject to (9) is $\sigma_k = 1$ corresponding to the maximum $\operatorname{Tr}(U^T Q^{(i)} U)$ over different views, and $\sigma_k = 0$ otherwise, which means that only the best view is finally selected by this method. Consequently, this solution cannot meet the demand for exploring the complementary characteristics of different views to get a better low-dimensional embedding than that based on a single view. In order to avoid this problem, we set $\sigma_i \leftarrow \sigma_i^r$ with $r > 1$ by following the trick utilized in [16–19]. In this condition, $\sum_{i=1}^{m} \sigma_i^r = 1$ achieves its maximum when $\sigma_i = 1/m$ according to $\sum_{i=1}^{m} \sigma_i = 1$ and $\sigma_i \geq 0$. Similarly σ_i for different views can be obtained by setting $r > 1$; thus each view makes a specific contribution to obtaining the final low-dimensional embedding. Consequently, the new objective function of MDGPP can be defined as follows:

$$\arg\max_{U,\sigma} \sum_{i=1}^{m} \sigma_i^r \operatorname{Tr} \left(U^T Q^{(i)} U \right) \tag{10}$$

subject to

$$U^T U = I, \qquad \sum_{i=1}^{m} \sigma_i = 1, \quad \sigma_i \geq 0. \tag{11}$$

The above optimization problem is a nonlinearly constrained nonconvex optimization problem, so there is no direct approach to find its global optimal solution. In this paper, we derive an alternating-optimization-based iterative algorithm to find a local optimal solution. The alternating optimization iteratively updates the projection matrix U and weight vector $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_m]$.

First, we update σ by fixing U. The optimal problem (10) subject to (11) becomes

$$\arg\max_{U,\sigma} \sum_{i=1}^{m} \sigma_i^r \operatorname{Tr}\left(U^T Q^{(i)} U\right) \tag{12}$$

subject to

$$\sum_{i=1}^{m} \sigma_i = 1, \quad \sigma_i \geq 0. \tag{13}$$

Following the standard Lagrange multiplier, we construct the following Lagrangian function by incorporating the constraint (13) into (12):

$$L(\sigma, \lambda) = \sum_{i=1}^{m} \sigma_i^r \operatorname{Tr}\left(U^T Q^{(i)} U\right) - \lambda \left(\sum_{i=1}^{m} \sigma_i - 1 \right), \tag{14}$$

where the Lagrange multiplier λ satisfies $\lambda \geq 0$.

Taking the partial derivation of the Lagrangian function $L(\sigma, \lambda)$ with respect to σ_i and λ and setting them to zeros, we have

$$\frac{\partial L(\sigma, \lambda)}{\partial \sigma_i} = r\sigma_i^{r-1} \operatorname{Tr}\left(U^T Q^{(i)} U\right) - \lambda = 0, \tag{15}$$

$$\frac{\partial L(\sigma, \lambda)}{\partial \lambda} = \sum_{i=1}^{m} \sigma_i - 1 = 0. \tag{16}$$

Hence, according to (13) and (14), the weight coefficient σ_i can be calculated as

$$\sigma_i = \frac{\left(1/\operatorname{Tr}\left(U^T Q^{(i)} U\right)\right)^{1/(r-1)}}{\sum_{i=1}^{m} \left(1/\operatorname{Tr}\left(U^T Q^{(i)} U\right)\right)^{1/(r-1)}}. \tag{17}$$

Then, we can make the following observations according to (15): If $r \to \infty$; then the values of different σ_i will be close to each other. If $r \to 1$, then only $\sigma_i = 1$ corresponding to the maximum $\operatorname{Tr}(U^T Q^{(i)} U)$ over different views, and $\sigma_i = 0$ otherwise. Thus, the choice of r should respect to the complementary property of different views. The effect of the parameter r will be discussed in the later experiments.

Second, we update U by fixing σ. The optimal problem (10) subject to (11) can be equivalently transformed into the following form:

$$\arg\max_{U} \operatorname{Tr}\left(U^T Q U\right) \tag{18}$$

subject to

$$U^T U = I, \tag{19}$$

where $Q = \sum_{i=1}^{m} \sigma_i^r Q^{(i)}$. Since $Q^{(i)}$ defined in (7) is a symmetric matrix, Q is also a symmetric matrix.

