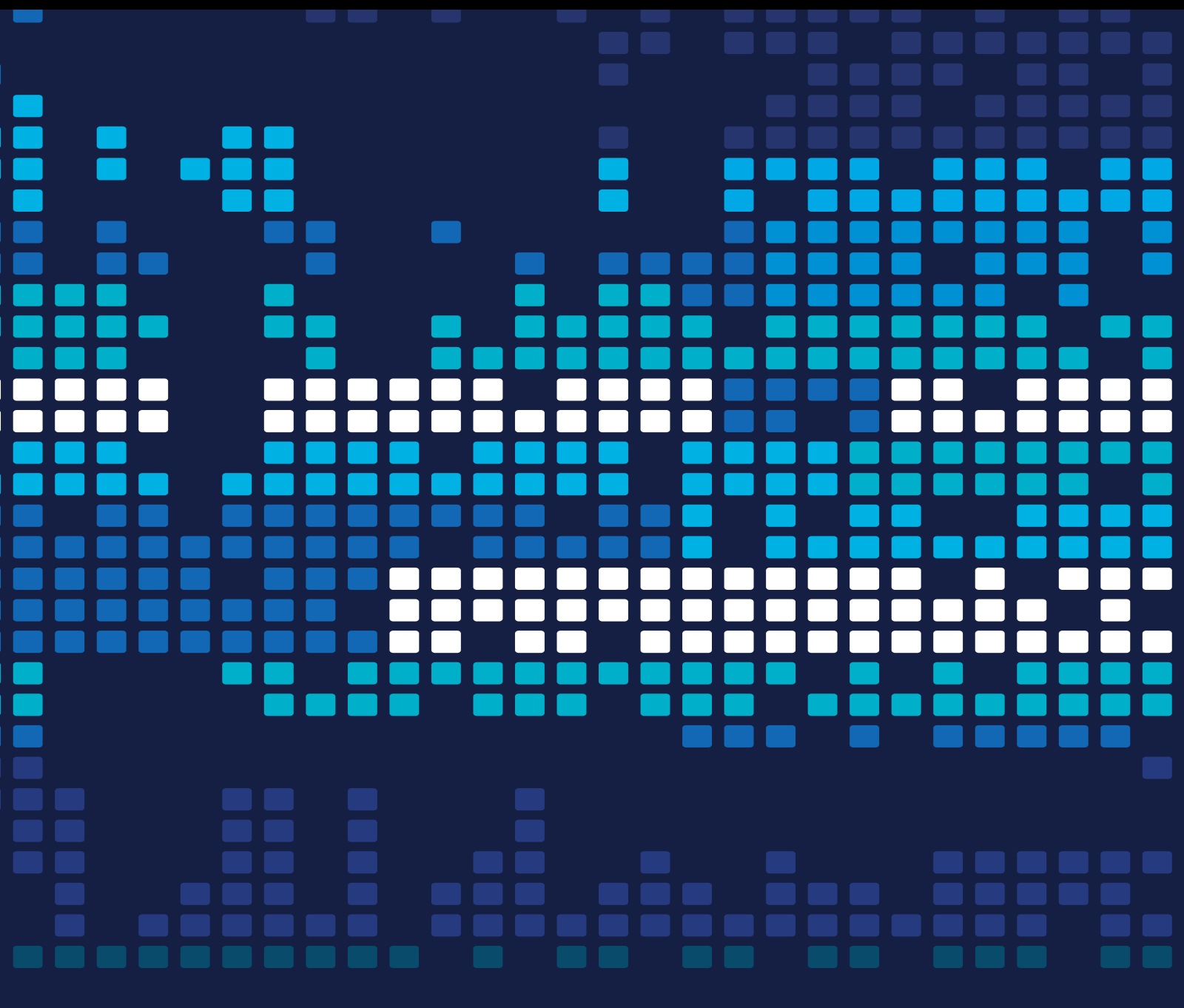
Scientific Programming

Scientific Programming towards a Smart World

Guest Editors: Wenbing Zhao, Xiong Luo, Huaping Liu, Kun Hua, and Chaomin Luo



Scientific Programming towards a Smart World

Scientific Programming towards a Smart World

Guest Editors: Wenbing Zhao, Xiong Luo, Huaping Liu, Kun Hua, and Chaomin Luo

Editorial Board

Contents

Hindawi
Scientific Programming
Volume 2017, Article ID 3706232, 2 pages
http://dx.doi.org/10.1155/2017/3706232

Editorial

Scientific Programming towards a Smart World

Wenbing Zhao,[1] Xiong Luo,[2] HuaPing Liu,[3] Kun Hua,[4] and Chaomin Luo[5]

[1]*Cleveland State University, Cleveland, OH, USA*
[2]*University of Science and Technology Beijing, Beijing, China*
[3]*Tsinghua University, Beijing, China*
[4]*Lawrence Technological University, Southfield, MI, USA*
[5]*University of Detroit Mercy, Detroit, MI, USA*

Correspondence should be addressed to Wenbing Zhao; w.zhao1@csuohio.edu

Received 26 February 2017; Accepted 27 February 2017; Published 20 March 2017

In the past several years, we have seen dramatic advancement in many application domains enabled by the integration of Internet of Things and cloud computing, computational intelligence algorithms, and soft computing methodologies. This development has resulted in many emerging highly multidisciplinary research areas typically termed as smart-* technologies and systems, including smart-healthcare, smart-home, and smart-grid, as well as smart vehicles and intelligent transportation systems. These new technologies are transforming our society towards a smart world.

This special issue attracted 32 high quality submissions from countries around the globe, including China, USA, UK, Korea, Saudi Arabia, Greece, and Algeria. After a rigorous review process, 18 papers were accepted in this issue. The research presented in these papers broadly covers the interesting scope of this special issue, including computer networking (3 papers), intelligent transportation and navigation (3 papers), computer vision (2 papers), knowledge discovery (2 papers), scientific computing (2 papers), cloud computing (2 papers), gesture recognition and authorization (2 papers), and Internet of Things and machine learning (1 each).

In the area of computer networking, X. Yang et al. proposed a novel localization algorithm based on Voronoi diagram and rank sequence in wireless sensor networks; X. Liu et al. introduced a new program code dissemination scheme (referred to as FPCD) for wireless software defined networks; and K.-H. Kim et al. reported a cascade control based method for queue management in smart-grid IP networks.

In the areas of intelligent transportation and navigation, M. Dong et al. described a sophisticated navigation method using sonar point cloud data for underwater remotely operated vehicles; Y. Xing et al. proposed a new traffic flow forecasting method based on extreme learning machine; and H. Gao et al. proposed a novel lateral control method for intelligent vehicles.

On computing vision research, H. Wang et al. introduced a robust automatic target recognition algorithm for large scene images, and Y. Li et al. reported a novel approach for scene reconstruction based on metric online monocular SLAM.

On the topic of cloud computing, Y. Lu et al. proposed a new workload forecasting model for cloud services, which is based on random variable learning rate backpropagation neural network, and X. Li et al. introduced a novel resource allocation mechanism for cloud workflow systems, where the resource allocation problem is treated as a market-oriented reverse auction problem.

On human gesture related research, B. Fang et al. provided a comprehensive report on how to use magnetic and inertial measurement units to accurately recognize hand and arm gestures, and A. M. H. Wong and D.-K. Kang proposed a novel stationary hand gesture authentication scheme using edit distance on finger pointing direction interval.

This issue also includes two papers on knowledge discovery. Y. Zhu et al. reported a research on using neural network based machine learning methods to build a geological knowledge graph. X. Han et al. described how to use fragment based semantic graph model to perform text summarization.

In addition, this special issue includes exciting works on using computer programming algorithms to help solve science problems. P. Shi et al. introduced a novel experiment platform for materials correction study. The throughput of the platform was significantly improved by using state of the art computing vision algorithms. Furthermore, X. Wang et al. proposed an efficient and simple rigid-fluid coupling scheme with programming algorithms for particle-based fluid simulation and visualization.

Finally, Z. Liu et al. proposed a novel sparse representation method for the class-imbalance problem. Last, but not least, L. Ding et al. reported a very interesting system based on Internet of Things that can be used to monitor dangerous goods in container yards.

Acknowledgments

The guest editors would like to thank the authors for contributing to this special issue and thank all the reviewers for their time and rigorous reviews.

Wenbing Zhao
Xiong Luo
HuaPing Liu
Kun Hua
Chaomin Luo

Hindawi
Scientific Programming
Volume 2017, Article ID 2073680, 12 pages
http://dx.doi.org/10.1155/2017/2073680

Research Article

Probabilistic Forecasting of Traffic Flow Using Multikernel Based Extreme Learning Machine

Yiming Xing,[1,2] Xiaojuan Ban,[1] and Chong Guo[1]

[1]*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China*
[2]*Engineering Training Center of Shenyang Aerospace University, Shenyang 110136, China*

Correspondence should be addressed to Yiming Xing; b20120409@xs.ustb.edu.cn

Received 7 August 2016; Accepted 1 December 2016; Published 16 March 2017

Academic Editor: Wenbing Zhao

Real-time and accurate prediction of traffic flow is the key to intelligent transportation systems (ITS). However, due to the nonstationarity of traffic flow data, traditional point forecasting can hardly be accurate, so probabilistic forecasting methods are essential for quantification of the potential risks and uncertainties for traffic management. A probabilistic forecasting model of traffic flow based on a multikernel extreme learning machine (MKELM) is proposed. Moreover, the optimal output weights of MKELM are obtained by utilizing Quantum-behaved particle swarm optimization (QPSO) algorithm. To verify its effectiveness, traffic flow probabilistic prediction using QPSO-MKELM was compared with other learning methods. Experimental results show that QPSO-MKELM is more effective for practical applications. And it will help traffic managers to make right decisions.

1. Introduction

Recently, the traffic flows maintain a steady growth in both urban and rural traffic leading to pollution, accidents, and congestion. To solve the problems, the intelligent transportation systems (ITS) are developed by many countries. The effectiveness of ITS is improved by using a lot of modern information technologies. According to the prediction period of time, traffic flow prediction can be divided into long-term, mid-term, and short-term prediction. Short-term traffic flow forecasting has become one of the main research areas of ITS. Prediction of real-time and accurate traffic flow becomes extremely important for effective traffic management systems, including traffic control, traffic induction, and vehicle routing. The problem induced by the randomness, nonlinearity, and complexity of traffic flow has compelled us to search for more reliable techniques to forecast traffic flow.

A lot of short-term traffic prediction algorithms are proposed in the literature [1–7]. Conventionally, a majority of study focus on developing accurate point-prediction method structures and learning algorithms for short-term traffic flow prediction, but these methods cannot be used for the quantitative analysis of the uncertainty of the prediction. In fact, there are lots of traffic variables influencing the results of traffic flow prediction such as weather, date and time, types, and flow parameters. The aim of the traffic flow forecasting model is to utilize these variables to predict the traffic flow.

Because of the chaotic property of traffic flow, mistakes in prediction are simply inevitable. In other words, forecast with exact-point value becomes inadequate to describe the real world information. To deal with the problem, prediction intervals (PIs) are vital for quantifying the underlying risks and uncertainties. PIs are a list of ranges including upper and lower bounds where the targets will lie. On the basis of the PIs with corresponding confidence level, the quantitative uncertainties of traffic flow predictions can supply valuable information to traffic managers for good preparation for the most severe and excellent situations in advance.

2. Literature Review

The exact point-prediction methods only supply the point forecasting value, while PIs work as intervals that consist of upper and lower bounds as well as pointing out the probability of correct forecast. PIs not only indicate the prediction accuracy, but also provide a range that targets will locate

[8, 9]. Some investigations in regression problems along with interval outputs have been conducted and they are able to be sorted into two classifications based on the amount of models they utilized. Studies in the first category use only one model, whereas they bring in other approaches to acquire interval outputs. These PIs construction approaches are usually set after a point-prediction model with particular preceding assumptions. Bayesian [10], mean-variance estimation [11], and bootstrap [12] are frequently used to obtain PIs. The main disadvantage of these existing methods is the high computational requirement. The other category focuses on learning a twin model; the PIs construction method named lower upper bound estimation (LUBE) approach is brought forward in [13]. Nevertheless, traditional neural networks (NNs) applied in the LUBE approach have the issue of high computational cost and overtraining.

ELM [14] is brought forward to train single-hidden layer feedforward neural networks (SLFNs). Although ELM extensively enhances the training effectiveness, fluctuation caused by the random input and hidden layer weights influences the steadiness of ELM in the situation of identical training data as well as model parameter [15]. The method that replaces ELM hidden layer with a kernel function makes ELM avoid choosing input and hidden layer weights randomly since the computation for hidden input is conducted by kernel function. The problem in ELM caused by random input and hidden layer parameters is solved by Kernel-ELM which gains higher stability by sacrificing the training rate [15]. Moreover, one kernel function is usually used in the standard kernel learning algorithms [16]. In our last paper [17], we proposed a single kernel extreme learning machine- (KELM-) based probabilistic forecasting method of traffic flow. On the basis of multikernel learning thought [18], the composites of two kernels may combine the good characteristics of them and have better performance than any other single kernel. The polynomial kernel function and the Gaussian kernel function are mixed as a kernel function which includes both kernels' advantages in this study.

A novel probabilistic traffic flow prediction approach is brought forward based on the multikernel ELM in this paper, which is applied to construct PIs for traffic flow. Then the output weights from MKELM models are optimized by Quantum-behaved particle swarm optimization (QPSO) [19]. The proposed approach has been examined by the practical traffic flow data. Accurate prediction results have represented the good performance of the QPSO-MKELM approach for traffic flow.

The rest of this paper is organized as follows. In Section 3 the algorithms of framework of MKELM and QPSO are presented. Section 4 proposes a novel MKELM model to construct optimal PIs for traffic flow with QPSO. The experiments to verify the effectiveness of the model are carried out in Section 5. Finally, Section 6 draws the conclusion.

3. Methodology

3.1. Multiple Kernel-Based ELM. ELM, developed by Huang et al. [20], is a novel learning algorithm for SLFNs that randomly selects hidden nodes parameters as well as ascertaining the output layer parameters of SLFNs analytically.

In a specified training dataset including n samples, (x, y) is a training sample. The formula of the SLFN with N_h hidden nodes is able to be represented as

$$\sum_{i=1}^{N_h} \beta_i g\left(w_i x + b_i\right) = y. \tag{1}$$

The output equation of ELM is able to be represented as

$$\begin{aligned} y &= F_{\mathrm{ELM}}(x) = \sum_{i=1}^{N_h} \beta_i g\left(w_i x + b_i\right) = h(x)\beta, \\ Y &= H\beta, \\ Y &= \left[y_1, \ldots, y_n\right]^T,\ H = \left[h(x_1), \ldots, h(x_n)\right]^T. \end{aligned} \tag{2}$$

According to ELM theory [14], input as well as hidden layer parameters are able to be randomly distributed as long as the activation function becomes infinitely differentiable. As to fixed input weights w_i and the hidden layer biases b_i, to train an SLFN is simply equal to figuring out a least-squares solution $\widehat{\beta}$ of the linear system $Y = H\beta$.

The smallest norm least-squares solution of the above linear system is

$$\widehat{\beta} = H^{+}Y, \tag{3}$$

where H^{+} represents the Moore–Penrose generalized inverse of matrix H.

With a user-defined cost coefficient C, Huang et al. [15] optimized the computation for the output weights β. Among them, when the amount of the hidden nodes is larger compared to that of train data, the result of β is

$$\beta = H^T\left(\frac{I}{C} + HH^T\right)^{-1} Y. \tag{4}$$

The hidden layer output of every sample $h(x_i)$ is able to be considered as an nonlinear mapping of samples x_i. Then

$$\begin{aligned} HH^T &= \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix}_{N*L} \cdot \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix}_{N*L}^{T} \\ &= \begin{bmatrix} h(x_1)\cdot h(x_1) & \cdots & h(x_1)\cdot h(x_N) \\ \vdots & \ddots & \vdots \\ h(x_N)\cdot h(x_1) & \cdots & h(x_N)\cdot h(x_N) \end{bmatrix}_{N*N}. \end{aligned} \tag{5}$$

When the hidden layer characteristic mapping $h(x)$ could not be unidentified, Huang et al. [15] suggested applying a kernel function. According to theory of kernel function, the kernel matrix Ω for ELM is able to be described as follows, where $K(x_i, y_i)$ is kernel function:

$$\Omega_{\mathrm{ELM}i,j} = h\left(x_i\right)\cdot h\left(x_j\right) = K\left(x_i, x_j\right). \tag{6}$$

So (5) can be deduced:

$$HH^T = \Omega_{\mathrm{ELM}i,j} = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_N) \\ \vdots & \ddots & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{bmatrix} = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix}_{N*L} \cdot \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix}_{N*L}^T, \tag{7}$$

$$h(x) H^T = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}. \tag{8}$$

Then the output function of Kernel-ELM can be represented as

$$y = F_{\mathrm{ELM}}(x) = h(x)\beta = h(x) H^T \left(\frac{I}{C} + HH^T\right)^{-1} Y = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_n) \end{bmatrix} \left(\frac{I}{C} + \Omega_{\mathrm{ELM}i,j}\right)^{-1} Y. \tag{9}$$

Any kernel function which meets Mercer's theorem [21] is able to be a kernel function of KELM. The kernels have their own advantages and shortcomings, which can be divided into two types, global and local kernels [16]. In local kernels, only the data that are close to or in the proximity of each other have obvious effects on kernel values. The Gaussian kernel function is an example of a typical local kernel. In contrast, a global kernel allows data points that are far away from each other to have obvious effects on the kernel value as well. A typical example of a global kernel is the polynomial kernel. In this work, we proposed a multikernel function which is taking the advantages of the polynomial kernel function and the Gaussian kernel function. The multikernel function is described as

$$K(x_i, x_j) = (1 - \lambda) K_{\mathrm{poly}}(x_i, x_j) + \lambda K_g(x_i, x_j) = \lambda \cdot \exp\left(-g \left\|x_i - x_j\right\|^2\right) + (1 - \lambda)\left(x_i \cdot x_j + 1\right)^p. \tag{10}$$

In multikernel function, λ is an adjusting coefficient between K_g and K_{poly}. When λ approximates to 0, the value of K approaches K_{poly}. In the same way, when λ approximates to 1, the value of K approaches K_g. Moreover, the width g and the degree p are the kernel parameters of the Gaussian kernel and the polynomial kernel functions, separately.

For achieving better generalization performance, the four variables λ, g, and p of MKELM and the punish parameter C need to be chosen appropriately.

3.2. QPSO. As an advanced method, QPSO [18] is an intelligence searching method; in the meanwhile, the status of every particle is illustrated by using wave function when all the particles are transferred into the quantum space, rather than the position as well as velocity in traditional PSO [19]. The particles of QPSO could appear much far away from current locations according to characteristics of the wave function, which enhances the possibility of escaping from the local optimal value. Considering that the QPSO algorithms include S-sized particle population in the D-dimensional searching space, the ith particle x_i is renewed based on

$$x_{id} = P_d \pm \delta \cdot \left|m_{\mathrm{best}} - x_{id}\right| \cdot \ln\left(\frac{1}{u}\right), \tag{11}$$

$$P_d = \frac{\varphi_{1d} p_{id} + \varphi_{2d} p_{gd}}{\varphi_{1d} + \varphi_{2d}}, \tag{12}$$

$$m_{\mathrm{best}} = \frac{\sum_{i=1}^{S} P_i}{S} = \left(\sum_{i=1}^{s} \frac{p_{i1}}{s}, \sum_{i=1}^{s} \frac{p_{i2}}{s} \cdots \sum_{i=1}^{s} \frac{p_{iD}}{s}\right). \tag{13}$$

where $d = 1, 2, \ldots, D$, $i = 1, 2, \ldots, S$, μ is a random number distributed equally on (0, 1), δ is a positive number as well as ranges from $\delta_{\min}$ to $\delta_{\max}$ called the contraction expansion (CE) coefficient, which balances the local and the global optimum, P_d is known as the local attractor of every particle at d-dimension on the basis of the trajectory analyses in [22], m_{best} is the mean best position, p_{id} is the best previous position of particle i, p_{gd} is the position of global best particle, and φ_{1d} and φ_{2d} are two different random vectors.

4. PIs Model Construction by QPSO-KELM

According to the theory of PIs, when targets lie in the PI nominal confidence (PINC), the prediction values need to be in the constructed PIs at the possibility of PINC equals $100(1 - \alpha)\%$. It can be expressed as

$$t_i \in \left[\widetilde{L}_t^{\alpha}(x_i), \widetilde{U}_t^{\alpha}(x_i)\right], \tag{14}$$

where $\widetilde{L}_t^{\alpha}(x_i)$ and $\widetilde{U}_t^{\alpha}(x_i)$ are the lower borders and upper borders of the prediction PIs at the nominal confidence level of the input x_i.

To construct PIs for the traffic flows, the KELM-based PIs establishing approach can be employed, which is illustrated in Figure 1. The KELM-based probabilistic prediction model targets are shown to directly create the upper borders and lower borders of two outputs. In Figure 1, $U(x)$ and $L(x)$ are the two output vectors of SLFN with respect to the input sample x_i, $\omega_i = [\omega_{i1}, \omega_{i2}, \ldots, \omega_{in}]^T$ is the weight vector connecting the ith hidden node and the input nodes, and $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{in}]^T$ is the weight vector connecting the ith hidden node and the output nodes. Besides, the determination for reliable PIs of the traffic flow via straight optimization for both sharpness and dependability will be demonstrated in the following part.

4.1. Construction of Optimal PIs for Traffic Flow. To assess the properties of the PIs acquired, the PI coverage probability

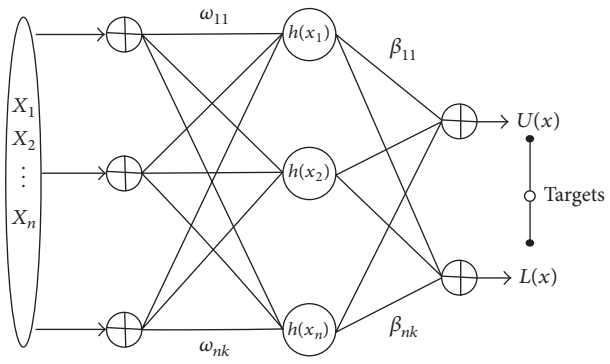


Figure 1: KELM-based PIs construction method.

(PICP) as well as the PI normalized average width (PINAW), which represents the dependability and the sharpness of PIs, is utilized in [11]. PICP is a crucial indicator regarding the dependability for the established PIs, which can be presented by

$$\text{PICP} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} c_i, \tag{15}$$

where N_{test} refers to the number for testing data. If the target t_i locates within the ith lower border and the ith upper border, then $c_i = 1$, or else $c_i = 0$.

In the procedure of interval prediction, the future targets are anticipated to locate within borders of constructed PIs at the PINC level. Nevertheless, it is able to be simply achieved by enlarging PIs from upper or lower bound. Actually, such PIs are meaningless for making a decision. Among the publications, PINAW is defined to indicate the PIs average width quantitatively, which is presented by

$$\text{PINAW} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left(U_i - L_i\right), \tag{16}$$

where R refers to the targets' range and R is applied to standardize the average width of PIs in terms of percentage.

To guarantee the high properties of created traffic flow's PIs, the output layer parameters β of the MKELM model are improved to lead to higher PICP as well as lower PINAW value. The extensively utilized coverage width-based criterion (CWC) [11] in (11) can be considered for training the KELM model:

$$\text{CWC} = \text{PINAW} \cdot \left(1 + \gamma_{\text{PICP}} \cdot e^{-\tau(\text{PICP}-\mu)}\right). \tag{17}$$

In (11), μ denotes the confidence probability $100(1-\alpha)\%$. τ works to magnify any tiny discrepancy between PICP and μ, and γ_{PICP} is a function of PICP and defined as

$$\gamma_{\text{PICP}} = \begin{cases} 0, & \text{PICP} > \mu \\ 1, & \text{PICP} < \mu. \end{cases} \tag{18}$$

When PICP is more than the given confidence probability μ, $\gamma_{\text{PICP}} = 0$ and CWC inclines to be identical to PINAW. Otherwise, $\gamma_{\text{PICP}} = 1$ and the function is going to be calculated by CWC. However, it is difficult to decide the value of τ when the CWC function is used.

In this paper, a width deviation rule (WDR) is brought forward to form a new objective function where the exponential punishment term in CWC will be substituted by the equation called WD as follows:

$$\begin{aligned} f(\beta) &= \text{PINAW} + \gamma_{\text{PICP}} \cdot \sigma \cdot \text{WD} \\ &= \text{PINAW} + \gamma_{\text{PICP}} \cdot \sigma \cdot \sum_{i=1}^{N_{\text{test}}} \text{WD}_i, \end{aligned} \tag{19}$$

$$\text{WD}_i = \begin{cases} \dfrac{L(x_i) - t_i}{U(x_i) - L(x_i)}, & \text{if } t_i < L(x_i) \\ 0, & \text{if } t_i \in I(x_i) \\ \dfrac{t_i - U(x_i)}{U(x_i) - L(x_i)}, & \text{if } t_i > U(x_i). \end{cases} \tag{20}$$

The PIs width deviation for each sample can be expressed in (19). σ is set to $1/\alpha^2$ as a penalty coefficient in the present work. For PICP $> \mu$, $f(\beta)$ will be equal to PINAW which is the same as CWC. Otherwise, the width deviation information of all samples is considered as a penalty to describe the sharpness more comprehensively.

4.2. QPSO Algorithm for PIs Optimization. To obtain the effective interval forecasts of traffic flow, the optimal output weights β of MKELM need to be calculated by using QPSO algorithm. The properties of acquired PIs are mirrored by the fitness values of every particle, which is obtained by computing the aim function in (11). The main steps of PIs optimization by using QPSO are demonstrated as follows.

(1) Preprocess. Standardize the training samples and testing samples to $[0, 1]$.

(2) Initialization. The interval prediction model based on MKELM with two outputs needs to be constructed. The upper bounds are set 30% higher than the targets. In the same way, the lower bounds are set 30% lower than the targets. Calculate the initial output matrix of hidden layer β_{int}. The particles position $[x_{i1}, x_{i2}, \ldots, x_{id}]$ is initialized by the initial output weights β_{int}.

(3) Construction of PIs and Evaluation of Cost Function. With the initialization parameters, the PIs model based on MKELM is constructed and the fitness and sharpness are calculated for each particle using (15) and (16).

(4) Renew the Position of Each Particle. All particles' locations are updated in the light of (11)–(13).

(5) Renew P_i and P_g. With the renewed β, establish a novel model and assess the fitness values of fresh particles. If present fitness value is better compared to that of P_i, afterwards P_i will

(a) Overall view of road pieces in Nanning city

(b) Enlarged view of part of road pieces in Nanning city

Figure 2: Roads are split into many small pieces. Road pieces are the basic unit to evaluate or predict.

be renewed. Moreover, if the fitness is better compared to P_g, then P_g will be renewed.

(6) Loop. On condition that the largest amount of repetition has not been achieved, go back to step (4). Or else, the process ends and the optimal output weights are acquired to establish MKELM-based model.

5. Application Studies

5.1. Database. The basic data source comes from traffic information detection system of Traffic Information Center of Nanning, which covers traffic data from main ways, minor roads, and branch roads of Nanning in Figure 2. Data collection is from April 15, 2015, to May 16, 2015. For instance, even if 2,400,000 training data are generated from May 15 to May 16, only 1,515,447 different training data are valid, and the redundancy of it needs to be eliminated. To get an effective database, the original traffic data need to be processed in advance. For example, when the speed is more than 100 km/h, the sample is invalid and needs to be eliminated. When flow, speed, and occupancy are zeros at the same time, the sample also will be removed. Each simulation runs 4-fold cross validation apart from particular account. The training samples are randomly divided into 4 same scale subsamples, where a single one is kept to be the validation data to test the model, and the rest are utilized to be training data every time. The detailed sets on May 15 are illustrated in Table 1.

The 258 attributes include the most important seven kinds of features such as date of week, time of day, traffic density, speed, flow, weather condition, and road logical region in Table 4. Generally, attributes either point out measurements for certain continuous scale such as traffic flow for the last time under this circumstance or demonstrate information about certain definite or separate features such as date type within the above characteristics. The characteristics utilized in this paper blend definite characteristics with real-valued characteristics and are supposed to be converted to all definite or real-valued characteristics. In our conditions, all definite characteristics possess only single value, which is set to be either 0 or 1. For example, characteristic date type from Monday to Friday is going to be converted to 7 new characteristics, each of which refers to either 0 or 1 for pointing out whether it occurs. Therefore, all the characteristics are real-valued and are able to be equally evaluated.

Table 1: Information of data on May 15.

Training data	Testing data	Attributes
1,136,585	378,862	258

5.2. Feature Extraction. The traffic features such as flow parameters, date types, and environmental situations will be described in detail in the following part. Also the most relevant features for high prediction precision are selected by sensitivity analysis.

5.2.1. Feature Description. Actually, the roads are divided into minor road segments as fundamental components for assessment and forecast. The road in Nanning city consists of 18, 041 minor segments. Figure 2(a) illustrates the road segments in Nanning city, and Figure 2(b) refers to the enlarged vision of road segments.

For traffic flow prediction, feature extraction is regarded as a significant part and will impact the forecasting outcomes directly. It is essential to decide appropriate input variables for precise predictions. Apparently, some factors affect the traffic condition of roads, for example, whether there are traffic lights or not, where the road is, and if a school or a shopping center is located nearby. However, it is difficult to obtain these characteristics automatically. Consequently, these characteristics will not be taken into consideration. We chose the following characteristics.

(i) Present time t: categorical, 06:05, 06:10, ..., 21:55. 191 in total.

Determining how to forecast the traffic flow accurately within a 5 min time interval is critical for ITS because the cycle of transportation induction is usually 5 minutes.

(ii) Date type: categorical, $1, 2, 3, \ldots, 7$, 7 in total.

(iii) Weather condition: categorical, $1, 2, 3, \ldots, 8$, 8 in total; details are shown in Table 2.

(iv) Traffic flow of the last time interval: continuous, normalization data varying from 0 to 1.

$$\text{Volume}^* = \frac{\text{Volume} - \text{Volume}_{\min}}{\text{Volume}_{\max} - \text{Volume}_{\min}}, \tag{21}$$

$$\text{Volume}_{\max} = \frac{1}{n}\sum_{j=1}^{n}\text{Volume}_{\max,j}, \tag{22}$$

Table 2: Quantificational disposal of weather factors.

Weather condition	Sunny	Cloudy	Windy	Foggy	Rainy	Thunder storm	Snowy	Snow storm
Quantificational value	8	7	6	5	4	3	2	1

$$\text{Volume}_{\min} = \frac{1}{n}\sum_{j=1}^{n}\text{Volume}_{\min,j}, \tag{23}$$

where $\text{Volume}_{\max,j}$ represents the maximum flow volume of the jth day and $\text{Volume}_{\min,j}$ is the minimum flow volume of the jth day. In addition, n donates the amount of the sample days.

We define $x_{i,j}$ $(0 \le x_{i,j} < 100)$ as the traffic flow at jth time interval and on the ith road (is collected every 5 minutes and 191 indexes per day totally, from 6:00 to 22:00). For the specific ith road, $[x_{i,0}, x_{i,1}, x_{i,2}, \ldots, x_{i,190}]^T$ is the vector of traffic flow on one day. Thus vectors of all roads are able to be described as follows:

$$\begin{aligned} x_0 &= [x_{0,0}, x_{0,1}, x_{0,2}, \ldots, x_{0,190}]^T \\ x_1 &= [x_{1,0}, x_{1,1}, x_{1,2}, \ldots, x_{1,190}]^T \\ &\vdots \\ x_n &= [x_{n,0}, x_{n,1}, x_{n,2}, \ldots, x_{n,190}]^T. \end{aligned} \tag{24}$$

(v) Road logical region: categorical, $1, 2, 3, \ldots, 50$, 50 in total.

Physical region, which partly represents the circumstance of roads, can play an important role in road congestion. For example, roads nearby the school could usually stay in a good situation apart from the time when students head for school on mornings and when the students go back home on afternoons. At the same time, traffic situation possesses strong space-time periodicity and area characteristic is supposed to be taken into consideration. Regrettably, the information of physical region cannot be automatically obtained by ITS, and it is easy to transform, particularly in the situation where a new road is in construction when ITS runs. In this paper, a logical region is utilized rather than physical region to demonstrate roads' circumstance characteristics because of the inaccuracy of the area labeled by individuals.

In this paper, the K-means algorithm is used to cluster the logical regions into k clusters. The value of k is set to be 50 in this paper.

(vi) Road type: categorical, 1, 2, 3, 3 in total.

This eigenvalue is defined with the numerical value. From branch way to major road the numerical values are 1, 2, and 3 in order. The higher the road grade, the higher the road's standard speed and maximum speed.

(vii) Number of lanes.

The number of lanes represents the road capacity. Under the circumstances of same traffic flow, the more lanes there are, the higher the speed is.

(viii) The average speed of all cars on the road section.

Firstly, we calculate the average speed of every car in the interval of ΔT on the road section. Then the average speed of all cars is attained. In the interval of ΔT, the velocity measurement sites of the floating car r on the terminal road section are distributed as shown in the figure below.

Table 3: Five levels of the speed of floating cars.

Speed grades	1	2	3	4	5
Range (km/h)	<15	15~35	35~55	55~75	>75

Sequence $\{t_0, t_1, \ldots t_p\}$ and sequence $\{u_0, u_1, \ldots u_p\}$ are floating car's time sequence and speed sequence on the road section. The floating car's driving distance S_r is defined by

$$\begin{aligned} S_r &= \int_{t_0}^{t_p} u\,dt \\ &\approx u_0\left(\frac{t_1 - t_0}{2}\right) + u_p\left(\frac{t_p - t_{p-1}}{2}\right) \\ &\quad + \sum_{i=1}^{p-1} u_i\left(\frac{t_{i+1} - t_{i-1}}{2}\right). \end{aligned} \tag{25}$$

Therefore, the floating car r's average speed can be represented as $U_r = S_r/(t_p - t_0)$.

If the number of floating cars on the road section at the moment is n, the average speed of all cars on the road section can be written as $\widetilde{U} = (\sum_1^n U_r)/n$.

(ix) The speed distribution of all cars on the road section.

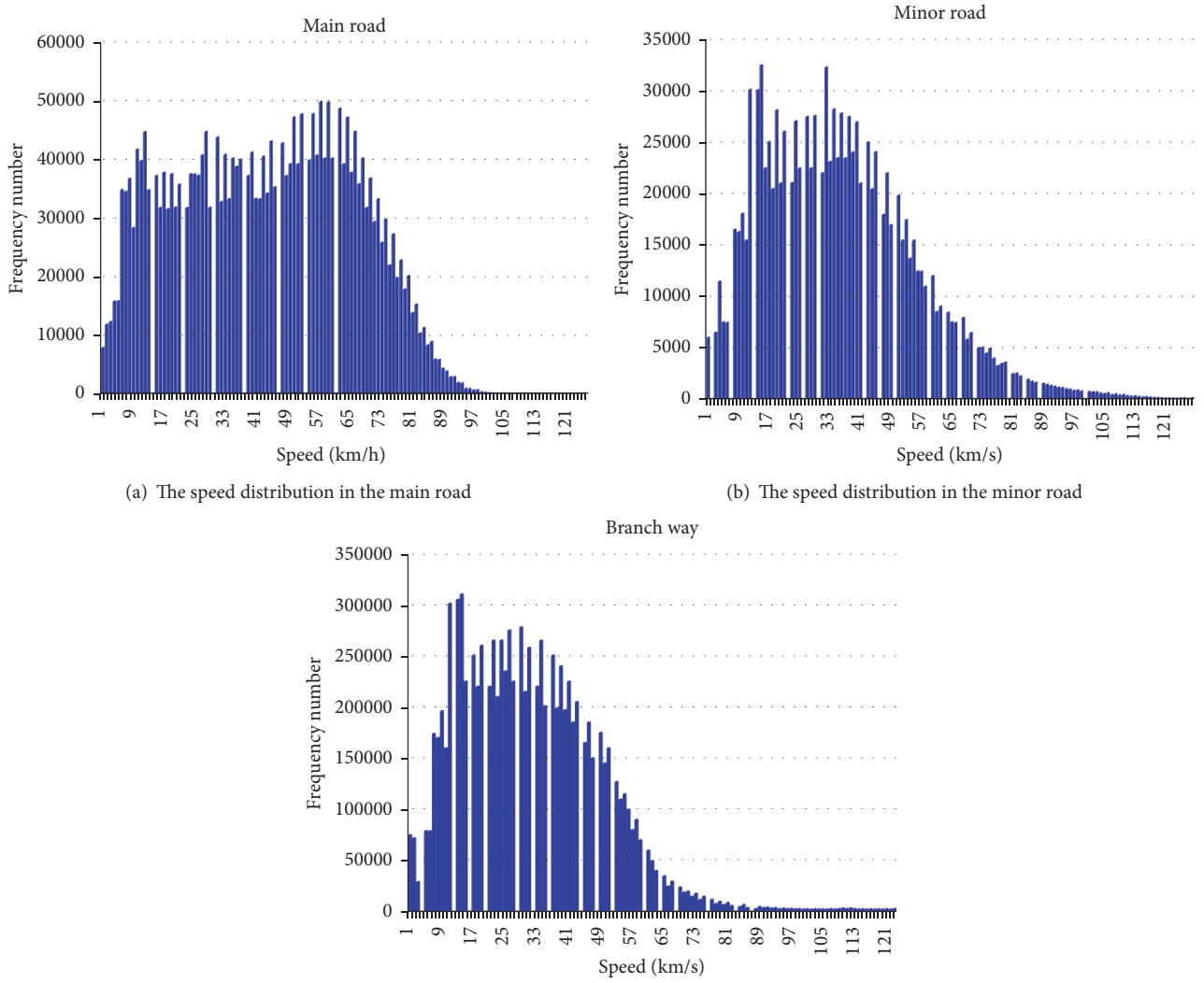
The car speed is distinguished into different levels and the histogram is used to represent the distribution of speed data of all floating cars on the read section. The standard of division is based on the speed distribution of floating cars in the city (Figure 3). Car speed data is mainly no more than 75 km/h. Therefore, the car speed is distributed into 5 grades as shown in Table 3.

(x) The average stopping time of all cars on the road section.

When a car's speed is below 5 km/h, it is identified as a stopped car. The floating car r's stopping time can be represented as $T_r = \sum_{i=0}^{p-1}(t_{I+1} - t_i)(u_i < 5)$.

Thus the average stopping time of all cars on the road section can be written as $\widetilde{T} = (\sum_1^n T_r)/n$.

5.2.2. Input Dimension Reduction. Determining the most relevant features is an issue of research in itself. Taking all the factors into consideration will lead to computation complexity, dimensionality course, and overtraining. There are three sorts of methodologies for decreasing the input



(a) The speed distribution in the main road

(b) The speed distribution in the minor road

(c) The speed distribution in the Branch way

Figure 3: The speed distribution of floating cars in the city on 23rd January 2015.

dimensionality including NNs sensitivity analysis, transformation, and correlation among features [23]. The correlation approaches lead to their high computational complexity. The problem with the transformation approaches is that the features of the original input no longer exist. To address the problems, we selected the sensitivity analysis method as the feature dimension reduction tool in this paper. Sensitivity analysis concentrates on how a network's output is influenced by its input perturbations. Irrelevant inputs are found and eliminated to reduce data collection cost and improve the network's performance. The underlying relationships between the inputs and outputs are found after sensitivity analysis. Sensitivity of an output o_k corresponding to its input x_i is especially

$$S_{ki} = \frac{\partial o_k}{\partial x_i}. \tag{26}$$

Inputs which affect outputs most significantly are determined by examination of the mean square average (MSA) sensitivity matrix $S_{ki,\text{avg}}$. A minor value of $S_{ki,\text{avg}}$ compared with others suggests that, as to the special kth output for the network, the ith input cannot remarkably donate per average to the output k, and consequently could be probably negligible. The one input varies up or below the mean while all other inputs were kept fixed at their respective means. This process was the repeated for each input variable.

The abbreviations of the chosen characteristics are shown in Table 4. And outcomes of sensitivity analysis are presented in Figure 4 in which seven most significant input parameters are date of week (DW), time of day (TD), traffic density (DEN), speed (SP), flow (FL), weather condition (WC), and road logical region (RLR).

When either one or more inputs possess comparably minor sensitivity compared to others, the NNs' dimension could be decreased by eliminating them; meanwhile a smaller-scale NNs are able to be retrained under most circumstances successfully.

Table 4: List of input variables.

Number	Variables	Abbreviation
1	Day of week	DW
2	Time of day	TD
3	Traffic density	DEN
4	Speed	SP
5	Flow	FL
6	Weather condition	WC
7	Road logical region	RLR
8	Road type	RT
9	Number of lanes	NL
10	Speed distribution	SD
11	Stopping time	ST
12	Number of vehicles	NV
13	Number of traffic lights	NTL

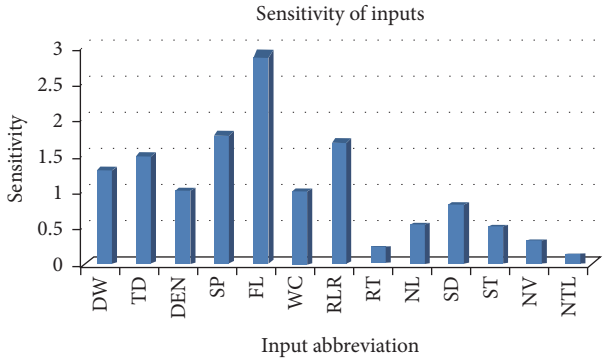


Figure 4: Sensitivity analysis of input variables.

5.3. Determination of Model's Parameters and Its Application to Traffic Congestion Prediction

(a) Optimization of Multikernel Model's Parameters. In (7), it can be observed that the prediction precision of MKELM is mainly affected by the four variables λ, g, and p of MKELM and the punish parameter C. The four variables need to be selected with optimizing method. Then the problem of how to choose proper parameters of MKELM can be converted into optimizing four variables: $x = [g, \lambda, p, C]$. The vector x which has the best fitness is chosen in MKELM. In our study, QPSO is used to obtain optimized parameters. The function for fitness calculation is

$$f(x) = \sqrt{\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} [t_i - y(x_i)]^2}, \tag{27}$$

where $y(x_i)$ is the predicted value, t_i is the target, and m is the amount of training data.