Obviously, the solution of (18) subject to (19) can be obtained by solving the following standard eigendecomposition problem

$$QU = \lambda U. \tag{20}$$

Let the eigenvectors $U_1, U_2, \ldots, U_d$ be solutions of (20) ordered according to eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_d$. Then, the optimal projection matrix U is given by $U = [U_1, U_2, \ldots, U_d]$. Now, we discuss how to determine the reduced feature dimension d by using the Ky Fan theorem [23].

Ky Fan Theorem. Let H be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and the corresponding eigenvectors $U = [U_1, U_2, \ldots, U_n]$. Then

$$\lambda_1 + \lambda_2 + \cdots + \lambda_k = \arg\max_{X^T X = I} \operatorname{Tr}\left(X^T H X\right). \tag{21}$$

Moreover, the optimal X^* is given by $X^* = U = [U_1, U_2, \ldots, U_k]R$, where R is an arbitrary orthogonal matrix.

From the above Ky Fan theorem, we can make the following observations. The optimal solution to (18) subject to (19) is composed of the largest d eigenvectors of the matrix Q, and the optimal value of objective function (18) equals the sum of the largest d eigenvalues of the matrix Q. Therefore, the optimal reduced feature dimension d is equivalent to the number of positive eigenvalues of the matrix Q.

Alternately updating σ and U by solving (17) and (20) until convergence, we can obtain the final optimal projection matrix U for multiple views. A simple initialization for σ could be $\sigma = [1/m, \ldots, 1/m]$. According to the aforementioned statement, the proposed MDGPP algorithm is summarized as follows.

Algorithm 1 (MDGPP algorithm).

Input. A multiview data set $X = \{X^{(i)} \in R^{p_i \times n}\}_{i=1}^{m}$, the dimension of the reduced low-dimensional subspace d ($d < p_i$, $1 \leq i \leq m$), tuning parameter $r > 1$, iteration number $T_{\max}$, and convergence error ε.

Output. Projection matrix U.

Algorithm.

Step 1. Simultaneously consider both intraclass geometry and interclass discrimination information to calculate $Q^{(i)}$ for each view according to (7).

Step 2 (initialization).

(1) Set $\sigma = [1/m, 1/m, \ldots, 1/m]$;

(2) obtain U^0 by solving the eigendecomposition problem (20).

Step 3 (local optimization). For $t = 1, 2, \ldots, T_{\max}$

(1) calculate σ as shown in (17);

(2) solve the eigenvalue equation in (20);

(3) sort their eigenvectors $U_1, U_2, \ldots, U_d$ according to their corresponding eigenvalues: $\lambda_1 > \lambda_2 > \cdots > \lambda_d$, and obtain $U^t = [U_1, U_2, \ldots, U_d]$;

(4) if $t > 2$ and $|U^t - U^{t-1}| < \varepsilon$, then go to Step 4.

Step 4 (output projection matrix). Output the final optimal projection matrix $U = U^t$.

We now briefly analyze the computational complexity of the MDGPP algorithm, which is dominated by three parts. One is for constructing the matrix $Q^{(i)}$ for different views. As shown in (7), the computational complexity of this part is $O((\sum_{i=1}^{m} p_i) \times n^2)$. In addition, each iteration involves computing view weight σ and solving a standard eigendecomposition problem; the computational complexity of running two parts in each iteration is $O((m + d) \times n^2)$ and $O(n^3)$, respectively. Therefore, the total computational complexity of MDGPP is $O((\sum_{i=1}^{m} p_i) \times n^2 + ((m + d) \times n^2 + n^3) \times T_{\max})$, where $T_{\max}$ denotes the iteration number and is always set to less than five in all experiments.

4. Experimental Results

In this section, we evaluate the effectiveness of our proposed MDGPP algorithm for two image classification tasks including face recognition and facial expression recognition. Two widely used face databases including AR [24] and CMU PIE [25] are employed for face recognition evaluation, and the well-known Japanese female facial expression (JAFFE) [26] database is used for facial expression recognition evaluation. We also compare the proposed MDGPP algorithm with some traditional single-view-based dimensionality reduction algorithms, such as PCA [2], LDA [2], LPP [3], MFA [10], DGPP [12], and the three latest multiview dimensionality reduction algorithms, including MVSE [18], MSNE [21], MSE [19], and SSMVE [22]. The nearest neighbor classifier with the Euclidean distance was adopted for classification. For a fair comparison, all the results reported here are based on the best tuned parameters of all the compared algorithms.