The main procedures are described as follows:

(1) Preprocesses the original samples and divide them into training and testing samples.

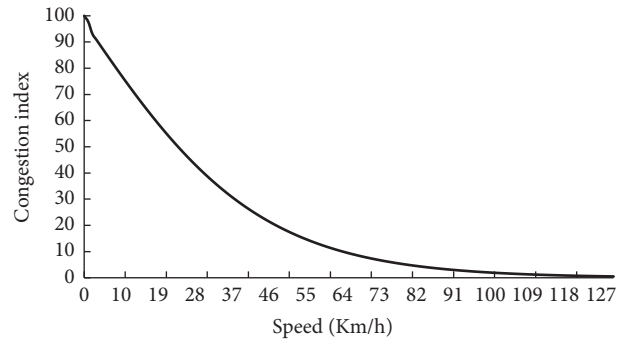


Figure 5: Graph of (21).

(2) Initialize the particles position $[x_{i1}, x_{i2}, \ldots, x_{id}]$. The range of parameters is $x = \{0, 1; 0.04, 100; 0.5, 1.3; 100, 300\}$.

(3) Calculate every particle's fitness of every particle and renew.

(4) Compare and update the top fitness value.

(5) If the ending conditions are met, the optimal parameters of MKELM are obtained.

(b) Traffic Congestion Prediction Using Optimized MKELM Model. Driving speed refers to a significant indicator for judging the road situation. It is simple to obtain average speed of vehicles via gathering 5-minute speed. We primarily utilize speed for determining the Traffic Congestion Indicator (TCI). Equation (28) is the function applied to assess transport situation while its plot is illustrated in Figure 5.

$$y(x) = 100 - 200 \cdot \left(\frac{1}{1 + e^{-0.46x}} - \frac{1}{2} \right), \tag{28}$$

where x refers to the speed of road segment (unit: km/h), whereas $y(x)$ refers to TCI. It can be seen that road situation becomes worst if the average speed turns to be zero, and consequently TCI is equal to 100. Otherwise, traffic situation becomes better as the speed increases. Thus, when speed inclines to be infinite, the value of TCI inclines to be 0, where $y = 0$ is the horizontal asymptote of $y(x)$.

The output of assessment refers to a continuous value which changes between 0 and 100. The feasible result indicates that the speed-TCI model turns out to be rational and be able to represent traffic status appropriately. The assessment outcomes satisfy the true condition as illustrated in Figure 6. In the map, green represents smooth traffic, yellow shows average condition, and red means the road is congested. Seen from the image taken by surveillance cameras, the traffic evaluation accurately reflects the road traffic congestion at that time. The system creates assessment output every 5 minutes so as to portray the present urban traffic situation for every road segment. These data gradually become historical. The historical assessment data are able to be applied to train forecasting model, and it is clear that a large number of training data are generated every day. In practice, almost 1,500,000 valuable data are created every day. Large data call for a proper model with fast speed. It is known that MKELM

(a) (b)

Figure 6: The comparison of camera results and evaluation results shows that the evaluation model is reasonable.

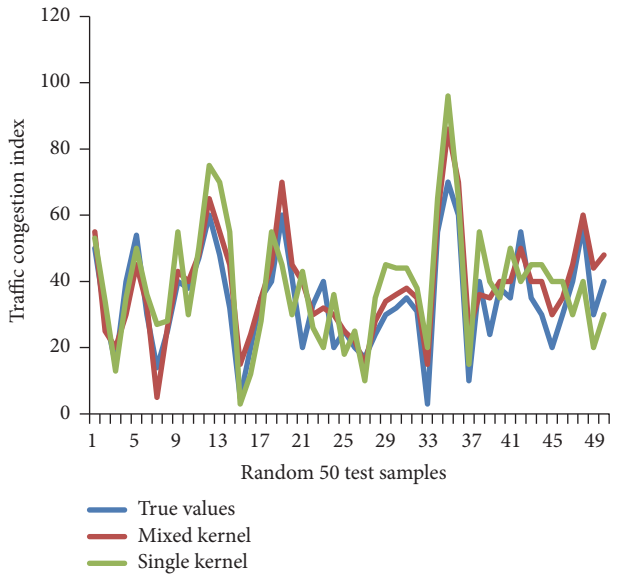


Figure 7: Part of regression prediction shows that mixed kernel model has a great fitting result.

algorithm inclines to offer good generalization feature at a fast learning rate.

With the optimal parameters of MKELM, we just draw target values of 50 testing data selected randomly, single Kernel-ELM predictions, and MKELM, respectively. In standard single kernel model, the radial basis function (RBF) is used of which the parameter γ and the cost coefficient C were selected in $\{20, 21, \ldots, 210\}$. Figure 7 shows that the multikernel model of KELM has more accurate prediction results on traffic training data than standard one-kernel ELM. The difference between the true values and the ones predicted by multikernel model are very small for most of the cases.

5.4. Study Results and Discussions. In this experiment, the kernel function of KELM is multikernel described in (10). For the parameters of QPSO, the population scale as well as the amount of repetitions is 200; the values of $\delta_{\min}$ and $\delta_{\max}$ are set to be 0.5 and 0.9.

The main target of the proposed QPSO-KELM approach is to originate dependable PIs. Moreover, traffic management needs valuable information along with higher confidence levels. Consequently, it is supposed to be more virtually significant to acquire high-confidence-level PIs for satisfying the requirements of traffic management. Diverse levels of PINC $100(1 - \alpha)\%$ from 80% to 95% are taken into account in this research.

Table 5: Comparison of MKELM-CWC and MKELM-WDR.

PINC%	Methods	PICP (%)	PINAW (%)
95%	MKELM-CWC	95.86	28.97
	MKELM-WDR	95.37	25.09
90%	MKELM-CWC	91.04	22.61
	MKELM-WDR	90.49	21.45
80%	MKELM-CWC	80.93	14.78
	MKELM-WDR	80.46	13.97

(1) Comparison of MKELM-Based Model Using CWC and WDR. To construct the PIs for traffic flow, whether the rule CWC or the rule WDR should be used in the MKELM-QPSO model needs to be decided firstly.

For the traffic flow, the dependability indicator PICP as well as the sharpness indicator PINAW of the constructed PIs at three different confidence levels 80%, 90%, and 95% is summarized in Table 5. Also the results using CWC and WDR are, respectively, summarized in Table 5. It is shown that, at every confidence level, the PICP values of WDR are closer to the confidence levels than the values of CWC. At the same time, CWC has a much higher sharpness than WDR which provides much more accurate PIs. The comparisons of CWC and WDR in Table 5 show that WDR generates the higher dependability and sharpness of the PIs. In the rest of the paper, the criterion WDR is applied to get the optimal PIs because of its better performance shown in Table 5.

To achieve a better visual effect, we draw the real values and forecasting PIs, respectively, of 100 testing samples which are selected randomly. The constructed optimal PIs at different confidence levels 80%, 90%, and 95% are shown in Figure 8. Most of the targets have been limited within the upper and the lower bound. For different confidence levels, most of the measured traffic flows are limited in the PIs established by the proposed method, which means that the

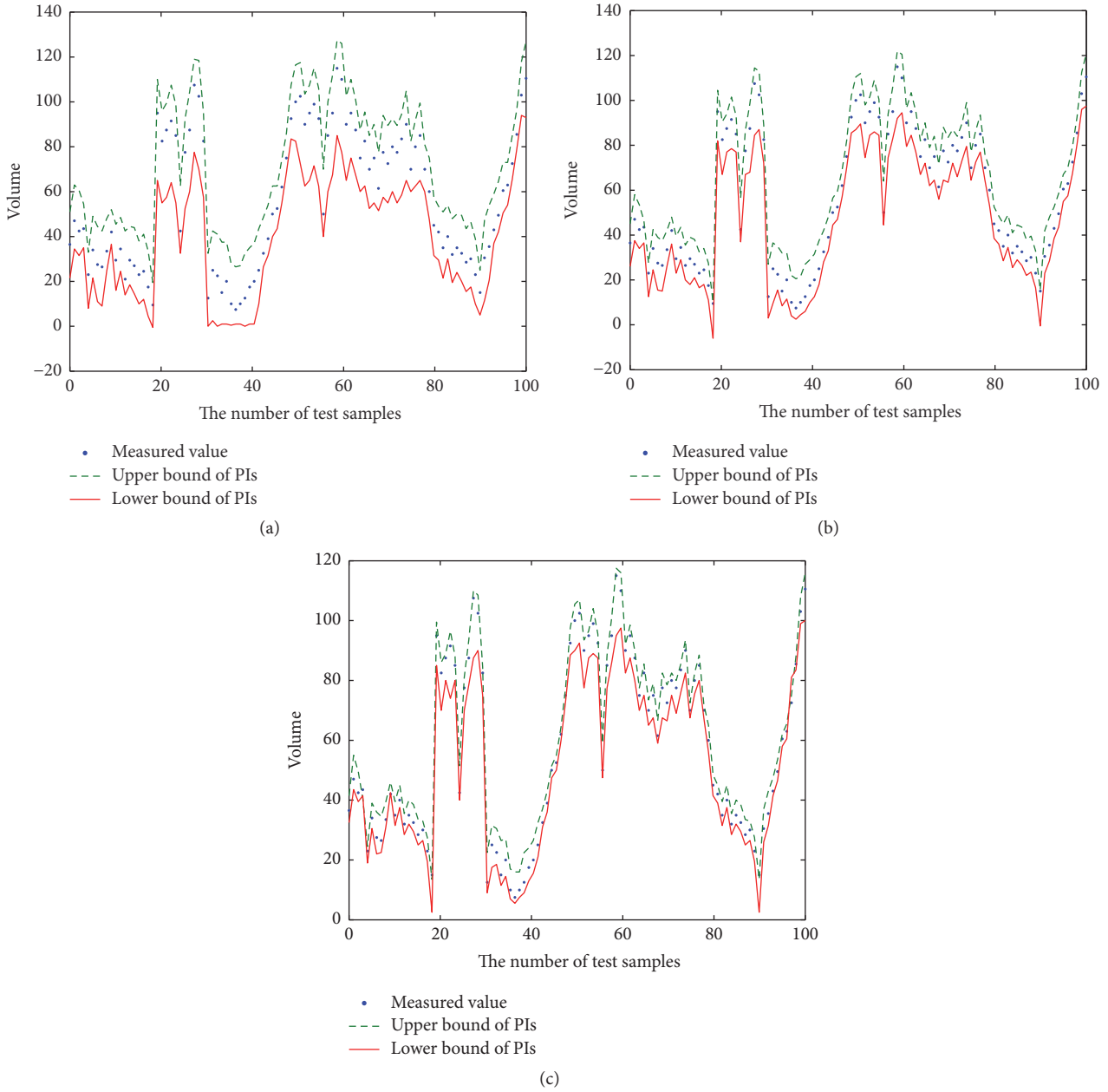


Figure 8: Overall optimal PIs by proposed QPSO-KELM approaches. (a) 95% confidence level case. (b) 90% confidence level case. (c) 80% confidence level case.

remarkable performance will satisfy the requirement of traffic management.

Also the nonstationary and nonlinearity characteristics of traffic flow are clearly demonstrated in these graphs.

(2) Comparison of Different Methods for Constructing the PIs. For the sake of further demonstrating the effectiveness of the QPSO-MKELM method, different intelligence search and learning algorithms are compared using the same dataset.

As to the PIs dependability test, corresponding PICPs and PINAWs are demonstrated in Table 6. The corresponding PICPs and PINAWs outcomes of diverse methods are described in Table 6. At different confidence levels, the dependability indicator PICP and the sharpness indicator PINAW are summarized in Table 6, which are acquired by QPSO-MKELM, PSO-MKELM, and GA-MKELM. From the outcomes of PICP as well as PINAW indices, it is shown that QPSO-KELM methods offers far more precise PIs combined dependability (indicated by higher values of PICP) as well as sharpness (reflected by lower values of PINAW) on the three various confidence levels.

The SVM, ELM, standard KELM, and MKELM are compared on the basis of the training time, PICP, and PINAW at the same confidence level 90% in Table 7. Although the

Table 6: Comparison of QPSO-KELM, PSO-KELM, and GA-KELM.

PINC%	QPSO-MKELM		PSO-MKELM		GA-MKELM	
	PICP%	PINAW%	PICP%	PINAW%	PICP%	PINAW%
95	95.37	25.09	94.32	30.54	93.86	31.07
90	90.49	21.45	89.54	29.17	88.93	29.66
80	80.46	13.97	79.37	22.89	78.86	23.85

Table 7: Comparison of SVM, ELM, KELM, and MKELM at confidence level 90%.

Algorithms	Training times (s)	PICP%	PINAW%
SVM	19204.86	90.56	23.43
ELM	300.39	87.34	37.57
KELM	811.05	90.84	21.76
MKELM	860.83	90.49	21.45

MKELM cannot compete with the ELM in terms of the training time, the MKELM provides much more accurate PIs than the ELM. We can also see that the SVM almost provides the same PICP and PINAW as the MKELM, but the training time for the SVM is more than 20000% of that for the KELM. Overall, the MKELM outperforms the SVM, the ELM, and the standard KELM. Thus, it is rational to draw a conclusion that the proposed method is an efficient probabilistic forecasting approach for traffic flow.

6. Conclusions

In order to develop an effective probabilistic forecasting method for traffic flow, a novel method on the basis of QPSO-MKELM has been described to establish the reliable PIs. MKELM has been developed to establish the reliable PIs, and the parameters of MKELM are optimized by using QPSO. The seven features including date of week, time of day, traffic density, speed, flow, weather condition, and road logical region are selected as inputs of KELM by sensitivity analysis. The experimental results have shown that QPSO-MKELM is an effective method to establish the optimal PIs. Moreover, the proposed method can offer far more precise PIs that combined higher dependability and sharpness at different confidence levels than other methods. Additionally, successful utilization in practical traffic flow prediction shows that QPSO-MKELM is an effective probabilistic forecasting method. In the current paper, only limited traffic flows are taken into consideration; other parameters such as accidents, traffic jams, or seasonal variation have not been discussed. In future study, more possible conditions will be included in a longer period of time.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 61272357, no. 61300074, and no. 61572075) and the National Key Research and Development Program of China (Grant no. 2016YFB0700502 and Grant no. 2016YFB1001404). The authors thank Dr. Ruoyi Liu for revising their English expressions.

References

[1] D. Ngoduy and A. Sumalee, "Adaptive estimation of noise covariance matrices in unscented Kalman filter for multiclass traffic flow model," *Transportation Research Record*, vol. 2188, pp. 119–130, 2010.

[2] R. K. Oswald, W. T. Scherer, and L. B. Smith, "Forecasting using approximate nearest neighbor nonparametric regression," Traffic Flow, 2000.

[3] D. Park, L. R. Rilett, and G. Han, "Forecasting multiple-period freeway link travel times using neural networks with expanded input nodes," in *Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering*, 2010.

[4] B. L. Smith and M. J. Demetsky, "Short-term traffic flow prediction: neural network approach," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1453, pp. 98–104, 1994.

[5] L. I. Bin, X. I. Tao, and M.-H. Shi, "Traffic flow combined forecast model of Support Vector Machine," *Journal of Tianjin Polytechnic University*, vol. 27, no. 2, pp. 73–76, 2008.

[6] N. L. Nihan and K. O. Holmesland, "Use of the box and Jenkins time series technique in traffic forecasting," *Transportation*, vol. 9, no. 2, pp. 125–143, 1980.

[7] Y.-N. Yang and H.-P. Lu, "Short-term traffic flow combined forecasting model based on SVM," in *Proceedings of the International Conference on Computational and Information Sciences (ICCIS '10)*, pp. 262–265, IEEE, Chengdu, China, December 2010.

[8] P. Pinson and G. Kariniotakis, "Conditional prediction intervals of wind power generation," *IEEE Transactions on Power Systems*, vol. 25, no. 4, pp. 1845–1856, 2010.

[9] C. A. M. da Silva Neves, M. Roisenberg, and G. S. Neto, "A method to estimate prediction intervals for artificial neural networks that is sensitive to the noise distribution in the outputs," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '09)*, pp. 2238–2242, IEEE, Atlanta, Ga, USA, June 2009.

[10] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. W. C. Van Lint, "Prediction intervals to account for uncertainties

in travel time prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 537–547, 2011.

[11] D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proceedings of the IEEE World Congress on Computational Intelligence International Conference on Neural Networks*, vol. 1, pp. 55–60, 1994.

[12] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Probabilistic forecasting of wind power generation using extreme learning machine," *IEEE Transactions on Power Systems*, vol. 29, no. 3, pp. 1033–1044, 2014.

[13] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Lower upper bound estimation method for construction of neural network-based prediction intervals," *IEEE Transactions on Neural Networks*, vol. 22, no. 3, pp. 337–346, 2011.

[14] G.-B. Huang and C.-K. Slew, "Extreme learning machine: RBF network case," in *Proceedings of the 8th International Conference on Control, Automation, Robotics and Vision (ICARCV '04)*, pp. 1029–1036, Kunming, China, December 2004.

[15] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[16] J. Kivinen, A. J. Smola, and R. C. Williamson, *Learning with Kernels*, MIT Press, 2002.

[17] Y. Xing, X. Ban, C. Guo, and Y. Wang, "Probabilistic forecasting of traffic flow using kernel based extreme learning maching and quantum-bahaved particle swarm optimization," in *Proceedings of the 4th International Conference on Cloud Computing and Intelligence Systems (CCIS '16)*, pp. 205–209, Beijing, China, August 2016.

[18] H. Q. Wang, F. C. Sun, Y. N. Cai, N. Chen, and L. G. Ding, "On multiple kernel learning methods," *Acta Automatica Sinica*, vol. 36, no. 8, pp. 1037–1050, 2010.

[19] J. Sun, B. Feng, and W. Xu, "Particle swarm optimization with particles having quantum behavior," in *Proceedings of the Congress on Evolutionary Computation (CEC '04)*, pp. 1571–1580, Portland, Ore, USA, June 2004.

[20] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[21] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 209, pp. 415–446, 1909.

[22] F. van den Bergh and A. P. Engelbrecht, "A study of particle swarm optimization particle trajectories," *Information Sciences*, vol. 176, no. 8, pp. 937–971, 2006.

[23] D. S. Yeung, I. Cloete, D. Shi, and W. Ng, *Sensitivity Analysis for Neural Networks*, Natural Computing Series, Springer, Berlin, Germany, 2009.

Hindawi
Scientific Programming
Volume 2017, Article ID 8502691, 11 pages
http://dx.doi.org/10.1155/2017/8502691

Research Article

Rigid Body Sampling and Individual Time Stepping for Rigid-Fluid Coupling of Fluid Simulation

Xiaokun Wang, Xiaojuan Ban, Yalan Zhang, and Xu Liu

University of Science and Technology Beijing, Beijing 100083, China

Correspondence should be addressed to Xiaojuan Ban; banxj@ustb.edu.cn

Received 4 August 2016; Accepted 14 February 2017; Published 2 March 2017

Academic Editor: Wenbing Zhao

In this paper, we propose an efficient and simple rigid-fluid coupling scheme with scientific programming algorithms for particle-based fluid simulation and three-dimensional visualization. Our approach samples the surface of rigid bodies with boundary particles that interact with fluids. It contains two procedures, that is, surface sampling and sampling relaxation, which insures uniform distribution of particles with less iterations. Furthermore, we present a rigid-fluid coupling scheme integrating individual time stepping to rigid-fluid coupling, which gains an obvious speedup compared to previous method. The experimental results demonstrate the effectiveness of our approach.

1. Introduction

Physically based fluid simulation is a popular issue in computer graphics and virtual reality while having a huge research and application demand in three-dimensional visualization and human- computer interactions. More realistic effects and higher simulation efficiency are the main goals; thus, reasonable, efficient, and scientific programming algorithms are needed to design and implement the animation. Two major schemes are employed for animating fluids: the grid-based Eulerian approach and particle-based Lagrangian approach. Eulerian method is particularly suited to simulate large volumes fluid, while being restricted by time step and computing time for small scale features. In contrast, Lagrangian method is suitable for capturing small scale effects such as spindrift and droplet. Among various particle-based approaches, Smoothed Particle Hydrodynamics (SPH) is the most popular method for simulating fluid due to computational simplicity and efficiency.

In reality, rigid-fluid interaction widely exists in many scenarios. As a result, the interesting fluid behaviors emerge once rigid objects are added to fluid simulation. While interaction between particle-based fluids and rigid objects seems to be straightforward, there are still several issues not well resolved. For one thing, rigid bodies must be sampled to particles in order to interact with particle-based fluids, but only a few rigid boundary sampling methods can be directly employed in rigid-fluid coupling simulation. For another thing, the computational expenses of rigid-fluid coupling are considerable. To deal with the increasing demands for more detailed fluids and high efficiency, we present rigid sampling and individual time stepping for rigid-fluid coupling and design a practical and easy rigid-fluid animation simulation scheme with our scientific programming algorithms.

2. Related Work

Desbrun and Gascuel introduced SPH to computer graphics for simulating deformable objects [1]. SPH became popular in computer graphics for various fluid phenomena. Monaghan addressed simulating free surface flows with SPH [2] that serves as a basis for SPH fluid simulation. Muller et al. [3] proposed using gas state equation with surface tension and viscosity forces for fluid simulation, which also bring compressibility issue. Becker and Teschner [4] proposed WCSPH employing Tait equation to reduce compressibility. It significantly increased realistic effects but the efficiency is limited by time step. As incompressibility expenses computation time, many improved algorithms were addressed to enhance the efficiency. Solenthaler and Pajarola [5] presented PCISPH using a prediction-correction scheme to determine

particle's pressure and large time steps which significantly improved efficiency comparing to WCSPH. Another similar method that ensures incompressibility by iterative process is LPSPH [6]. Afterwards, Ihmsen et al. addressed a more efficient approach IISPH [7]. It carefully constructed pressure Poisson equation and solved the linear system using Relaxed Jacobi, which has a great improvement in stability as well as convergence speed and is particularly suitable for large-scale scene. Recently, a promising approach for impressible SPH has been proposed by Bender and Koschier [8]. It combines two pressure solvers which enforce low volume compression and a divergence-free velocity field and permits large time steps that yields a considerable performance gain.

Besides, adaptive method, either spatial resolution or time discretization, is another way to promote efficiency. They allot computing resources to regions with complex flow behavior. Space adaptive methods [9–11] adaptively sample particle and employ less particles to produce similar details. Large particles are divided into small particles if high resolution is needed, and vice versa. However, difficulties exist in reproducing quantity when refining particles, and neighbor searching is usually the bottleneck. As an alternate to adaptive spatial discretization, the time domain can be adaptively sampled as well. Globally adaptive time stepping methods [1, 12, 13] employ a single time step to adjust each step for all particles in consideration of CFL condition. Though each particle has the current smallest time step, it is not the most efficient way. Locally adaptive time stepping methods [9, 14, 15] use different time steps for particles. Desbrun and Cani proposed that each particle evaluates forces depending on its current individual time step [9]. He et al. [16] adopt this idea and implement stable simulation of stiff fluids. It updates position, velocity, and density for active particles and interpolates for inactive particles. In this paper, we integrate it to rigid-fluid coupling to reduce the computational time.

For boundary handling in SPH fluid simulation, distance-based penalty methods were commonly employed [17–19]. Nonetheless, these methods require large penalty forces which limit the time step and make particles stick to the boundary on account of lacking fluid neighbors. Frozen or ghost particles based models are used to solve the problem of sticking particles [20]. In order to avoid penetration, more than one layer of frozen particles were used [21], or the positions of penetrating particles should be corrected [13]. However, handling two-way interaction is troublesome since the elevated density near boundary in one phase affects particles in the other phase. For this reason and for the lack of fluid neighbors, Ghost SPH [22] solved those problem using a narrow layer of ghost particles and Akinci et al. employed boundary particles to correct the calculation of fluid density [23]. Because Ghost SPH is more time consuming, we use Akinci's boundary handling method which is simple and easy to achieve in this paper.

For rigid-fluid coupling, several approaches were presented up to now. In [24], the fluid is represented as rigid spheres and switching impulses with rigid bodies. In [25, 26], the pressure at the boundary is taken into consideration for two-way fluid-rigid coupling. Then, Oh et al. proposed an impulse-based scheme for two-way coupling of SPH fluids with rigid bodies [27]. Becker et al. presented direct forcing for rigid-fluid coupling [28] which employs a prediction-correction scheme to enforce particle positions and velocities to specific values. Akinci et al. presented a momentum-conserving two-way coupling approach based on hydrodynamic forces that use boundary particles to sample the surface of rigid bodies [23]. We present a rigid-fluid coupling scheme by integrating individual time stepping to Akinci's boundary handling method that gains an obvious speedup [29].

Rigid bodies sampling includes particle-based methods and polygonization-based methods. Turk repelled particles on surfaces to get a uniform sample [30] and also simplified a polygonization through reducing the number of polygons [31]. Witkin and Heckbert employed local repulsion to make particles spread uniform [32]. Nehab and Shilane [33] presented algorithm of stratified point sampling. Cook addressed stochastic sampling of Poisson disk distributions with blue noise [34]. Blue sampling has the ability to generate random points and get uniform distribution of sampling points set. Therefore, the following sampling methods always have blue noise characteristics. Corsini et al. sampled triangular meshes with blue noise properties [35]. Dunbar and Humphreys [36] modified Poisson disk sample using a spatial data structure. Bridson [37] simplified Dunbar's approach with rejection sampling and extending it to higher dimensions. Then, Schechter [22] modified Bridson's approach and employed it to Ghost SPH. Inspired by Schechter's approach, we address sampling method improved by SPH equation that is more efficient and easy to implement.

3. Particle-Based Fluid Simulation Framework

In particle-based fluid simulation, the forces acting on particles are derived from the Navier-Stokes equations. The conservation of mass and momentum are written as

$$\begin{aligned} \frac{d\rho_i}{dt} &= -\rho_i \nabla \cdot \mathbf{v}_i, \\ \rho_i \frac{D\mathbf{v}_i}{Dt} &= -\nabla p_i + \rho_i g + \mu \nabla^2 \mathbf{v}_i, \end{aligned} \tag{1}$$

where $\mathbf{v}_i$ is the velocity, ρ_i is the density, p_i is the pressure, μ is the viscosity coefficient, and g is the external force field.

SPH works by obtaining approximate numerical solutions of fluid dynamics equations by expressing fluids with particles. In SPH, the representation of a field variable A at location $\mathbf{x}_i$ is defined as

$$\langle A(\mathbf{x}_i) \rangle = \sum_j m_j \frac{A_j}{\rho_j} W(\mathbf{x}_i - \mathbf{x}_j, h), \tag{2}$$

where m_j and ρ_j represent particle mass and density, respectively, $W(\mathbf{x}_i - \mathbf{x}_j, h)$ is smoothing kernel, and h is smoothing radius.

It can be easily derived from the basic SPH equation by substituting fluid density ρ into (2), that is,

$$\rho_i = \sum_j m_j W(\mathbf{x}_i - \mathbf{x}_j, h). \tag{3}$$

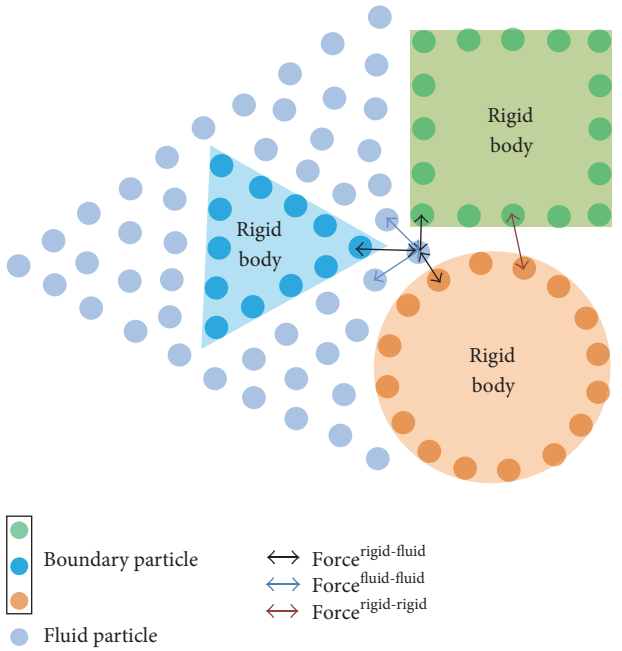


FIGURE 1: Forces between boundary particles and fluid particles.

Therefore, particles' pressure force $\mathbf{f}_i^P$ and viscous force $\mathbf{f}_i^v$ can be written as

$$\mathbf{f}_i^P = -\sum_j m_j \left(\frac{P_i}{\rho_i^2} + \frac{P_j}{\rho_j^2} \right) \nabla W_{ij},$$
$$\mathbf{f}_i^v = \mu \sum_j m_j \frac{\mathbf{v}_{ji}}{\rho_j} \nabla^2 W_{ij}. \tag{4}$$

In this paper, we use Tait equation [4] to calculate the pressure; that is, $p_i = (\rho_0 c_S^2/\gamma)((\rho_i/\rho_0)^\gamma - 1)$, where $\rho_0 = 1000$ is the rest density of the fluid, $\gamma = 7$ is stiffness parameter, and c_S is velocity of sound. We use the equation in [23] to compute viscous force.

4. Boundary Handling for Particle-Based Fluids

In rigid-fluid interaction, there are three types of forces among particles: the forces between fluid particles and boundary particles $\text{Force}^{\text{rigid-fluid}}$, the forces between fluid particles $\text{Force}^{\text{fluid-fluid}}$, the forces between boundary particles $\text{Force}^{\text{rigid-rigid}}$, which is shown in Figure 1. In our simulation, we sample rigid body to obtain boundary particles which is described in Section 5. Moreover, for the boundary handling way of rigid-fluid interaction, we implement our simulation based on the work of [23]. The following will briefly introduce boundary handling in this section.

Considering influence of boundary particles, density formula of fluid particles in (3) needs to introduce weighted summation influence of boundary particle [23], that is,

$$\rho_{f_i} = \sum_j m_{f_j} W_{ij} + \sum_k m_{b_k} W_{ik}, \tag{5}$$

where f_j and b_k denote fluid particle j and boundary particle k, respectively. The first summation calculates the affection of adjacent fluid particles, while the second summation computes the influence of adjacent boundary particles. This formula can overcome the problem of boundary defects in SPH fluid simulation to some extent.

Due to the use of boundary particle mass in (5), the density of fluid particles is incorrect or unstable when the boundary particle density is set unreasonably or unevenly distributed. Hence, consider the contribution of boundary particles to fluid particle by the volume of boundary particles is

$$\Psi_{b_i}(\rho_0) = \rho_0 V_{b_i} \tag{6}$$

where ρ_0 denotes the remaining density of fluid and V_{b_i} is the estimation value of boundary area volume of corresponding boundary particles. Applying $\Psi_{b_i}(\rho_0)$ to replace the boundary particle mass can guarantee the stability.

Thus, (5) can be rewritten as

$$\rho_{f_i} = \sum_j m_{f_j} W_{ij} + \sum_k \Psi_{b_k}(\rho_{0i}) W_{ik}. \tag{7}$$

The most important interaction between fluid particles and boundary particles is the pressure. The pressure acceleration generated by boundary particles to fluid particles can be computed as

$$\frac{d\mathbf{v}_{f_i}}{dt} = -\frac{k p_{f_i}}{\rho_{f_i}^2} \sum_k \Psi_{b_k}(\rho_{0i}) \nabla W_{ik}, \tag{8}$$

where $p_{f_i} > 0$ takes $k = 2$. When $p_{f_i} < 0$, boundary particles and fluid particles attract each other; then, we can adjust parameter k $(0 \le k \le 2)$ to realize different adsorption effects, and we choose $k = 1$ in our experiment.

To simulate the friction between fluid and container wall or the interaction of rigid body and fluid, we have to compute the friction of boundary particles with fluid particles. The friction is calculated by the artificial viscosity; that is,

$$\frac{d\mathbf{v}_{f_i}}{dt} = -\sum_k \Psi_{b_k}(\rho_{0i}) \Pi_{ik} \nabla W_{ik}, \tag{9}$$

where $\Pi_{ik} = -\nu(\mathbf{v}_{ik}{}^T \mathbf{x}_{ik}/(\mathbf{x}_{ik}{}^2 + \varepsilon h^2))$, $\nu = 2\alpha h c_s/(\rho_k + \rho_j)$.

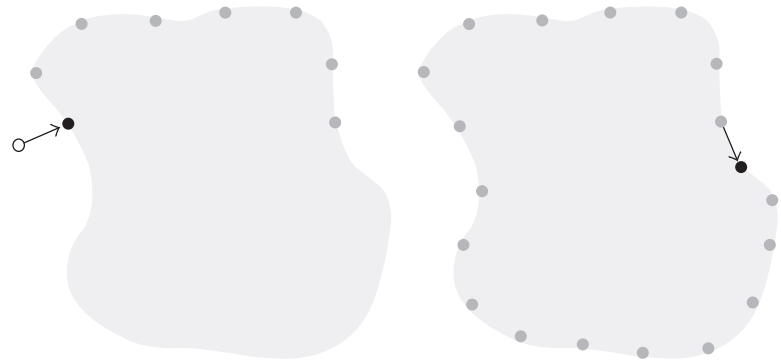
On account of (8) and (9) that listed the forces for fluid particles, we can get the forces of boundary particles using Newton's third law. The forces generated by fluid particles to boundary particles are

$$\mathbf{F}_{b_k} = \sum_i \left(\frac{k p_{f_i}}{\rho_{f_i}^2} + \Pi_{ik} \right) m_{f_i} \Psi_{b_k}(\rho_{0i}) \nabla W_{ik}, \tag{10}$$

where i denotes the fluid neighbors of boundary particle k. It is the counteracting force of (8) and (9).

For a rigid body, the total force and torque need to be calculated. This can be separately written as

$$\mathbf{F}_{\text{rigid}} = \sum_k \mathbf{F}_{b_k},$$
$$\tau_{\text{rigid}} = \sum_k \left(\mathbf{x}_k - \mathbf{x}_{\text{rigid}}^{\text{cm}} \right) \times \mathbf{F}_{b_k}, \tag{11}$$



(a) (b)

Figure 2: Surface sampling and relaxation. (a) Surface sampling. (b) Surface relaxation. Black points: newly added sample. Gray points: surface sampling points. White points: exterior points before being projected to the surface.

Input: Level set ϕ, radius r, count t, constant e
Output: Sample set S
(1) **for** all grid cells C that ϕ changes sign **do**
(2) **for** each t **do**
(3) Generate random point $\mathbf{p}$ in C
(4) Project $\mathbf{p}$ to surface of ϕ
(5) **if** $\mathbf{p}$ satisfies Poisson Disk criterion in S **then**
(6) $S \leftarrow S \cup \{\mathbf{p}\}$
(7) Break
(8) **if** no point was found in C **then**
(9) Continue
(10) **while** new samples are found **do**
(11) Generate random tangential direction $\mathbf{d}$ of surface at $\mathbf{p}$
(12) $\mathbf{q} \leftarrow \mathbf{p} + \mathbf{d} \cdot e \cdot r$
(13) Project $\mathbf{q}$ to surface of ϕ
(14) if $\mathbf{q}$ satisfies the Poisson Disk criterion in S **then**
(15) $S \leftarrow S \cup \{\mathbf{q}\}$
(16) $\mathbf{p} \leftarrow \mathbf{q}$

Algorithm 1: Surface sampling.

where $\mathbf{x}_k$ is the location of boundary particle k and $\mathbf{x}^{\text{cm}}_{\text{rigid}}$ denotes the mass center of a rigid body. The total force and torque will be transmitted to the physics engine to handle the motion of rigid bodies.

5. Rigid Boundary Sampling

Rigid body sampling is the first issue in rigid-fluid coupling which we have to handle. We propose a rigid body sampling algorithm which is an extension of Poisson disk method and sampling method in [22] for rigid-fluid coupling.

For rigid objects sampling, boundary particles are used to sample the surface of rigid bodies that has several merits. For one thing, using particles permits us to get a rigid model that can handle different shapes even with complex geometry structure. For another thing, the use of boundary particles successfully alleviates sticking artifacts and makes sampling uniform.

There are two components in our sampling: surface sampling and surface relaxation. As shown in Figure 2, it first samples the surface of rigid object image, and then it improves initial sampling with surface relaxation. In order to realize the first procedure, it needs fast projection of points to the surface. Hence, level set method is employed to express surface geometry with $\phi > 0$ and $\phi < 0$ denoting exterior and interior of rigid objects, respectively, while $\phi = 0$ denotes surface of rigid objects.

After obtaining the surface geometry of signed distance function, we use surface sampling method proposed in [22] (as shown in Algorithm 1), first, searching seed points on the surface by checking each grid cell intersecting with the surface; the details are as follows: projecting random points from the cell to the surface and stopping when the point satisfies the Poisson disk criterion, which operates k attempts in a cell; when obtaining a seed sample, continuing to sample it and taking a step of size $e \cdot r$ from the previous samples along a random tangential direction $\mathbf{d}$; then projecting to the surface and checking the Poisson disk criterion again. Parameters were chosen as $k = 30$ and $e = 1.085$.

Input: sample set S, Level set ϕ, radius r, count t, constant f
Output: relaxed sample set S
(1) **for** each t **do**
(2) **for** each $p_i \in S$ **do**
(3) compute density $\rho_i(t)$, average density $\overline{\rho}(t)$
(4) compute density gradient $\nabla\rho_i(t)$
(5) $d \leftarrow r \cdot |(\rho_i(t) - \overline{\rho}(t))/\overline{\rho}(t)| \cdot f$
(6) $p^{\text{new}} \leftarrow p + d \cdot \nabla\rho_i(t)$
(7) if p^{new} outside ϕ or came from surface sample
(8) Project p^{new} to surface of ϕ
(9) if p^{new} satisfies the Poisson Disk criterion in S
(10) $p \leftarrow p^{\text{new}}$

Algorithm 2: Modified surface relaxation.

In order to optimize the position of sample points, reduce noises, and get a uniform distribution set of sampling points, we need to further improve sampling set with a surface relaxation step. Inspired by the relaxation algorithm proposed in [22] and SPH interpolation method, surface relaxation algorithm is presented in Algorithm 2. Unlike employing random testing way in [22], we compel particles to move according to the density gradient. This ensures that the particles move to sparse place, so as to insure uniform distribution of particles.

It starts with the initial particles seed in Algorithm 1 and attempts to reposition each sample through density gradient. Next it computes density $\rho_i(t)$ and density gradient $\nabla\rho_i(t)$ of each surface particles and employs deviation of density $\rho_i(t)$ and average density $\overline{\rho_i(t)}$ as a coefficient to tune distance d. Then, it employs $d \cdot \nabla\rho_i(t)$ to adjust particle locations. Surface sample candidates are projected to the surface once again and are reserved which satisfies the Poisson disk criterion. Parameter t is iterations and f is distance coefficient. According to SPH gradient formula, particle's density gradient can be written as

$$\langle \nabla\rho_i \rangle = \sum_{j=1}^{N} m_j \nabla W \left(\left| \mathbf{x}_i - \mathbf{x}_j \right|, h \right). \tag{12}$$

On the basis of signed distance field ϕ, it is fairly convenient for calculation of steps (7) and (8). If $\phi(p^{\text{new}}) \neq 0$, it means p^{new} is not on the surface. While for projecting particles to surface, we compute signed distance field gradient $\nabla\phi(p^{\text{new}})$. Projection formula is

$$p^{\text{new}} = p + d \cdot \nabla\rho_i(t) - \nabla\phi\left(p^{\text{new}}\right). \tag{13}$$

We have done the experiment in 2 dimensions contrast to the relaxation algorithm in [22]. We randomly generate 100 points in a 0.1×0.1 square shown in Figure 3. The red points represent that it does not meet the conditions of Poisson disk.

Figure 4 shows the relaxation results of Figure 3; the first row is our method and the second row is the method in [22]. The column (a) reveals the distribution of points after relaxation of two algorithms and the red points mean it does not satisfy Poisson disk condition. Each algorithm iterates 100 times, respectively, while column (b) illustrates the number

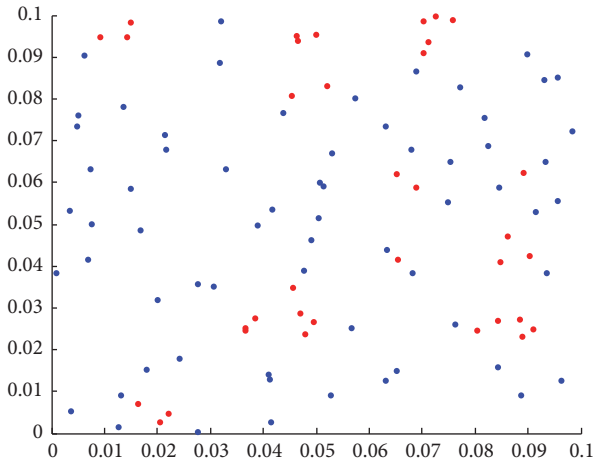


Figure 3: Initial samples in a 0.1×0.1 square.

of points that do not satisfy Poisson disk conditions for each iteration. It is obvious that our method can get a better effect with a slight concussion. Compared to the method in [22], the optimization effect is basically the same after 30 times' iteration using our method while it is the optimal results of fast Poisson disk method. In addition, our method is more efficient. In MATLAB environment, all parameters are the same as mentioned in [22]; our method takes 2.44691 s while relaxation method in [22] costs 56.44153 s for 100 iterations.

6. Individual Time Stepping for Rigid-Fluid Coupling

In this section, we propose a rigid-fluid coupling method employing individual time stepping. As a result, larger time steps can be used comparing to previous methods and the overall computation time is reduced. In particle-based fluids, particles only interact with their neighbors. So permitting particles to have different time steps is more efficient than using a global time step for all particles. The individual time stepping computes time step for each particle and updates time step asynchronously.