4.1. Data Sets and Experimental Settings. We conducted face recognition experiments on the widely used AR and CMU PIE face databases and facial expression recognition experiments on the well-known Japanese female facial expression (JAFFE) database.

The AR database [24] contains over 4,000 color images corresponding to 126 people (70 men and 56 women), which include frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). Each person has 26 different images taken in two sessions (separated by two weeks). In our experiments, we used a subset of 800 face images from 100 persons (50 men and 50 women) with eight face images of different expressions and lighting conditions per person. Figure 1 shows eight sample images of one individual from the subset of the AR database.

Figure 1: Sample face images from the AR database.

Figure 2: Sample face images from the CMU PIE database.

The CMU PIE database [25] comprises more than 40,000 facial images of 68 people with different poses, illumination conditions, and facial expressions. In this experiment, we selected a subset of the CMU PIE database which consists of 3060 frontal face images with varying expression and illumination from 68 persons with 45 images from each person. Figure 2 shows some sample images of one individual from the subset of the CMU PIE database.

The Japanese female facial expression (JAFFE) database [26] contains 213 facial images of ten Japanese women. Each facial image shows one of seven expressions: neutral, happiness, sadness, surprise, anger, disgust, or fear. Figure 3 shows some facial images from the JAFFE database. In this experiment, following the general setting scheme of facial expression recognition, we discard all the neutral facial images and only utilize the remainder 183 facial images which include six basic facial expressions.

For all the face images in the above three face databases, the facial part of each image was manually aligned, cropped, and resized into 32×32 according to the eye's positions. For each facial image, we extract the commonly used four kinds of low-level visual features to represent four different views. These four features include color histogram (CH) [27], scale-invariant feature transform (SIFT) [28], Gabor [29], and local binary pattern (LBP) [30]. For the CH feature extraction, we used 64 bins to encode a histogram feature for each facial image according to [27]. For the SIFT feature extraction, we densely sampled and calculated the SIFT descriptors of 16×16 patches over a grid with spacing of 8 pixels according to [28]. For the Gabor feature extraction, following [29], we adopted 40 Gabor kernel functions from five scales and eight orientations. For the LBP feature extraction, we followed the parameter settings in [30] and utilized 256 bins to encode a histogram feature for each facial image. For more details on these four feature descriptors, please refer to [27–30]. Because these four features are complementary to each other in representing facial images, we empirically set the tuning parameter r in MDGPP to be five.

In this experiment, each facial image set was partitioned into the nonoverlap training and testing sets. For each database, we randomly selected 50% data as the training set and use the remaining 50% data as the testing set. To reduce statistical variation for each random partition, we repeated these trials independently ten times and reported the average recognition results.

Figure 3: Sample facial images from the JAFFE database.

4.2. Compared Algorithms. We compared our proposed MDGPP algorithm with the following dimensionality reduction algorithms.

(1) PCA [2]: PCA is an unsupervised dimensionality reduction algorithm.

(2) LDA [2]: LDA is a supervised dimensionality reduction algorithm. We adopted a Tikhonov regularization term μI rather than PCA preprocessing to avoid the well-known small sample size (singularity) problem in LDA.

(3) LPP [3]: LPP is an unsupervised manifold learning algorithm. There is a nearest neighbor number k to be tuned in LPP and it was empirically set to be five in our experiments. In addition, the Tikhonov regularization was also adopted to avoid the small sample size (singularity) problem in LPP.

(4) MFA [10]: MFA is a supervised manifold learning algorithm. There are two parameters (i.e., k_1 nearest neighbor number and k_2 nearest neighbor number) to be tuned in MFA. We empirically set $k_1 = 5$ and $k_2 = 20$ in our experiments. Meanwhile, the Tikhonov regularization was also adopted to avoid the small sample size (singularity) problem in MFA.

(5) DGPP [12]: DGPP is a supervised manifold learning algorithm. There is a tradeoff parameter λ to be tuned in DGPP and it was empirically set to be one in our experiments.

(6) MVSE [18]: MSVE is an initially proposed multiview algorithm for dimensionality reduction.

(7) MSNE [21]: MSNE is a probability-based unsupervised multiview algorithm. We followed the parameter setting in [21] and set the tradeoff coefficient λ to be five in our experiments.