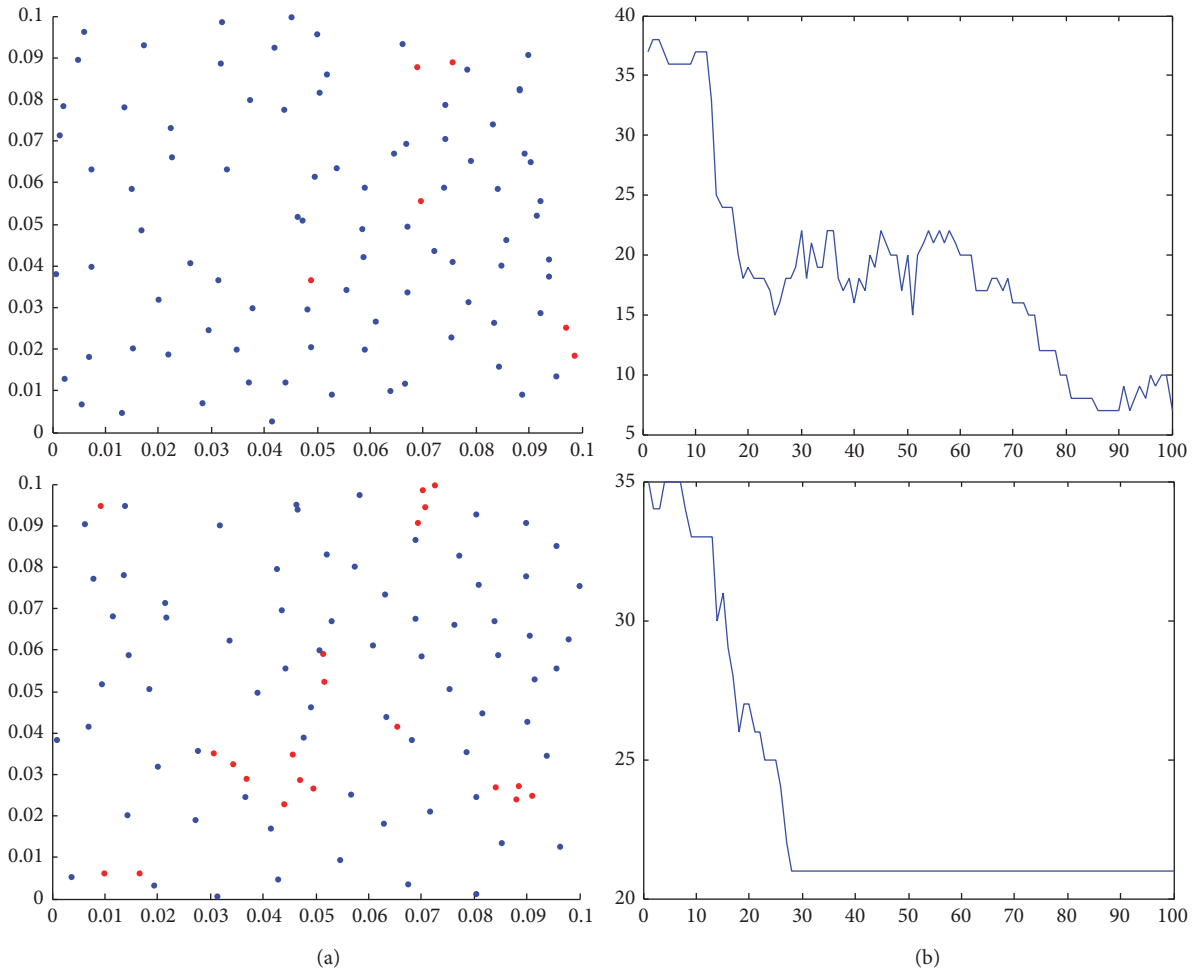


FIGURE 4: Relaxation results comparison of our method with fast Poisson disk. (a) Relaxation results. (b) The relationship between iteration times and points not conforming to the conditions.

6.1. Time Steps. For particle-based fluids, the time step must satisfies Courant-Friedrich-Levy (CFL) condition for numerical stability, that is,

$$\Delta t_{\mathrm{CFL}} \leq \lambda_v \left(\frac{h}{v_{\max}} \right), \tag{14}$$

where $v_{\max} = \max_i \|\mathbf{v}_i\|$ is the maximum velocity of particles and coefficient $\lambda_v < 1$. In addition, it also has to consider particles' maximum acceleration. Thus, the time step must also meet the condition

$$\Delta t_f \leq \lambda_f \left(\frac{h}{f_{\max}} \right), \tag{15}$$

where $f_{\max} = \max_i \|d\mathbf{v}_i/dt\|$ denotes the maximum force per unit mass of particles and $\lambda_f < 1$. In [13], $\lambda_v = 0.4$ and $\lambda_f = 0.25$ are used for PCISPH, while we use $\lambda_v = 0.1$ and $\lambda_f = 0.05$ for WCSPH. Instead of using a constant time step, we adjust time step dynamically as

$$\Delta t \leq \min \left(\Delta t_{\mathrm{CFL}}, \Delta t_f \right). \tag{16}$$

Thus, the time step for each particle i is

$$\Delta t_i = \min_j \left(\lambda_v \frac{h}{\|\mathbf{v}_j\|}, \lambda_f \sqrt{\frac{h}{\|d\mathbf{v}_j/dt\|}} \right), \tag{17}$$

where j denotes that it iterates all the neighbors.

However, the presented algorithm demands small coefficients for asynchrony, so we choose $\lambda_v = 0.05$ and $\lambda_f = 0.025$.

6.2. Asynchronism. Since each particle has individual time step, we enforce asynchronous time integration for the update. To save computing resources, the system time step is chosen as the minimized individual time step:

$$\Delta t = \min_i \left(\Delta t_i \right), \tag{18}$$

where Δt_i is computed by (17).

The particle i will be updated if it satisfies the condition

$$t_i^{\mathrm{last}} + \Delta t_i < t, \tag{19}$$

where t_i^{last} denotes the last update time of particle i and t is the system time. It means that if system time is larger than

```
(1) while animating do
(2)     select active
(3)     for each active fluid particle i do
(4)         find fluid and boundary neighbors
(5)     for each fluid particle i do
(6)         if active then
(7)             compute ρ_i(t), p_i(t)
(8)         else
(9)             interpolate ρ_i(t), p_i(t) using dρ_i(t_i^last)/dt
(10)    for each active fluid particle i do
(11)        compute dv_i(t)/dt
(12)        compute dρ_i(t)/dt
(13)        compute time step condition Δt_i' (Eq. (15))
(14)        t_i^last = t
(15)    for each boundary particle k do
(16)        compute forces (Eq. (10))
(17)    for each fluid particle i do
(18)        compute time step Δt_i = min_j(Δt_i')
(19)    Δt = min_i(Δt_i)
(20)    for each rigid body do
(21)        compute total forces, torques (Eq. (11))
(22)        pass forces and torques to physics engine
(23)        update rigid body
(24)        update boundary particles of rigid body
(25)    for each fluid particle i do
(26)        Δt' = t + Δt − t_i^last
(27)        v_i(t_i^last + Δt') = v_i(t_i^last) + Δt'(dv_i(t_i^last)/dt)
(28)        x_i(t_i^last + Δt') = x_i(t_i^last) + Δt'v_i(t_i^last + Δt')
(29)    t = t + Δt
```

Algorithm 3: Individual time stepping for rigid-fluid coupling.

the individual time step, particle i will be set as active particle and be updated.

Semi-implicit Euler integration is generally used in SPH simulation. To accommodate for asynchronism, the semi-implicit Euler integrations can be expressed as

$$\begin{aligned}\mathbf{v}_i\left(t_i^{\text{last}}+\Delta t'\right)&=\mathbf{v}_i\left(t_i^{\text{last}}\right)+\Delta t'\mathbf{a}_i\left(t_i^{\text{last}}\right),\\ \mathbf{x}_i\left(t_i^{\text{last}}+\Delta t'\right)&=\mathbf{x}_i\left(t_i^{\text{last}}\right)+\Delta t'\mathbf{v}_i\left(t_i^{\text{last}}+\Delta t'\right),\end{aligned}\tag{20}$$

where $\Delta t'$ is an independent integral time step which is different from the global time step Δt. For inactive particles, (20) is equivalent to interpolation, while, for active particles, they are semi-implicit Euler integration equations.

6.3. Algorithm. The individual time stepping for rigid-fluid coupling algorithm is shown in Algorithm 3. In this algorithm, particle i has several extra variables; that is, $d\rho_i(t)/dt$ denotes density derivative, Δt_i is time step, $\Delta t_i'$ is individual condition time step, and t_i^{last} is last updated time. In the algorithm, t is the system time and Δt is system update time step. Particle i is active if $t_i^{\text{last}}+\Delta t_i < t$.

In order to analyse the proposed algorithm, we implement the breaking dam with obstacles experiment. The setting of this experiment is shown in Table 1.

Table 1: The setting of breaking dam with obstacles.

Item	Value
Simulation domain size	12 m × 12 m × 8 m
Fluid particles	153600
Boundary particles	73585
Smoothing kernel function	Cubic splines
Smoothing radius	0.2 m
Fluid particle width	0.1 m

Table 2: Comparison of experimental results of breaking dam with obstacles.

Method	Total comp. time	Avg. Δt (avg. active pct)	Speedup
Constant steps	175 min	0.11 ms	—
Globally adaptive	41 min	0.46 ms	—
Individual stepping	27 min	0.23 ms (31%)	1.5 (6.4)

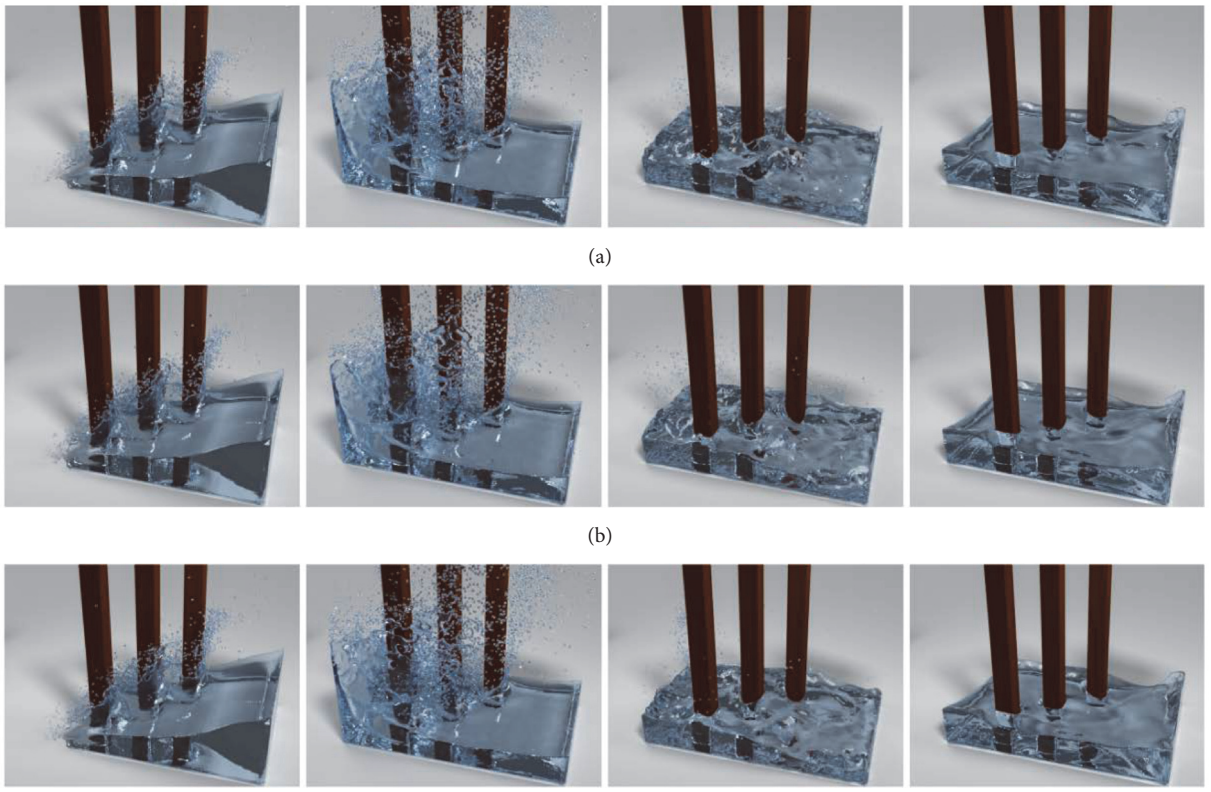
We compare individual stepping method to adaptive stepping and constant stepping method in breaking dam with obstacles scene. The rending results are shown in Figure 5 and the time statistics are listed in Table 2. From Figure 5, the fluid simulation results are almost not different using three methods, while in Table 2, we can find that our method gains 1.5 and 6.4 times speedup comparing to globally adaptive stepping method and constant stepping method, respectively. In addition, the average active particles percent of individual stepping method is 31%.

7. Implementation and Results

All the experiments in this paper are implemented on Intel 3.50 GHz CPU with 4 cores. The simulation algorithms (Algorithms 1, 2, and 3) and surface reconstruction [38, 39] are actualized with C++ language and multithreading technology. Bullet is used to simulate rigid objects while OpenMP served as parallelization. Images were rendered with Blender.

To implement fluid-rigid coupling animation efficiently and scientifically, we design a fluid-rigid coupling programming simulation scheme shown in Figure 6. We firstly initialize rigid and fluid particles configuration. Then, we do neighbor search using spatial hashing algorithm for each particle. Next, the governing equations of fluid, rigid boundary sampling, and boundary handling are solved as described in the previous sections. Then, total force and torque of rigid are calculated and provided to Bullet. After total force acting on particles is computed, particles are integrated to next time step using asynchronous update scheme introduced in Section 6.

In order to demonstrate the validity of the entire fluid-rigid coupling simulation system, we designed a scene of dropping multiple small squares into water. The setting and statistics are shown in Table 3, and the experimental results



(a)

(b)

(c)

Figure 5: Rendering results of breaking dam with obstacles. (a) Individual stepping method. (b) Globally adaptive stepping method. (c) Constant stepping method.

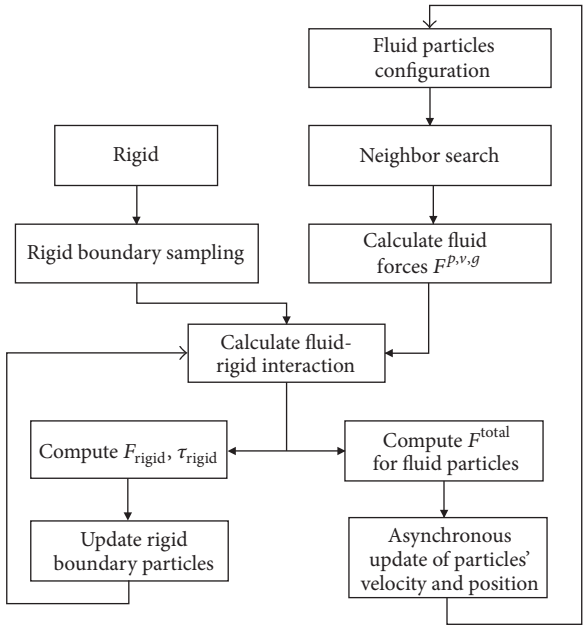


Figure 6: Flowchart of the programming scheme.

are displayed in Figure 7. It can be seen from the diagram that small squares fall into water and splash water while they are rotated and inclined by water. Finally, small squares are force balance and floating on the water.

We realize another fluid-rigid coupling experiment displayed in Figure 8. Figure 8(a) is the results in particle view while Figure 8(b) is the rendering results. The setting and statistics are illustrated in Table 3. In this scenario, the breaking dam of water hits the sculpture which is knocked down and pushed for some distances due to kinetic energy of water. The motions of sculpture are in line with expectations which proved that the simulation and calculation of fluid-rigid coupling system accord with physics laws.

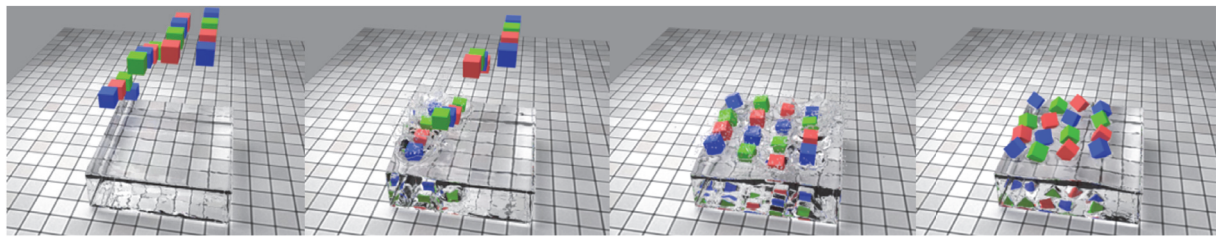
This experiment further proved that our method can implement vivid fluid-rigid coupling animation simulation system with high realistic effects. It can be expected that this animation system can be used in virtual reality domain and special effects in film and game.

8. Conclusion

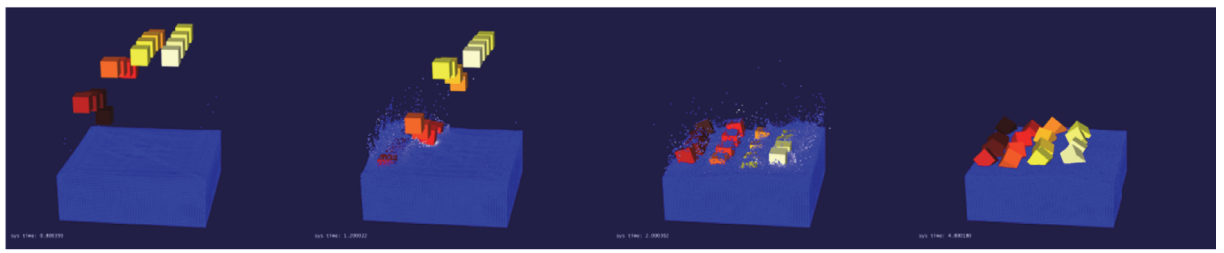
We proposed an efficient and simple rigid-fluid coupling scheme for particle-based fluid simulation. It samples rigid surface with boundary particles which are used to interact with fluids. It insures uniform distribution of particles which requires less iterations. In addition, we present an efficient rigid-fluid coupling approach combining individual time stepping with rigid-fluid coupling. Neighbors and forces of

Table 3: The setting and statistics of fluid-rigid coupling.

Item	Cubes fall into water	Water shock sculpture
Simulation domain size	12 m × 12 m × 12 m	10 m × 8 m × 5 m
Fluid particles	320 k	873 K
Boundary particles	47 K	156 K
Smoothing kernel function	Cubic splines	Cubic splines
Smoothing radius	0.05	0.1
Artificial viscosity coefficient	0.05	0.05
Surface tension coefficient	0	0
Rigid body mass	65 kg	1100 kg
Rigid body volume	0.125 m^{-3}	18.4 (4 × 2 × 2.3) m^{-3}
Fluid rest density	1000 $kg{\cdot}m^{-3}$	1000 $kg{\cdot}m^{-3}$
SPH computing time (1 frame)	0.9008 min	8.46 min
Surface reconstruction time (1 frame)	74.0 s	261 s
Rendering time (1 frame)	13.85 min	8.86 min

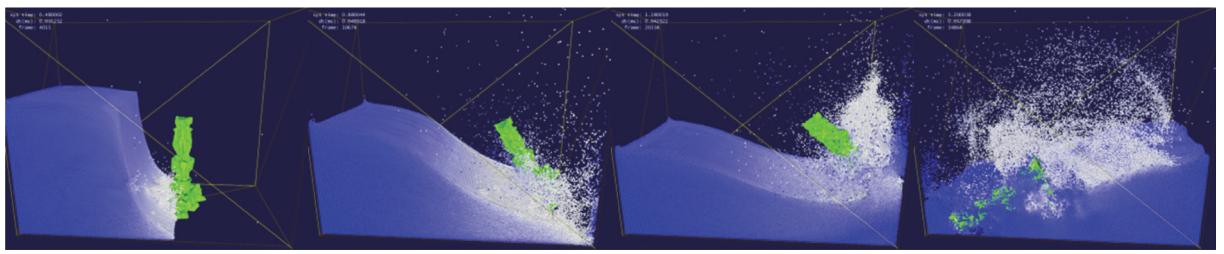


(a)

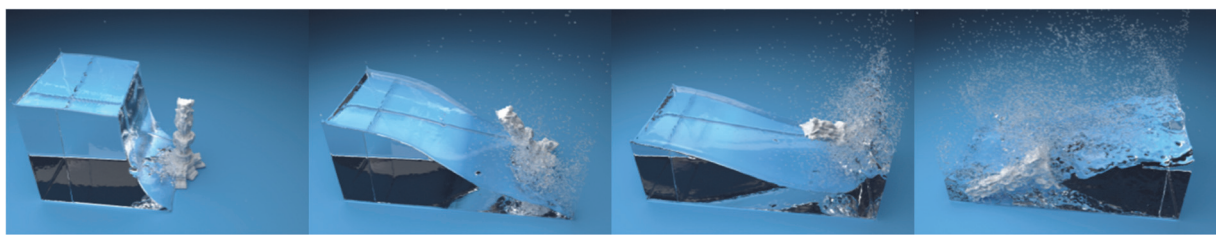


(b)

Figure 7: 16 cubes fall into water. (a) Simulation in particle view; (b) rendering results.