(8) MSE [19]: MSE is a supervised multiview algorithm. There are two parameters (i.e., the nearest neighbor number k and the tuning coefficient r) to be tuned in MSE. We empirically set $k = 5$ and $r = 5$ in our experiments.

(9) SSMVE [22]: SSMVE is a sparse unsupervised multiview algorithm. We followed the parameter setting method in [22] and set the regularized parameter λ to be one in our experiments.

It is worth noting that since PCA, LDA, LPP, MFA, and DGPP are all single-view-based algorithms, these five algorithms adopt the conventional feature concatenation-based strategy to cope with the multiview data.

Table 1: Comparisons of recognition accuracy on the AR database.

Algorithm	Accuracy	Dimensions
PCA	81.5%	120
LDA	90.4%	99
LPP	90.8%	100
MFA	91.9%	100
DGPP	93.2%	100
MVSE	93.4%	100
MSNE	93.7%	100
SSMVE	94.1%	100
MSE	95.0%	100
MDGPP	98.8%	100

Table 2: Comparisons of recognition accuracy on the CMU PIE database.

Algorithm	Accuracy	Dimensions
PCA	56.8%	80
LDA	63.4%	67
LPP	65.2%	68
MFA	73.5%	68
DGPP	78.4%	68
MVSE	79.6%	68
MSNE	82.7%	68
SSMVE	84.5%	68
MSE	87.3%	68
MDGPP	95.2%	68

Table 3: Comparisons of recognition accuracy on the JAFFE database.

Algorithm	Accuracy	Dimensions
PCA	52.6%	20
LDA	65.8%	9
LPP	71.5%	10
MFA	83.2%	10
DGPP	86.3%	10
MVSE	87.8%	10
MSNE	89.2%	10
SSMVE	91.6%	10
MSE	93.4%	10
MDGPP	96.5%	10

4.3. Experimental Results. For each face image database, the recognition performance of different algorithms was evaluated on the testing data separately. The conventional nearest neighbor classifier with the Euclidean distance was applied to perform recognition in the subspace derived from different dimensionality reduction algorithms. Tables 1, 2, and 3 report the recognition accuracies and the corresponding optimal dimensions obtained on the AR, CMU PIE, and JAFFE databases, respectively. Figures 4, 5, and 6 illustrate the recognition accuracies versus the variation of reduced dimensions on the AR, CMU PIE, and JAFFE databases,

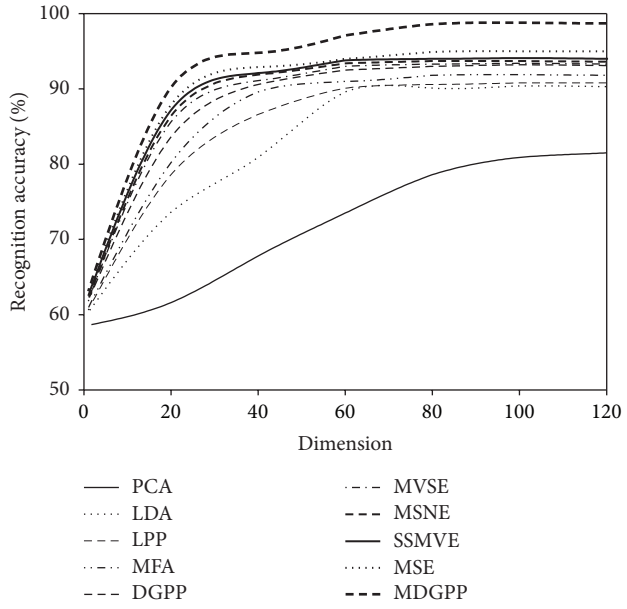


Figure 4: Recognition accuracy versus reduced dimension on the AR database.

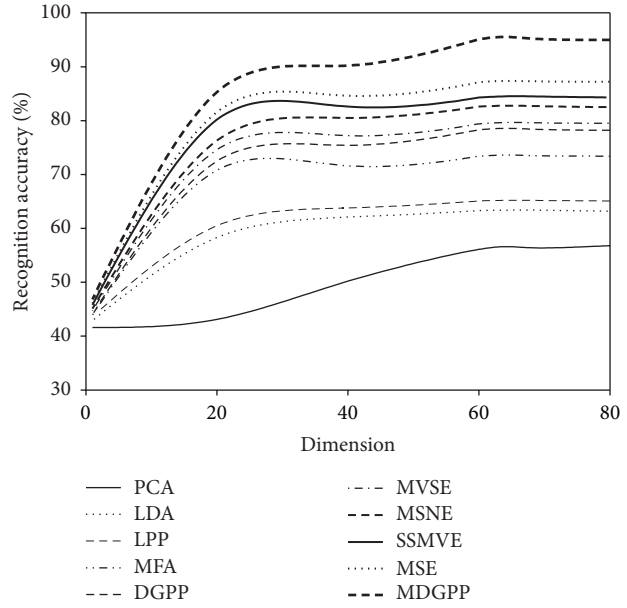


Figure 5: Recognition accuracy versus reduced dimension on the CMU PIE database.

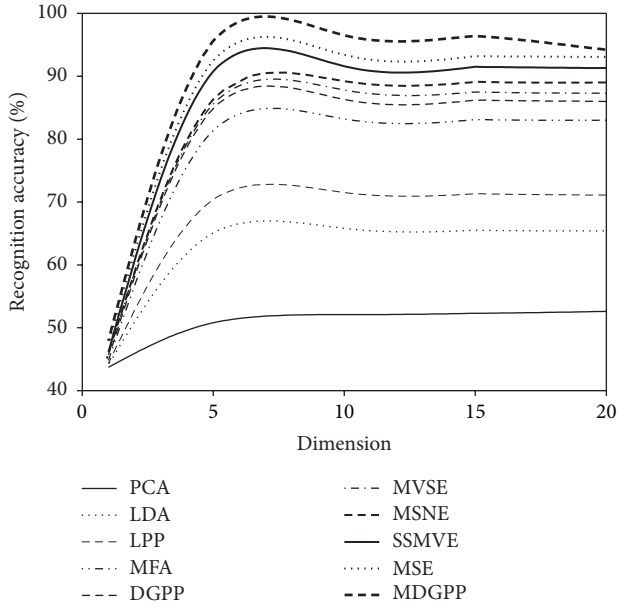


Figure 6: Recognition accuracy versus reduced dimension on the JAFFE database.

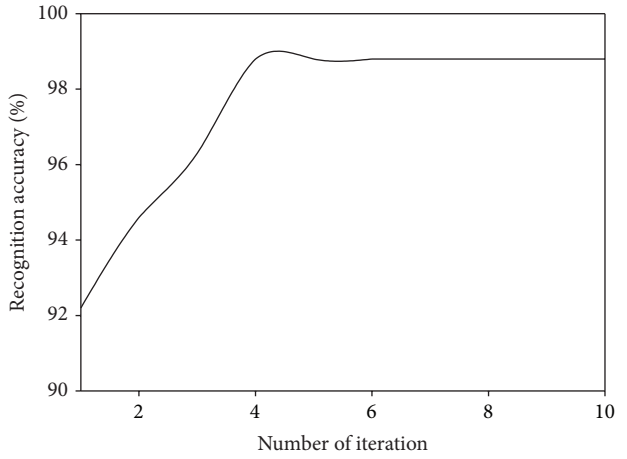


Figure 7: Recognition accuracy of MDGPP versus different numbers of iteration on the AR database.

respectively. According to the above experimental results, we can make the following observations.

(1) As can be seen from Tables 1, 2, and 3 and Figures 4, 5, and 6, our proposed MDGPP algorithm consistently outperforms the conventional single-view-based algorithms (i.e., PCA, LDA, LPP, MFA, and DGPP) and the latest multiview algorithms (i.e., MVSE, MSNE, MSE, and SSMVE) in all the experiments, which implies that extracting a discriminative feature subspace by using both intraclass geometry and interclass discrimination and explicitly considering the complementary information of different facial features can achieve the best recognition performance.

(2) The multiview learning algorithms (i.e., MVSE, MSNE, MSE, SSMVE, and MDGPP) perform much better than single-view-based algorithms (i.e., PCA, LDA, LPP, MFA, and DGPP), which demonstrates that simple concatenation strategy cannot duly combine features from multiple views, and the recognition performance can be successfully improved by exploring the complementary characteristics of different views.