(a)



(b)

Figure 8: Water shock sculpture. (a) Simulation in particle view; (b) rendering results.

particles are updated only when needed, while computing resources are allocated to complex regions. It obtains an obvious speedup compared to previous methods. Besides, this scheme was integrated with rigid body coupling simulation with several scenes which has a good sense of visual reality. Overall, our method is efficient to compute while the sampling and coupling algorithm can be applied to other particle-based simulation or relevant approaches. Future work would be extending the proposed method to IISPH [7] or DFSPH [8] as well as large-scale scenarios.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by National Natural Science Foundation of China (nos. 61272357, 61300074, and 61572075).

References

[1] M. Desbrun and M. P. Gascuel, "Smoothed particles: a new paradigm for animating highly deformable bodies," in *Computer Animation and Simulation '96*, pp. 61–76, Springer Vienna, 1996.

[2] J. J. Monaghan, "Simulating free surface flows with SPH," *Journal of Computational Physics*, vol. 110, no. 2, pp. 399–406, 1994.

[3] M. Muller, D. Charypar, and M. Gross, "Particle-based fluid simulation for interactive applications," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 154–159, Eurographics Association, San Diego, Calif, USA, July 2003.

[4] M. Becker and M. Teschner, "Weakly compressible SPH for free surface flows," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 209–217, San Diego, Calif, USA, August 2007.

[5] B. Solenthaler and R. Pajarola, "Predictive-corrective incompressible SPH," *ACM Transactions on Graphics*, vol. 28, no. 3, article no. 40, 2009.

[6] X. He, N. Liu, S. Li, H. Wang, and G. Wang, "Local poisson SPH for viscous incompressible fluids," *Computer Graphics Forum*, vol. 31, no. 6, pp. 1948–1958, 2012.

[7] M. Ihmsen, J. Cornelis, B. Solenthaler, C. Horvath, and M. Teschner, "Implicit incompressible SPH," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 426–435, 2014.

[8] J. Bender and D. Koschier, "Divergence-free smoothed particle hydrodynamics," in *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '15)*, pp. 147–155, Los Angeles, Calif, USA, August 2015.

[9] M. Desbrun and M. P. Cani, *Space-Time Adaptive Simulation of Highly Deformable Substances*, INRIA, 1999.

[10] B. Adams, M. Pauly, R. Keiser, and L. J. Guibas, "Adaptively sampled particle fluids," *ACM Transactions on Graphics*, vol. 26, no. 3, Article ID 1276437, 2007.

[11] B. Solenthaler and M. Gross, "Two-scale particle simulation," *ACM Transactions on Graphics*, vol. 30, no. 4, article 81, 2011.

[12] J. J. Monaghan, "Smoothed particle hydrodynamics," *Annual Review of Astronomy and Astrophysics*, vol. 30, no. 1, pp. 543–574, 1992.

[13] M. Ihmsen, N. Akinci, M. Gissler, and M. Teschner, "Boundary handling and adaptive time-stepping for PCISPH," in *Proceedings of the Workshop on Virtual Reality Interaction and Physical Simulation*, pp. 79–88, Eurographics Association, Copenhagen, Denmark, November 2010.

[14] P. Goswami and C. Batty, "Regional time stepping for SPH," in *Eurographics 2014*, pp. 45–48, Eurographics Association, 2014.

[15] P. Goswami and R. Pajarola, "Time adaptive approximate sph," in *Proceedings of the 8th Workshop on Virtual Reality Interactions and Physical Simulations (VRIPHYS '11)*, pp. 19–28, December 2011.

[16] L. He, X. Ban, X. Liu, and X. Wang, "Individual time stepping for SPH fluids," in *Proceedings of the 36th Annual Conference of the European Association for Computer Graphics (Eurographics '15)*, Eurographics Association, Zurich, Switzerland, May 2015.

[17] M. Müller, S. Schirm, M. Teschner, B. Heidelberger, and M. Gross, "Interaction of fluids with deformable solids," *Computer Animation and Virtual Worlds*, vol. 15, no. 3-4, pp. 159–171, 2004.

[18] T. Harada, S. Koshizuka, and Y. Kawaguchi, "Smoothed particle hydrodynamics on GPUs," *Structure*, vol. 4, no. 4, pp. 671–691, 2007.

[19] J. J. Monaghan and J. B. Kajtar, "SPH particle boundary forces for arbitrary boundaries," *Computer Physics Communications*, vol. 180, no. 10, pp. 1811–1820, 2009.

[20] X. Y. Hu and N. A. Adams, "A multi-phase SPH method for macroscopic and mesoscopic flows," *Journal of Computational Physics*, vol. 213, no. 2, pp. 844–861, 2006.

[21] R. A. Dalrymple and O. Knio, "SPH modelling of water waves," in *Proceedings of the 4th Conference on Coastal Dynamics*, pp. 779–787, ASCE, June 2001.

[22] H. Schechter and R. Bridson, "Ghost SPH for animating water," *ACM Transactions on Graphics*, vol. 31, no. 4, article 61, 2012.

[23] N. Akinci, M. Ihmsen, G. Akinci, B. Solenthaler, and M. Teschner, "Versatile rigid-fluid coupling for incompressible SPH," *ACM Transactions on Graphics*, vol. 31, no. 4, article 62, 2012.

[24] S. Clavet, P. Beaudoin, and P. Poulin, "Particle based viscoelastic fluid simulation," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 219–228, ACM, Los Angeles, Calif, USA, July 2005.

[25] G. Oger, M. Doring, B. Alessandrini, and P. Ferrant, "Two-dimensional SPH simulations of wedge water entries," *Journal of Computational Physics*, vol. 213, no. 2, pp. 803–822, 2006.

[26] R. Keiser, B. Adams, P. Dutre, L. Guibas, and M. Pauly, "Multiresolution particle-based fluids," *ETH Department of Computer Science*, vol. 31, no. 6, pp. 1797–1809, 2006.

[27] S. Oh, Y. Kim, and B.-S. Roh, "Impulse-based rigid body interaction in SPH," *Computer Animation and Virtual Worlds*, vol. 20, no. 2-3, pp. 215–224, 2009.

[28] M. Becker, H. Tessendorf, and M. Teschner, "Direct forcing for Lagrangian rigid-fluid coupling," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 3, pp. 493–503, 2009.

[29] R. Li and X. Wang, "Individual time-stepping for rigid-fluid coupling of particle based fluids," in *Proceedings of the International Conference on Cyberworlds (CW '16)*, pp. 235–238, Chongqing, China, September 2016.

[30] G. Turk, "Generating textures on arbitrary surfaces using reaction-diffusion," *ACM SIGGRAPH Computer Graphics*, vol. 25, no. 4, pp. 289–298, 1991.

[31] G. Turk, "Re-tiling polygonal surfaces," *ACM Transactions on Graphics*, vol. 26, no. 2, pp. 55–64, 1992.

[32] A. P. Witkin and P. S. Heckbert, "Using particles to sample and control implicit surfaces," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, pp. 269–277, ACM, Orlando, Fla, USA, July 1994.

[33] D. Nehab and P. Shilane, "Stratified point sampling of 3D models," in *Proceedings of the 1st Eurographics conference on Point-Based Graphics*, pp. 49–56, Eurographics Association, Zurich, Switzerland, June 2004.

[34] R. L. Cook, "Stochastic sampling in computer graphics," *ACM Transactions on Graphics (TOG)*, vol. 5, no. 1, pp. 51–72, 1986.

[35] M. Corsini, P. Cignoni, and R. Scopigno, "Efficient and flexible sampling with blue noise properties of triangular meshes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 6, pp. 914–924, 2012.

[36] D. Dunbar and G. Humphreys, "A spatial data structure for fast poisson-disk sample generation," in *Proceedings of the ACM (SIGGRAPH '06)*, pp. 503–508, ACM, Boston, Mass, USA, August 2006.

[37] R. Bridson, "Fast poisson disk sampling in arbitrary dimensions," in *Proceedings of the ACM SIGGRAPH Sketches (SIGGRAPH '07)*, San Diego, Calif, USA, August 2007.

[38] J. Yu and G. Turk, "Reconstructing surfaces of particle-based fluids using anisotropic kernels," *ACM Transactions on Graphics*, vol. 32, no. 1, article no. 5, 2013.

[39] X. Wang, X. Ban, X. Liu, Y. Zhang, and L. Wang, "Efficient extracting surfaces approach employing anisotropic kernels for SPH fluids," *Journal of Visualization*, vol. 19, no. 2, pp. 301–317, 2016.

Hindawi Publishing Corporation
Scientific Programming
Volume 2017, Article ID 5072427, 13 pages
http://dx.doi.org/10.1155/2017/5072427

Research Article

Intelligent Learning for Knowledge Graph towards Geological Data

Yueqin Zhu,[1,2] Wenwen Zhou,[2,3,4] Yang Xu,[2,3,4] Ji Liu,[2,3,4] and Yongjie Tan[1,2]

[1] *Development and Research Center, China Geological Survey, Beijing 100037, China*
[2] *Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Beijing 100037, China*
[3] *School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing 100083, China*
[4] *Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China*

Correspondence should be addressed to Yueqin Zhu; yueqin_zhu@126.com and Yang Xu; b20160304@xs.ustb.edu.cn

Received 14 September 2016; Revised 21 December 2016; Accepted 12 January 2017; Published 16 February 2017

Academic Editor: HuaPing Liu

Knowledge graph (KG) as a popular semantic network has been widely used. It provides an effective way to describe semantic entities and their relationships by extending ontology in the entity level. This article focuses on the application of KG in the traditional geological field and proposes a novel method to construct KG. On the basis of natural language processing (NLP) and data mining (DM) algorithms, we analyze those key technologies for designing a KG towards geological data, including geological knowledge extraction and semantic association. Through this typical geological ontology extracting on a large number of geological documents and open linked data, the semantic interconnection is achieved, KG framework for geological data is designed, application system of KG towards geological data is constructed, and dynamic updating of the geological information is completed accordingly. Specifically, unsupervised intelligent learning method using linked open data is incorporated into the geological document preprocessing, which generates a geological domain vocabulary ultimately. Furthermore, some application cases in the KG system are provided to show the effectiveness and efficiency of our proposed intelligent learning approach for KG.

1. Introduction

Geological data is a variety of data and information accumulated in the geological research work and practical activities. Generally, the types of geological data are in a wide variety, including geological documents, geological books, geological information and journals, physical specimens, and electronic file data [1–3]. Due to the technical reasons, traditional storage modes may lead to the operations inefficient in queries, statistics, and updates; then they are not conducive to the application of checking, querying, and mining, which means the low ability of data services.

With the increasing economic and society, in the field of geological survey, geological data sharing service has become an important tool to measure the level of social and business management, which is significant in ensuring the sustainable development of geological work. The features of geological data include increasing volume, complex type, and long response time. Aiming at the geological application problems, the intelligent analysis and deep mining of geological data could reduce the repetitive working and the risk of geological survey [4, 5].

In recent years, knowledge service based on the knowledge graph (KG) technology and the search technology of semantic web has become a research hot spot in information service. In this case, the KG arises at the historic moment [6–10]. Drawing KG and conducting intelligent search based on KG have formed a mature methodology. For example, in Chinese, Sogou Knowledge Cube is the first KG introduced into the domestic search engine [11], which makes a reoptimization of search results through the integration of massive Internet fragmented information, and presents the core of the information to users. Baidu Zhixin is a new generation of Baidu search engine technology based on its KG [12]. There are four steps in the process of constructing the KG, including the named entity mining, attribute-value pair (AVP) mining, upper and lower relation mining, and related entity mining. Although there are some successful applications, it

still has room for developing KG, and the applications still should be further strengthen, especially for the geological data.

In this article, the KG construction technology is applied in geology to implement intelligent analysis and deep mining of geological data. Through an unsupervised knowledge learning method for open data sources, we not only achieve self-learning process for a set of documents, but also form a geology glossary and complete the construction of KG. Through the research along this topic, promoting the geological materials information and social services has important value for the realization of intelligent geological survey.

The contributions of this article are as follows:

(1) *On the basis of linked open data and ontology learning strategy, we achieve the unsupervised learning and geological knowledge extraction for geological documents*. Through the processing steps of word segmentation, web crawler, keywords extraction, and relation extraction, the processing of geological documents and the deep mining of geological information are implemented.

(2) *Through the use of geological data sample, the geological ontology library, including entities, geology dictionary, and semantic link are obtained*. Firstly, we analyze the features of the geological data and acquire the geological ontology based on the geological knowledge extraction. Meanwhile, considering the specificity of the geological data, we design the corresponding entities combining with the geological dictionary and other specific documents. Secondly, through the processing of documents and web crawler using online encyclopedia, an expanded dictionary of geology and complete interconnection of semantic relations are proposed.

(3) *The KG towards geological data is proposed and the application system under Browser/Server (B/S) schema is also achieved*. Through the optimization of semantic relations and the storage of our knowledge base, we develop a framework for the application of the KG towards geological data, on the basis of the self-processing and self-expanding technologies towards geological documents. Combined with HTML5, JSP, Servlet, JDBC, and other advanced technologies, a B/S-based application system of KG is designed to realize the documents importing and processing, the intermediate results presenting, and expert intervening.

The rest of this article is organized as follows. Section 2 provides an analysis for background regarding the features of geological data and KG. The details of the intelligent learning scheme for KG, including framework, key technologies, and algorithms, are proposed in Section 3. Furthermore, the evaluation for our developed KG is conducted on two application experiments in Section 4. Finally, a conclusion is provided in Section 5.

2. Backgrounds

2.1. The Features of Geological Data. In recent years, with the increasing demand of geological data in production units and social masses, the geological data services are facing the dual demands of "digitization" and "socialization." It is necessary to improve the contents and methods of geological data services, promoting the government departments to adapt to the development of the situation and achieve the transformation from the archival results to the service product [13]. Due to the features of reusability, reprocessing, and long-term service, the data information in mining industry, which has been accumulated over the years, could be called "big data." Generally, geological data is mainly composed of structured and unstructured diversity data generated from geological actions. Its features are summarized as "5V," that is, volume, variety, value, velocity, and veracity.

The requirement of maintaining the type and quantity of geological data grows with the long-term accumulation for data. It includes various types of electronic file data, such as documents, maps, database (map database, spatial database, and attribute database), pictures, charts, videos, audio, which might be structured, semistructured, and unstructured. Due to the technical reasons, this storage mode makes data query, statistics, updates, and other operations to the data not only inefficient, but also detrimental to the application, such as check, query, and mining, which leads to the low capability to the data service. Hence, it is significant to exploring how to apply the concept and technology of big data to organize massive geological data in the field of geology effectively and achieve the corresponding services [14].

Generally, the diversified fragmentation of complex geological unstructured data is one of the most striking features. There are mainly three contents that reflect to the data analysis and mining processing, including the establishment of content index library, search, and clustering recommendation [15]. Although it has achieved some results on this aspect [16, 17], with the development of the intelligent geological survey, multicategory whose content extension organization and search application are based on geological domain ontology will be an important direction of geological data repository construction in the future [18]. In view of this aspect, we can try to use the semantic link technology based on KG to eliminate the ambiguity of search, which could make the search engine use the search based on entity instead of character string. In addition, the Internet could also be used to provide rich resources for the KG, in order to realize the semantic link of big data, intelligent analysis, and mining of the geological large data accurately and effectively.

2.2. Knowledge Graph (KG). KG is also known as science knowledge graph, knowledge domain visualization, and knowledge domains map. It is a series of various graphs that show the development process of scientific knowledge and the structure relation [19]. It could describe the knowledge resource and its supporter, excavation, analysis, and construction and draw and display the knowledge and the mutual connection between them by using the visualization technology [6, 20, 21]. KG is a research method

combining the theory and method of applied mathematics, graphics, information visualization technology, information science, and other disciplines with metrology citation analysis, cooccurrence analysis, and other methods to show the core framework, the history of development, the frontier domain, and overall knowledge structure of the discipline through visual graph. It shows the dynamic development of knowledge and the complex domain knowledge through data mining, information processing, knowledge measurement, and graphics rendering.

Most works regarding the KG originated from Google KG. It is essentially a semantic network. The nodes represent entities or concepts and the edges represent a variety of semantic relations between entities and concepts. Moreover, the motivation of KG is from a series of practical applications, including semantic search, machine answering, information retrieval, electronic reading, and online learning. Now, some companies, such as Baidu, Sogou, have launched their own KG.

Our researchers have developed many applications around KG, while illustrating different perspectives in their process. For example, in the process of visual analysis of Chinese science literature, it showed the time sequence distribution, journal distribution, and author distribution of scientific literature during the past 30 years [22]. According to cocitation analysis for the authors of the quotation in 24 kinds of information science core journals, the KG of information science was drawn [23]. Based on the example of education in China, KG was drawn to evaluate excellent scientific research institutions through word frequency analysis, high frequency author statistics, high yield author cooperative network, and other methods [24].

In addition to the above applications, many scholars have also carried out some works in KG. Hook showed that KG has four purposes (i.e., discovery, understanding, communication, and education) and six aspects of application (i.e., microcosmic display of specific areas, macroscopic visualization of subject, assisting in the education course teaching, saving document knowledge in coordination, facilitating the use of digital library, and displaying knowledge dissemination) [25]. It indicated that KG could be used in displaying the overall structure of the domain knowledge, analyzing retrieval result visually, grasping the overall knowledge of discipline and evolution situation of visualization knowledge, and grasping the rapid variation of the knowledge [26]. Meanwhile, information fusion as a key issue plays an important role in developing a KG. To deal with this issue, a novel approach was proposed for exemplar extraction through structured sparse, considering not only the reconstruction capability and the sparsity, but also the diversity and robustness [27]. Furthermore, based on previous work, a joint kernel sparse coding model was developed to solve the multifinger tactile sequence classification problem, where all of the coding vectors were encouraged to share the same sparsity support pattern [28].

KG applications increase rapidly in recent years, which cover some disciplines of natural science and social science, and show the osmotic tendency towards other disciplines. Drawing KG and mining KG have formed a high mature methodology. However, the function of KG has not been fully applied, and the application still needs to be further strengthened. So far, only little attention has been paid to the geological data field. Hence, it is necessary and important to consider these particular objects.

3. Intelligent Learning for Knowledge Graph

3.1. The System Framework. The construction of KG towards geological data consists of two logical components: knowledge extraction and knowledge management. The former mainly learns the corresponding geological knowledge through unsupervised processing and including five steps, which are word segmentation, frequency statistics, web crawler, keywords extraction, and relation extraction. The latter is basically composed of two parts: knowledge graph storage and retrieval. The specific processes are shown in Figure 1.

3.2. Knowledge Extraction. Knowledge extraction is a key step in the construction of knowledge graph, as well as in the processing of geological documents. Knowledge extraction in this article, through an unsupervised knowledge learning method based on an open source, and the geological domain vocabulary and knowledge graph would be formed through the automatic learning of a large number of geological documents. The flow of knowledge extraction is shown in Figure 2.

Knowledge extraction has three major steps, including data sources analysis, entity/concept extraction, and relation extraction.

3.2.1. The Analysis of the Available Data Sources

(1) Text. Texts are the most abundant data source. It is difficult to learn knowledge from texts due to their nonstructural property. In this article, we obtain a large number of geological professional texts from library.

(2) Internet Encyclopedia. Internet encyclopedias (e.g., Wikipedia, Baidu baike, and Baike.com) are the large-scale free encyclopedias that allow users to edit almost any article accessible. Through technical tools such as web crawler, we obtain knowledge from Internet encyclopedias continuously, which could be updated and expanded automatically.

Although the contents of encyclopedias exist with the form of web pages, there are still a lot of structured information. Since all of encyclopedias have their own classification system, category labels are used to organize a large number of entries. In general, each entry has category label, which could be used to label its own type. In addition, most of entries have multiple labels. For example, the category labels of "Steve Jobs" could be "20th-century American business people," "American billionaires," "American computer business people," and many others in Wikipedia.