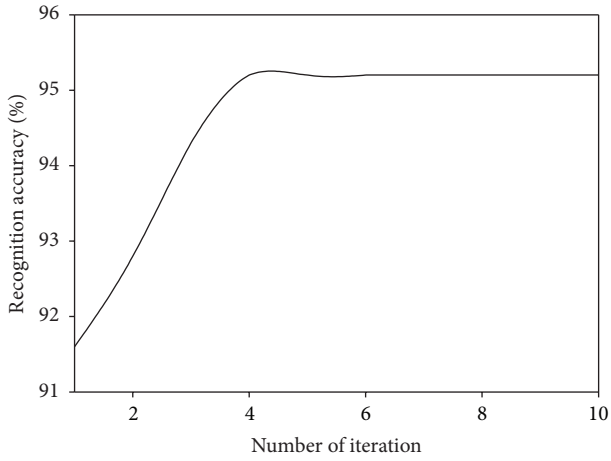


Figure 8: Recognition accuracy of MDGPP versus different numbers of iteration on the CMU PIE database.

Figure 9: Recognition accuracy of MDGPP versus different numbers of iteration on the JAFFE database.

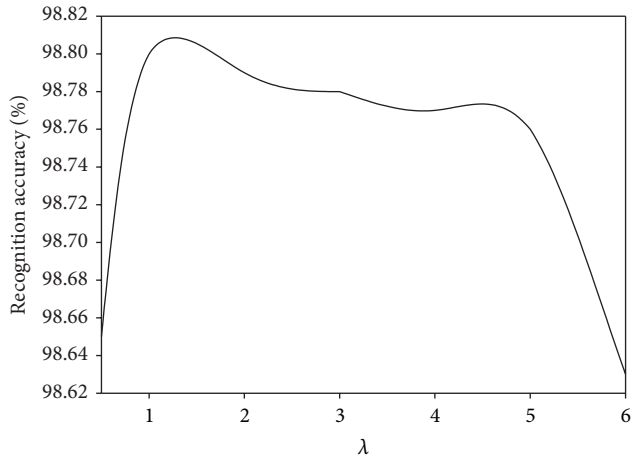


Figure 10: Recognition accuracy of MDGPP versus varying parameter λ on the AR database.

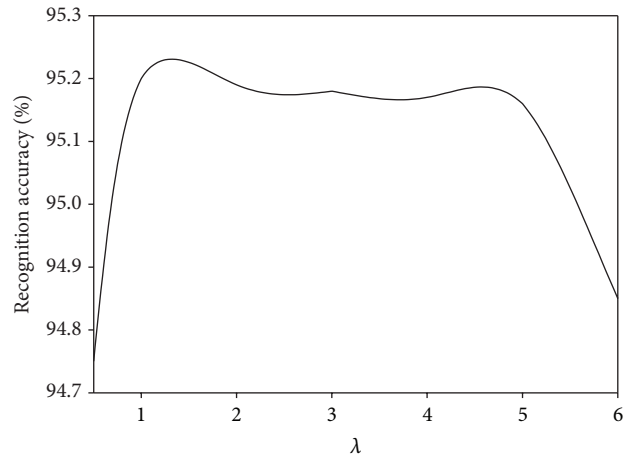


Figure 11: Recognition accuracy of MDGPP versus varying parameter λ on the CMU PIE database.

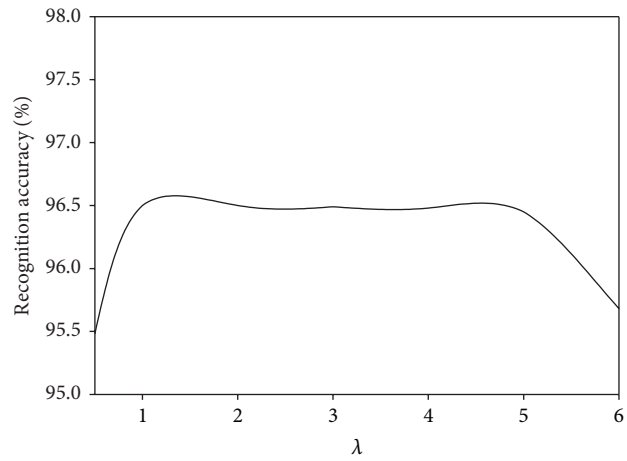


Figure 12: Recognition accuracy of MDGPP versus varying parameter λ on the JAFFE database.