This article mainly focuses on Chinese information in Internet encyclopedias. Wikipedia is considered the Internet's largest and most popular general reference book. However, Chinese content in Wikipedia is not perfect. On the one

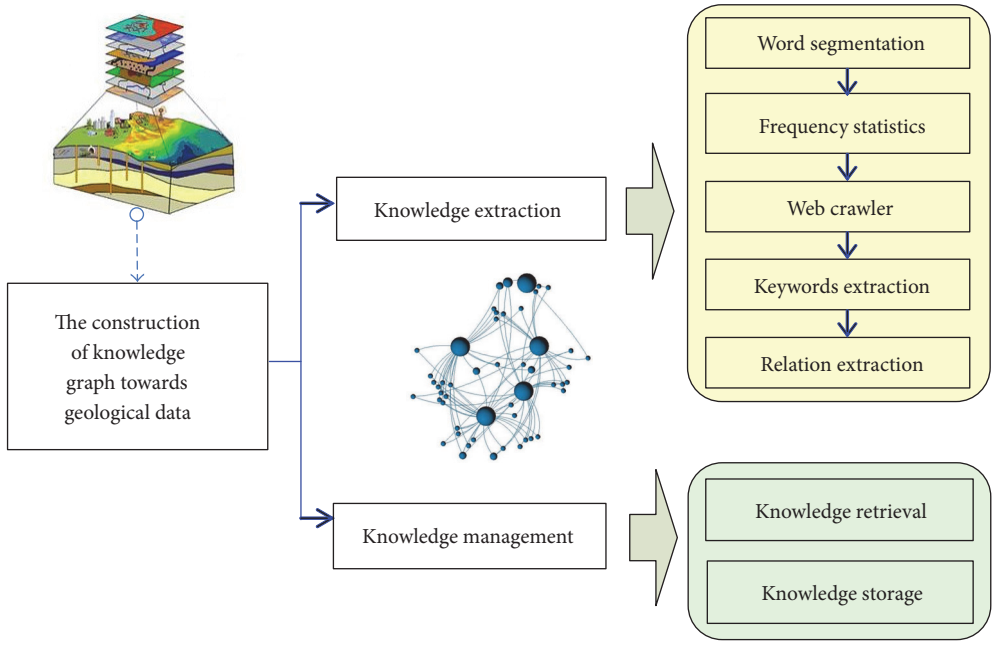


Figure 1: The logical structure of knowledge graph construction towards geological data.

hand, the total number of entries is insufficient. And, the contents of the articles in Wikipedia are also relatively short, and some parts of them are translated from other languages directly, which are lacking the expression exactly in Chinese. Consequently, we make use of Baike.com instead of Wikipedia as the data source of web crawler in this article.

3.2.2. Entity/Concept Extraction. Entity/concept extraction mainly starts from these two data sources. We could filter out entities or concepts of geology directly by combining the information after text processing with category labels of Baike.com. Therefore, the entity/concept extraction includes four bottom-up steps: word segmentation, frequency statistics, web crawler, and keywords extraction.

The technology of HanLP could be used in the word segmentation, stop word filtration, and frequency statistics. Motivated by the TextRank algorithm, word segmentation used in this article is as follows. First of all, we use HanLP standard tokenizer to process documents, which are divided into different parts of speech words. Secondly, the custom data dictionary and extended stop list are designed. Finally, we filter out the word with little relevance to retrieving content and only retain the designated part of speech through the method of TextRank algorithm. Meanwhile, we also filter out the stop words, so as to achieve the effect of keyword extraction.

In terms of web crawler, we mainly consider to crawl the category labels of entries in the Internet encyclopedia by an automated tool Selenium, which could open the HtmlUnit browser, search entries, and access to class label information via programming by custom. Specifically, the method for online encyclopedia crawler is as follows. When we want to get information about a word "*n*," we should open our browser first. Then, we search and open encyclopedia interface of "*n*." We can locate and save category label elements by XPath finally.

In terms of the keywords extraction, according to the geological dictionary and the category labels, we could exactly determine whether the words in the segmentation results belong to the geological keywords or not. Through the statistical characteristics of Wikipedia category labels, we extract some keywords, including geography, mining, marine, rock, hydrology, environment, natural disasters, biology, city, air, oil, roads, plants, energy, metallurgy, and civil. We put all crawled category labels into a map collection. By calling the containsKey method of map, we can determine whether the collected object contains the keywords, if the answer is yes, this object is defined as a geological entity.

3.2.3. Relation Extraction. The purpose of relation extraction is nontaxonomic relation extraction of the association rule analysis in data mining and the Internet encyclopedia. The correlation between two geological terminologies is acquired by association rule analysis. And the category relationship of terminologies is acquired through crawling Internet encyclopedia.

The basic principle of association rule is that if the two concepts or entities frequently appeared in the same unit (e.g., a document, a paragraph, or a sentence), we could make sure there exist some relationships between them. We do not care about the specific semantic relations between two concepts, but the correlated degree between them. Hence, judging the correlated degree between two concepts through cooccurrence analysis in a document is more important. With the increase of the number of documents processed, there would be a higher correlated degree if the two concepts frequently appeared together. This method is also motivated by the process of human reading and learning. However, this method is just suitable to be

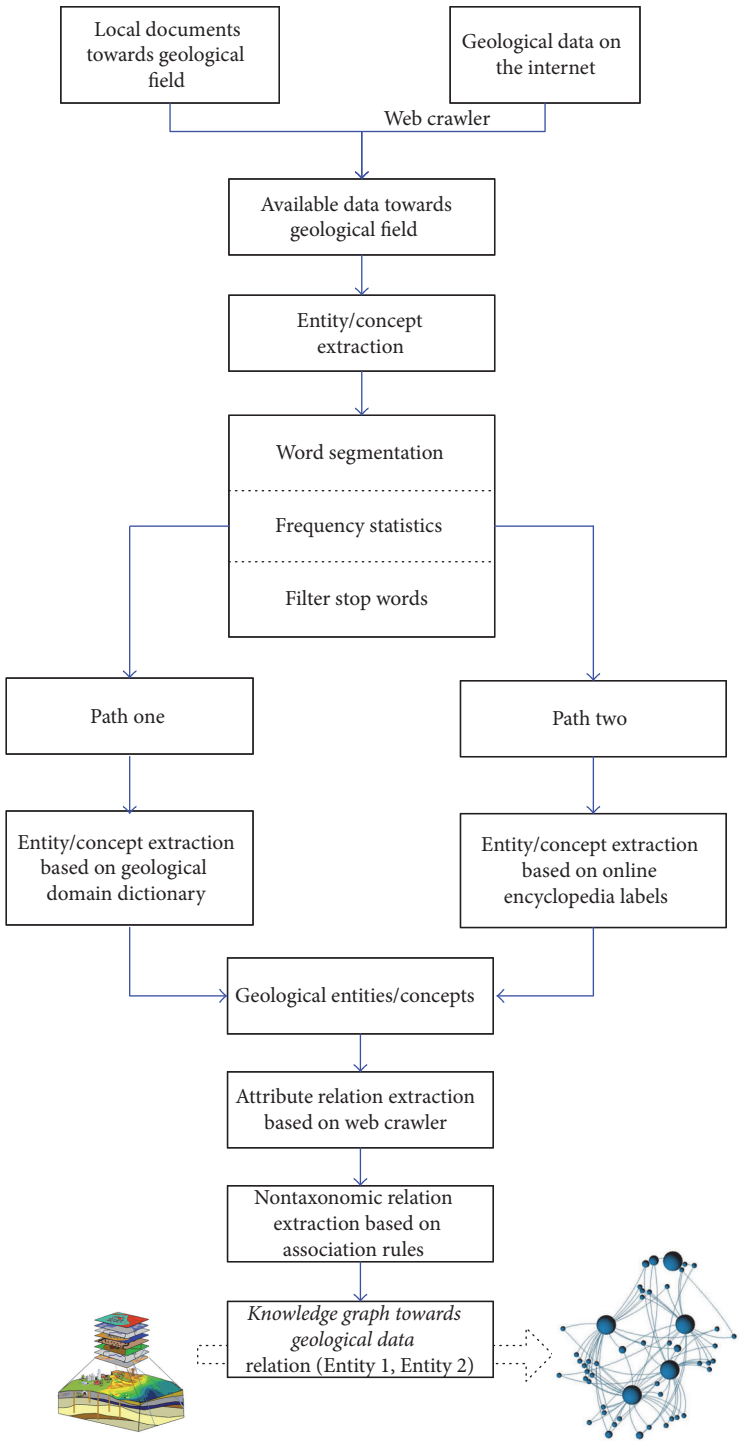


Figure 2: The flow of knowledge extraction.

employed for dealing with large number of documents; when the number of documents is small, this method would be inefficient.

Meanwhile, the purpose of crawling Internet encyclopedia is to obtain relationships between concepts and entities by making use of the open data source in the online encyclopedia. As mentioned above, here we mainly consider the category relationships.

Using the above two methods, the rule of our relation extraction is as follows. In terms of correlated degree, we set a relational degree R for each concept, where the initial value of R is 0. After processing a document, the correlation between all the words which appeared in the document is increased by 1. The value of R updates once in the process of dealing with document each time. Furthermore, each concept has category labels as their property.

Table 1: The attributes of tables in background database.

Table name	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
"articles"	ID	Content	Date	Path	—
"words"	ID	Content	Frequency	Label	—
"re_words_articles"	ID	ID1	ID2	Frequency	—
"re_words_words"	ID	ID1	ID2	ID3	Count
"dictionary"	Name	Label	—	—	—

3.3. Knowledge Management. Knowledge management considers how to show the knowledge acquired through above steps in a visualized way. The main technical methods are the database storage and retrieval.

3.3.1. Database Storage. Considering the actual needs of the geological field, the system uses MySQL database as the background database. MySQL database is one of the best relational database management systems in web application, which has a small size, fast storage and retrieval speed, and low cost.

In our system, the entities and relationships acquired by processing geological documents are stored in a special database. Through JDBC technology, the background database operations, such as CRUD, are allowed. There are five tables in our database. Table "articles" stores the information about the documents processed, including ID, name, added time, and local storage path of documents. Table "words" stores the information about the words filtered out from the results of segmentation, including ID, content, frequency, and category labels of words. Table "re_words_words" stores the correlation information between two geological terms.

The attributes of those tables in our background database are in Table 1.

3.3.2. Knowledge Graph Retrieval. Retrieval can be carried out by users only after storing the knowledge extracted from documents in our database. Based on B/S working schema, the browser makes a post request to the back-end server after users input the search words. Meanwhile, back-end server responds to the request, getting the words submitted and the number of nodes that need to be rendered (it is set to 20 as a default value). The retrieved words are set to the key nodes and retrieved in our database. Then, it returns the results to the browser. The returned contents include ID, content and category labels of node, and ID of correlated documents.

3.3.3. Backstage Management System of KG. Backstage management system of KG is designed to facilitate the process of documents and the operation of database for users, mainly including login page, geological documents processing page, and expert intervention page.

Two login modes could be chosen when users enter the login page by inputting the URL in the browser. Users could enter the geological document processing page if logged in as an administrator. Users also could enter the page of expert intervention if logged in as an expert. The browser submits the form, including name, password, and login mode. Subsequently, the users authorization would be checked by server, and users could enter the relevant page after verified.

On the page of geological documents processing, users could input the document name and storage path. And background module gets the form data submitted by users when the button "submit" is clicked on. The background module enters the stage of document processing and the results are stored in background database if all of this input data is valid. On the page of expert intervention, the experts have the right to add and delete the correlation between two words. For example, when adding a correlation, the experts enter the two words in the input box and click the button "submit." The browser submits these two words to the background module, and the background module judges whether there is a correlation between them or not. If the association does not exist, the background module would add a correlation, which is defined as "expert-defined."

3.4. The Key Algorithms. The prototype system of KG towards geological big data is designed and accordingly implemented using B/S architecture and HTTP protocol, which includes natural language processing (NLP), data mining, web application development, and other related technologies. Key technologies and solutions involved during the process of system development are described as follows.

3.4.1. The Automatic Chinese Segmentation Technology: HanLP. HanLP is a Java toolkit composed of a series of models and algorithms, whose target is to promote the application of NLP in the production environment. HanLP supports Chinese word segmentation. Its functions include N shortest path word segmentation, CRF word segmentation, index word segmentation, and user defined dictionary. Specifically, they are named entity recognition, keyword extraction, phrase extraction, Pinyin conversion, conversion between simplified and complex, and dependency parsing (i.e., MaxEnt dependency parsing, CRF dependency parsing). The characteristics of HanLP are perfect function, efficient performance, clear architecture, new corpus, and being customizable.

(1) TextRank Algorithm. Making the use of TextRank for Chinese word segmentation mainly includes word segmentation, delete stop words, and iterative voting. The basic idea of TextRank Chinese word segmentation is as follows: dividing the original text into sentences first, filtering out the withdrawal in each sentence, and only retaining the specified part of speech word. From it, we could get a set of sentences and a set of words. Then each word would be as a node

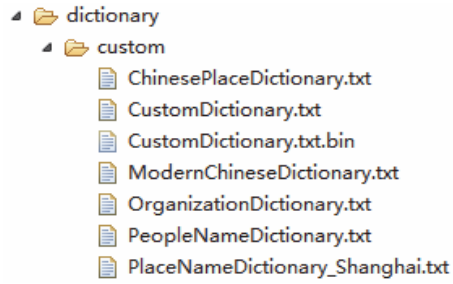


Figure 3: User defined dictionary.

in the TextRank by using the method of matrix iterative convergence [29]. The window size is set to K, and we assume that a sentence is constructed by the following words:

$$w_1, w_2, w_3, w_4, w_5, \ldots, w_n, \tag{1}$$

where $w_1, w_2, \ldots, w_k, w_2, w_3, \ldots, w_{k+1}, w_3, w_4, \ldots, w_{k+2}$ are all in a window. There is an undirected and unweighted edge between any two words corresponding to the node in a window.

With the above composition diagram, we could calculate the weight of each word node. Then, the iterative formula in TextRank algorithm is as follows:

$$\mathrm{WS}\left(V_i\right) = (1-d) + d \times \sum_{V_j \in \mathrm{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \mathrm{Out}(V_j)} w_{jk}} \mathrm{WS}\left(V_j\right), \tag{2}$$

where $d \in [0, 1]$ is a damping factor, V_i is a given node, $\mathrm{In}(V_i)$ are the set of nodes that point to it (predecessors), $\mathrm{Out}(V_i)$ is the set of nodes that node V_i points to (successors), and $\mathrm{WS}(V_i)$ is the weight of node V_i.

(2) User Defined Dictionary. HanLP word segmentation supports the function of custom dictionary, where our custom dictionary designed is shown in Figure 3.

We add a large number of words that could help the word segmentation of geological documents in the custom dictionary effectively. Here, the "CustomDictionary" includes 21,742 geological words, the "OrganizationDictionary" includes 31,926 institutional nouns, the "ChinesePlaceDictionary" includes 90,558 place names, the "PeopleNameDictionary" includes 50,192 personal names, and the "ModernChineseDictionary" includes 207,964 modern Chinese additional words. Among them, "CustomDictionary" is a dictionary defined by a global user which could add, delete, and affect all word segmentation at any time.

3.4.2. Internet Encyclopedia Crawler Based on Selenium. On the basis of our analysis mentioned above, it is effective and efficient to integrate the online encyclopedia crawler technology into the processing flow of geological documents, which need to get the category labels of words obtained by word segmentation in Wikipedia. The mainstream method of crawler is implemented using URL address webs which can be obtained through depth or breadth first search strategy. Here, the web site that we need to crawl is fixed (i.e., http://www.baike.com/), and we already have the target word (i.e., word segmentation results). Then, it is available to introduce the automated crawler technology. Here, we use Selenium automated crawler.

Selenium automation test browser is mainly applied to the automated testing of the web application, while supporting all management task automation based on web. By embedding the Selenium IDE plug-in into the browser, the recording and playback functions of a simple browser operation could be achieved.

It should be noted that Selenium provides a highly rapid and convenient way for the fixed web crawler. Here, we use Selenium to control HtmlUnit, a virtual browser that Java comes with, which serves the purpose of automated crawler. The specific process mainly includes opening the HtmlUnit browser, reading a search word "*n*," retrieving by opening the encyclopedia interface of the retrieved word "*n*," getting category labels according to category label elements by XPath, and finally closing the browser.

Implementation details of Internet encyclopedia crawler are as follows:

(1) Open HtmlUnit browser: *static final WebDriver driver = new HtmlUnitDriver()*

(2) Open the interface of search word "*n*": *driver.get ("http://www.baike.com/wiki/"+n)*

(3) Locate the label element:

List<WebElement> elements = driver.findElements(By. xpath("//dL[@id='show_tag']/dd/a"))

3.4.3. Java Web Development Based on Servlet and Java Server Pages (JSP). Java Servlet is a Java program that extends the capabilities of a server. Although Servlets could respond to any types of requests, they implement applications hosted on web servers usually. Such Web Servlets are the Java counterpart to other dynamic web content technologies, such as PHP and ASP.NET.

Servlets are often used to process and store a Java class in Java EE that conforms to the Java Servlet API, a standard for implementing Java classes which could respond to the requests. And, Servlets could communicate over any clientCserver protocol, but they are often used with the HTTP protocol. So, "Servlet" is often used as shorthand for "HTTP Servlet." Thus, a software developer should use a Servlet to add dynamic content to a web server by using the Java platform. The generated content is HTML but may be other data such as XML. Servlets could maintain state in session variables across many server transactions by using HTTP cookies or rewriting URLs.

Servlets could be generated from JSP by the JavaServer pages compiler automatically. Architecturally, JSP could be viewed as a high-level abstraction of Java Servlets. It allows Java code and certain predefined actions to be interleaved with static web markup content, such as HTML, with the resulting page being compiled and executed on the server to deliver a document. JSP are translated into Servlets at

Table 2: The Servlets and their key functions.

Servlet name	Key functions
"Myservlet.java"	It is used in the retrieval of KG, and it gets the form data submitted by the user and retrieves them.
"Myservlet2.java"	It is used in the second retrieval. When clicking on some word in the page, the user can get the graph of this word.
"LoginServlet.java"	It is used in the login function of the backstage management system of KG, and it gets the form data submitted by the user and enters the response page.
"AddServlet.java"	It is used while adding a relationship in expert intervention page.
"DelServlet.java"	It is used while deleting a relationship in expert intervention page.
"CoreServlet.java"	It is used while showing the intermediate processing for geological documents.

runtime, and each JSP Servlet is cached and reused until the original JSP is modified.

Servlets can complete the following tasks:

(1) The web container initializes the Servlet instance; then the Servlet instance could read data that has been provided in the HTTP request.

(2) The Servlet instance could create and return a dynamic response page to the client.

(3) The Servlet instance could access server resources, such as files and database.

(4) The Servlet instance could prepare dynamic data for the JSP and create a response page with JSP.

In this article, the Servlets and their key functions that we design under com.servlet package are shown in Table 2.

In summary, the software platforms and development environments in our system are as follows. Operating system is Windows 7. Programming language is Java. Programming environment is MyEclipse 10. Web development environment is Tomcat + Severlet + JSP. Web crawler environment is Selenium + HtmlUnit.

4. Experiments and Evaluation

4.1. Processing for a Single Document. Processing for a single geological document is as shown in Figure 4. We can see that user who logs as the administrator enters the document processing page. Then, the user inputs the name and storage path of document and clicks on the submit button. A background module gets the form data submitted by the administrator and determines if this document exists in the local path. The background module enters the stage of document processing when all of this input data are valid. The document is converted into a long string. Then the background module would cut word segmentation by HanLP, filter out stop word, and select out the geological terminologies. The result after intermediate processing is showed in the document processing page.

The document is processed using the similar method in [30].

4.1.1. The Result Analysis of Segmentation

(1) Some results of segmentation in our KG system are shown in Figure 5. After translating it from Chinese into English, the updated version of Figure 5 is shown in Figure 6.

(2) Some results of segmentation by NLPIR systems of Beijing Institute of Technology [31], which is a popular NLP system, are shown in Figure 7. After translating it from Chinese into English, the updated version of Figure 7 is shown in Figure 8.

According to the process in [30], some results of segmentation in our KG system are showed in Figure 5. Meanwhile, Figure 6 shows some results of segmentation in NLPIR systems. By comparing these two figures, we can find that the results processed in our system are more valuable and satisfactory. For example, to these geological terminologies, such as "华北陆块 (North China craton)," "高于庄组 (Gaoyuzhuang Formation)," "下马岭组 (Xiamaling Formation)," "铁岭组 (Tieling Formation)," and "吕梁运动 (Luliang Movement)," our KG system can accurately cut word segmentation. However, in NLPIR systems, many geological terminologies are cut inaccurately.

4.1.2. Word Frequency Statistics. The results of word frequency statistics are as shown in Figure 9. After translating it from Chinese into English, the updated version of Figure 9 is shown in Figure 10. We can see that our system can count word frequency correctly for the result set of segmentation, such as "杨庄组 (Yangzhuang Formation)/13," "下马岭组 (Xiamaling Formation)/8," "长石石英砂岩 (Feldspathic Quartz Sandstone)/1," and "同位素年龄 (Isotope Age)/1."

4.1.3. Keywords Extraction. Figure 11 shows the results of keywords extraction. We normally consider that the terminologies contained in the title and subtitles are basically the keywords for the document. Therefore, the result of keywords extraction in Figure 11 includes the critical position "华北陆块 (North China Craton)," three key stratigraphic units "高于庄组 (Gaoyuzhuang Formation)," "杨庄组 (Yangzhuang Formation)," and "下马岭组 (Xiamaling Formation)," the key stratigraphic unit "元古界 (Proterozoic Erathem)," and the major stratigraphic relationship "不整合面 (Unconformities)." In summary, our keywords extraction has satisfying results.

4.1.4. Internet Encyclopedia Crawler. The results of category labels crawled from the Internet encyclopedia (http://www.baike.com/) are as shown in Figure 12, including geological terminologies of segmentation result sets and the category labels. After translating it from Chinese into English, the updated version of Figure 12 is shown in Figure 13.