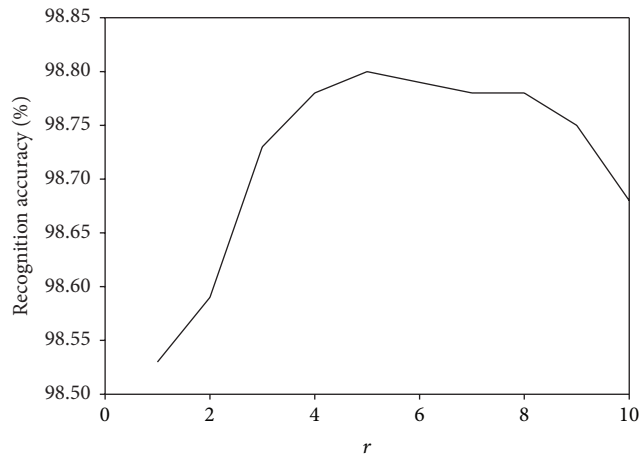


Figure 13: Recognition accuracy of MDGPP versus varying parameter r on the AR database.

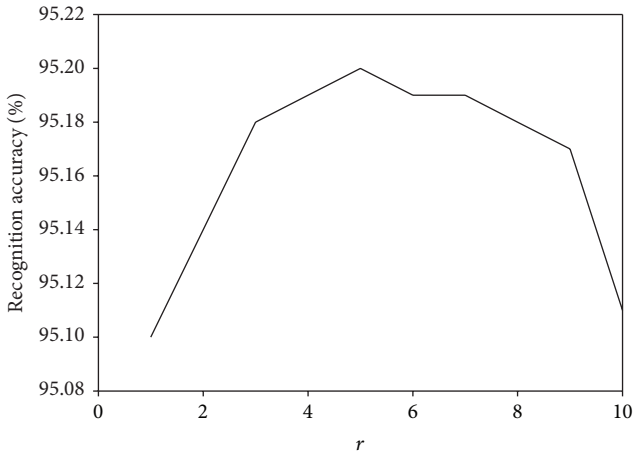


Figure 14: Recognition accuracy of MDGPP versus varying parameter r on the CMU PIE database.

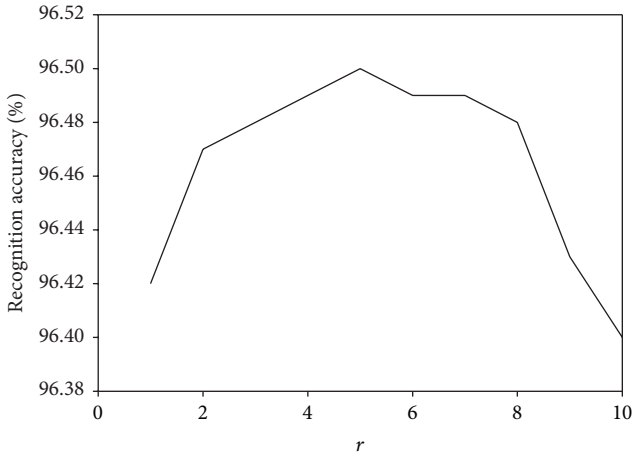


Figure 15: Recognition accuracy of MDGPP versus varying parameter r on the JAFFE database.

(3) For the single-view-based algorithms, the manifold learning algorithms (i.e., LPP, MFA, and DGPP) perform much better than the conventional dimensionality reduction algorithms (i.e., PCA and LDA). This observation confirms that the local manifold structure information is crucial for image classification. Moreover, the supervised manifold learning algorithms (i.e., MFA and DGPP) perform much better than the unsupervised manifold learning algorithm LPP, which demonstrates that the utilization of discriminant information is useful to improve the image classification performance.

(4) For the multiview learning algorithms, the supervised multiview algorithms (i.e., MSE and MDGPP) outperform the unsupervised multiview algorithms (i.e., MVSE, MSNE, and SSMVE) due to the utilization of the labeled facial images.

(5) Although MVSE, MSNE, and SSMVE are all unsupervised multiview learning algorithms, SSMVE performs much better than MVSE and MSNE. The possible explanation is that the SSMVE algorithm adopts the sparse coding technique, which is naturally discriminative in determining the appropriate combination of different views.

(6) Among the compared multiview learning algorithms, MVSE performs the worst. The reason is that MVSE performs a dimensionality reduction process on each view independently. Hence it cannot fully integrate the complementary information of different views to produce a good low-dimensional embedding.

(7) MDGPP can improve the recognition performance of DGPP. The reason is that MDGPP can make use of multiple facial feature representations in a common learned subspace such that some complementary information can be explored for recognition task.

4.4. Convergence Analysis. Since our proposed MDGPP is an iteration algorithm, we also evaluate its recognition performance with different numbers of iteration. Figures 7, 8, and 9 show the recognition accuracy of MDGPP versus different numbers of iteration on the AR, CMU PIE, and JAFFE databases, respectively. As can be seen from these figures, we can observe that our proposed MDGPP algorithm can converge to a local optimal optimum value in less than five iterations.

4.5. Parameter Analysis. We investigate the parameter effects of our proposed MDGPP algorithm: tradeoff coefficient λ and tuning parameter r. Since each parameter can affect the recognition performance, we fix one parameter as used in the previous experiments and test the effect of the remaining one. Figures 10, 11, and 12 show the influence of the parameter λ in the MDGPP algorithm on the AR, CMU PIE, and JAFFE databases, respectively. Figures 13, 14, and 15 show the influence of the parameter r in the MDGPP algorithm on the AR, CMU PIE, and JAFFE databases, respectively. From Figure 10 to Figure 15, we can observe that MDGPP demonstrates a stable recognition performance over a large range of both λ and r. Therefore, we can conclude that the performance of MDGPP is not sensitive to the parameters λ and r.

5. Conclusion

In this paper, we have proposed a new multiview learning algorithm, called multiview discriminative geometry preserving projection (MDGPP) for feature extraction and classification by exploring the complementary property of different views. MDGPP can encode different features from different views in a physically meaningful subspace and learn a low-dimensional and sufficiently smooth embedding over all views simultaneously with an alternating-optimization-based iterative algorithm. Experimental results on three face image databases show that the proposed MDGPP algorithm

outperforms other multiview and single view learning algorithms.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants no. 70701013 and 61174056, the Natural Science Foundation of Henan Province under Grant no. 102300410020, the National Science Foundation for Postdoctoral Scientists of China under Grant no. 2011M500035, and the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant no. 20110023110002.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[3] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[4] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[6] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[7] X. He and P. Niyogi, "Locality preserving projections," *Advances in Neural Information Processing Systems*, vol. 16, pp. 585–591, 2003.

[8] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 846–853, San Diego, Calif, USA, June 2005.

[9] D. Cai, X. He, K. Zhou, J. W. Han, and H. J. Bao, "Locality sensitive discriminant analysis," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*, pp. 708–713, Hyderabad, India, January 2007.

[10] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[11] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[12] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 174–184, 2010.

[13] J. Lu and Y.-P. Tan, "A doubly weighted approach for appearance-based subspace learning methods," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 1, pp. 71–81, 2010.

[14] J. Lu, J. Hu, X. Zhou, Y. Shang, Y. Tan -P, and G. Wang, "Neighborhood repulsed metric learning for kinship verification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2594–2601, Providence, RI, USA, June 2012.

[15] D. Zhou and C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 1159–1166, Corvalis, Ore, USA, June 2007.

[16] J. Yu, D. Liu, D. Tao, and H. S. Seah, "On combining multiple features for cartoon character retrieval and clip synthesis," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 42, no. 5, pp. 1413–1427, 2012.

[17] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4636–4648, 2012.

[18] B. Long, P. S. Yu, and Z. Zhang, "A general model for multiple view unsupervised learning," in *Proceedings of the 8th SIAM International Conference on Data Mining*, pp. 822–833, Atlanta, Ga, USA, April 2008.

[19] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 40, no. 6, pp. 1438–1446, 2010.

[20] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1299–1313, 2009.

[21] B. Xie, Y. Mu, D. Tao, and K. Huang, "M-SNE: multiview stochastic neighbor embedding," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 41, no. 4, pp. 1088–1096, 2011.

[22] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang, "Sparse unsupervised dimensionality reduction for multiple view data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1485–1496, 2012.

[23] R. Bhatia, *Matrix Analysis*, Springer, New York, NY, USA, 1997.

[24] A. M. Martinez and R. Benavente, *AR Face Database*, CVC Technology Report, 1998.

[25] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[26] M. J. Lyons, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[27] L. G. Shapiro and G. C. Stockman, *Computer Vision*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2003.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[29] T. S. Lee, "Image representation using 2d gabor wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[30] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.