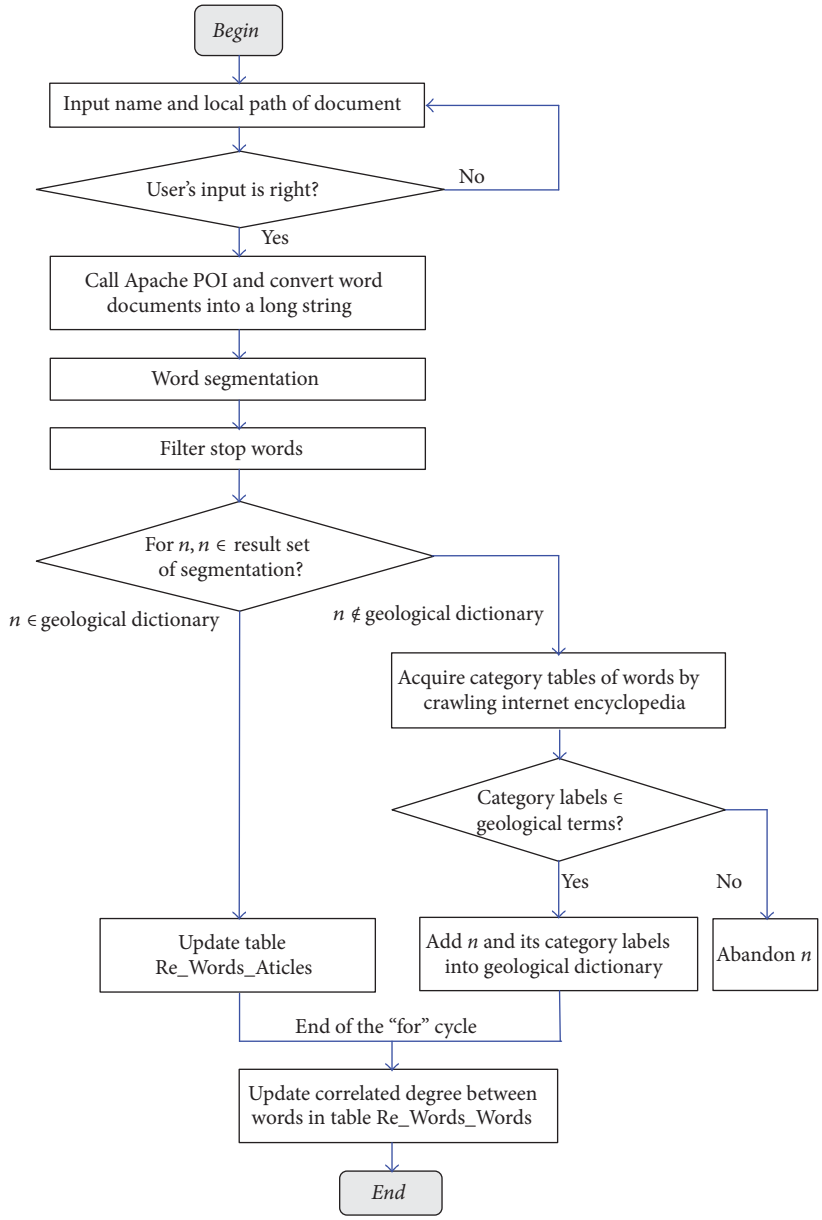


Figure 4: Processing for geological documents.

华北地块 /北缘 /元古界 /重要 /合面 /地质 /特征 /构造 /意义 /华北地块 /北缘 /元古代 /沉积层序 /出现 /沉积间断 /不整合 /研究 /不整合 /进行 /野外 /考察 /分析 /特征 /空间 /分布 /高于庄组 /底部 /杨庄组 /底部 /不整合 /海平面 /变化 /造成 /大陆边缘 /内部 /高地 /沉积间断 /海岸 /超覆 /沉积 /结果 /铁岭组 /顶部 /合面 /可能 /环境 /发生 /挤压 /抬升 /有关 /下马岭组 /沉积 /抬升 /可能 /相邻 /大陆 /地块 /碰撞 /作用 /产物 /经历 /吕梁运动 /华北地块 /成为 /全球 /Columbia /大陆 /自古 /元古代 /后期 /华北地块 /边缘 /伸展构造 /环境 /形成 /裂谷 /盆地 /华北地块 /北缘 /元古界 /长城 /蓟县 /正是 /伸展构造 /背景 /形成 /发展 /长城 /蓟县 /巨厚 /碎屑岩 /碳酸盐岩 /组合 /长城 /自下而上 /常州沟组 /串岭沟组 /团山子组 /组成 /浅海 /碎屑岩 /碳酸盐岩 /组合 /分布 /范围广 /蓟县 /自下而上 /高于庄组 /杨庄组 /雾迷山组 /洪水庄组 /铁岭组 /组成 /下部 /碳酸盐岩 /上部 /页岩 /泥质 /白云岩 /砂岩 /组合 /青白口 /自下而上 /景儿峪组 /岩性 /砂岩 /碳酸盐岩 /为主 /页岩 /值得注意 /最近 /下马岭组 /内部 /火山岩 /定年 /结果 /证明 /下马岭组 /属于 /元古界 /下马岭组 /最新 /年龄 /数据 /应对 /华北地块 /北缘 /新元 /古代 /地层 /进行 /重新 /划分 /下马岭组 /置于 /建立 /西山 /下部 /表示 /华北地块 /北缘 /新元 /古代 /地层 /层序 /同位素年龄 /数据 /合面 /位置 /成因 /解释 /对应 /构造 /环境 /高于庄组 /底部 /界面 /华北地块 /北缘 /剖面 /实测剖面 /高于庄组 /伏地

Figure 5: Some results of segmentation.

Gaoyuzhuang Formation / bottom / Yangzhuang Formation / bottom / unconformity / sea level / change / cause / continental margin / inside / highland / hiatus / coast / overlap / sedimentation / results / Tieling Formation / top / ormity surface / possible / environment / occur / extrusion / uplift / relating to / Xiamaling Formation / sedimentation / uplift / maybe / adjacent / continental / block / collision / effect / product / experience / Lvliang movement / North China Block / become / global /Columbia / supercontinent / since ancient times / Proterozoic / late stage / North China Block / edge / extensional tectonics / environment / formation / rift basin / North China block / north edge / Proterozoic / the Great Wall / Jixian County / is / the extensional tectonic / background / form / development / the Great Wall / Jixian County / thick / clastic rock / carbonate / combination

Figure 6: The updated version of Figure 5 after translating it from Chinese into English.

4.2. Searching in KG. The specific process of retrieval in KG is shown in Figure 14. We can see from this figure that the first step users need to do is to input the retrieved word

Figure 7: Some results of segmentation in NLPIR system.

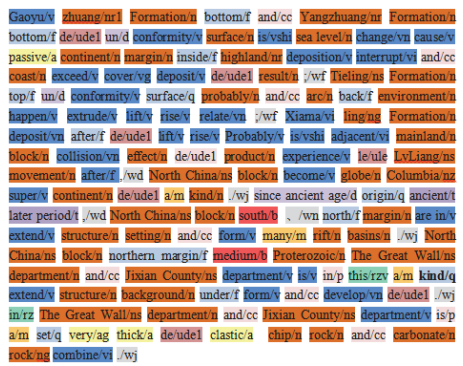


Figure 8: The updated version of Figure 7 after translating it from Chinese into English.

and click on the "search" button. Background module gets the form data submitted by the user and sets it as the key node. Furthermore, the background module retrieves in our database to filter out the terminologies that have a relational degree with the key node in top 20 (default) and shows them in a graph.

4.2.1. The Comparison of Different Retrieval Processing Stages

(1) After processing one geological document, the results of retrieving "变质岩 (metamorphic rock)" are in Figure 15.

(2) After processing 100 geological documents, the results of retrieving "变质岩 (metamorphic rock)" are in Figure 16.

Figures 15 and 16 show the results of the KG retrieval. The orange node represents the retrieved word "变质岩 (metamorphic rock)." The blue nodes represent the terminologies, which have a top-20 relational degree with the orange node, such as "同位素年龄 (isotope age)" and "砂岩 (sandstone)." When the mouse is placed on some node, we could acquire its ID and category labels.

From the comparison of two retrieval processing stages, we could see that the results of KG have been improved with a growing number of documents processed. When the number of processed documents is 1, the retrieval results have little relevance with the retrieved word. However, when the number is 100, we could get entities that have a very close relationship with "变质岩 (metamorphic rock)," such as

Figure 9: The results of word frequency statistics.

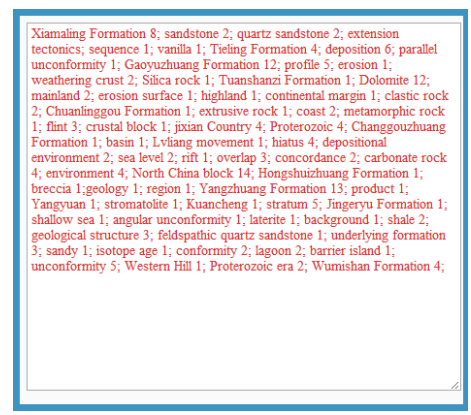


Figure 10: The updated version of Figure 9 after translating it from Chinese into English.

"花岗岩 (granite)," "岩浆 (magma)," and "火山岩 (volcanic rock)."

In addition, we could get the following information from the above results.

(1) The top 20 geological terminologies associated with "变质岩 (metamorphic rock)."

(2) The category labels for every geological terminology.

(3) The ID of documents in which both words appear.

4.2.2. Searching More Words. Furthermore, some complicated phrases and sentences can also be processed correctly. For example, when inputting "侵入岩和沉积岩 (intrusive rock and sedimentary rock)," the background module can cut word segmentation into two keywords "侵入岩 (intrusive rock)" and "沉积岩 (sedimentary rock)," retrieve them, and get the terminologies that have a relational degree with the key node in top 20. The results are as shown in Figure 17.

Similarly, we could get the following information from the above results.

(1) We can get the top 20 geological terminologies associated with "侵入岩 (intrusive rock)" and "沉积岩 (sedimentary rock)."

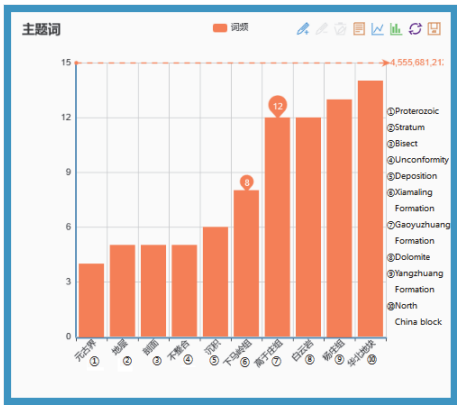


Figure 11: The results of keywords extraction.

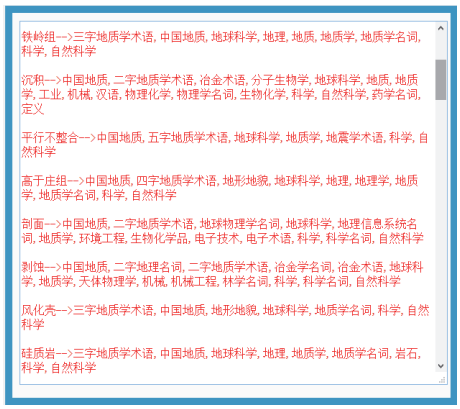


Figure 12: The results of Internet encyclopedia crawler.

(2) We can get the category labels for every geological terminology in KG.

(3) We can get the ID of documents in which both words appear.

(4) While retrieving the two words, we can get the documents in which both words appear, and it achieves mining of implicit related documents.

(5) In addition, we can see the following:

(i) In terms of "侵入岩 (intrusive rock)," there exists a connecting line between "侵入岩 (intrusive rock)" and "花岗岩 (granite)," which means that there exists a high degree of correlation between them. However, there is no connecting line between "侵入岩 (intrusive rock)" and "泥岩 (mudstone)," which means that there exists a low correlation between them.

(ii) In terms of "沉积岩 (sedimentary rock)," there exists a connecting line between "沉积岩 (sedimentary rock)" and "泥岩 (mudstone)." However, there is no connecting line between "沉积岩 (sedimentary rock)" and "花岗岩 (granite)."

Figure 13: The updated version of Figure 12 after translating it from Chinese into English.

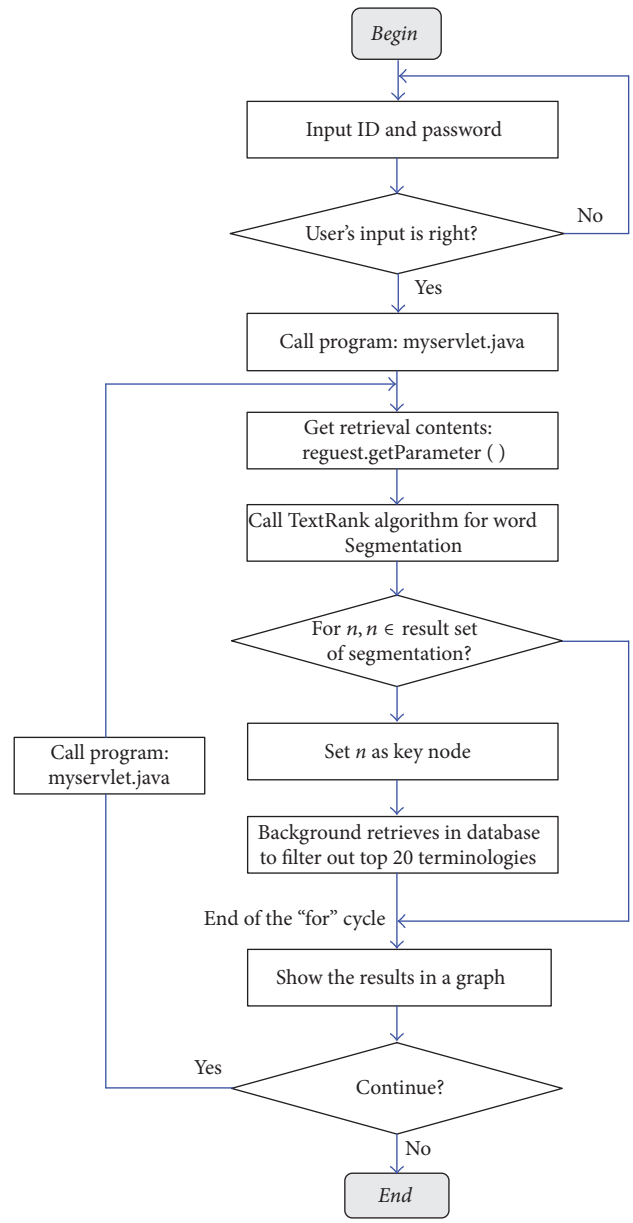


Figure 14: The specific process of retrieval.

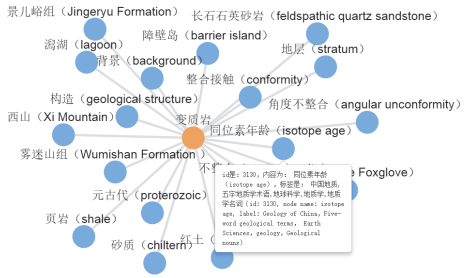


FIGURE 15: An experimental result after processing one geological document.

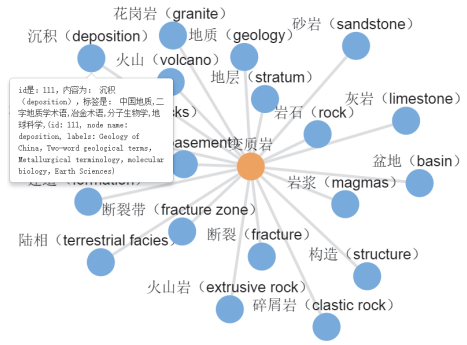


FIGURE 16: An experimental result after processing 100 geological documents.

Geological professionals know that "泥岩 (mudstone)" is a kind of "沉积岩 (sedimentary rock)" and "花岗岩 (granite)" is a kind of "侵入岩 (intrusive rock)." Hence, there is a high correlation between "侵入岩 (intrusive rock)" and "花岗岩 (granite)," as well as between "沉积岩 (sedimentary rock)" and "泥岩 (mudstone)." All the results we could see from the KG are our acquired learning information after processing 100 documents. Through this example, we could indicate that our KG system could provide valuable and accurate information in most cases. The more documents we process, the more accurate correlations we can get from the KG system.

4.2.3. The Formation of Geological Domain Dictionary. When processing geological documents, new geological terminologies and their category labels are obtained from web crawler. And they are added in our expanded geological domain dictionary.

In our experiments, the original number of words in geological domain dictionary is 11,062. And after processing 100 documents, the number of words in geological dictionary is 13,227. Some results of geological domain dictionary are in Figure 18, including geological terminologies and their corresponding category labels.

5. Conclusion

This article proposes a novel approach to constructing KG towards geological data. The proposed approach uses unsupervised learning method with linked open data to process geological documents and extract knowledge directly.

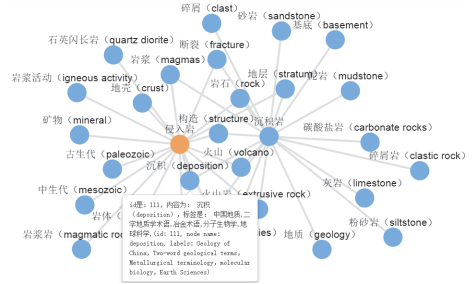


FIGURE 17: An experimental result after searching several key geological words.

Gname	Gtag
曝气(Aeration)	化学工程名词,土木工程名词,技术,水族箱装饰用品,环境工程,高新技术(chem
暴雨径流(Storm runoff)	公路科技名词,地理学,水文学(Highway science and technology terminology
边沟(Side ditch)	公路名词,公路科技名词,土木工程名词,应用科学,建筑术语,科学,铁道科技名词
冰坝(ice dam)	大陆2006至2007年剧情电影,大陆21世纪初剧情电影,大陆剧情电影,应用科学
产沙量(Sediment yield)	技术,环境工程,高新技术(technology，Environmental Engineering，high ar
常流河(Constant flow river)	技术,环境工程,高新技术,环境工程、河流(technology，Environmental Engir
常年河(Perennial river)	技术,环境工程,高新技术,环境工程、地质、水文(technology，Environmenta
潮差(Tidal range)	土木工程名词,天气百科,海洋学,环境工程,石油名词,石油工业,科学,科学名词,地
潮汐(Tide)	潮汐(Tide)
城市水文学(City hydrology)	地理,地理学,水文,生态学,自然科学(geography，geography，hydrology，ec

FIGURE 18: The geological domain dictionary.

Through this approach, we accordingly achieve an effective self-learning process for documents, form a geology glossary, and complete the construction of KG based on the technologies of documents processing and dictionary expanding. Furthermore, we design an application system of KG on the basis of B/S working schema. Finally, the test on a large number of geological documents is conducted and some satisfactory results are obtained. In the future work, aiming at the features of geological data, the knowledge extracting approach in the KG is further improved to get more accurate entities and relations.

Competing Interests

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

Acknowledgments

This work was partially supported by the Special Funds Project for Scientific Research of Public Welfare Industry from the Ministry of Land and Resources of China under Grant 201511079 and the National Key Technologies R&D Program of China under Grant 2015BAK38B01.

References

[1] T. Zhang, Y. Du, T. Huang, and X. Li, "Stochastic simulation of geological data using isometric mapping and multiple-point geostatistics with data incorporation," *Journal of Applied Geophysics*, vol. 125, pp. 14–25, 2016.

[2] M. G. Runge, M. S. Bebbington, S. J. Cronin, J. M. Lindsay, and M. R. Moufti, "Integrating geological and geophysical data to improve probabilistic hazard forecasting of Arabian Shield volcanism," *Journal of Volcanology and Geothermal Research*, vol. 311, pp. 41–59, 2016.

[3] L. Zhang, "Improvement of K-means algorithm and its applications in analysis of geological exploration seismic data," *Electronic Journal of Geotechnical Engineering*, vol. 20, no. 12, pp. 4423–4434, 2015.

[4] Y. Zhu, Y. Tan, R. Li, and X. Luo, "Cyber-physical-social-thinking modeling and computing for geological information service system," in *Proceedings of the 4th International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI '15)*, Beijing, China, October 2015.

[5] X. Luo, D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, "A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems," *Future Generation Computer Systems*, vol. 61, pp. 85–96, 2016.

[6] D. Le-Phuoc, H. Nguyen Mau Quoc, H. Ngo Quoc, T. Tran Nhat, and M. Hauswirth, "The graph of things: a step towards the live knowledge graph of connected things," *Journal of Web Semantics*, vol. 37-38, pp. 25–35, 2016.

[7] D. Danciulescu and M. Colhon, "Systems of knowledge representation based on stratified graphs. Application in natural language generation," *Carpathian Journal of Mathematics*, vol. 32, no. 1, pp. 49–62, 2016.

[8] A. Ballatore, M. Bertolotto, and D. C. Wilson, "A structural-lexical measure of semantic similarity for geo-knowledge graphs," *ISPRS International Journal of Geo-Information*, vol. 4, no. 2, pp. 471–492, 2015.

[9] N. Jayaram, A. Khan, C. Li, X. Yan, and R. Elmasri, "Querying knowledge graphs by example entity tuples," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 10, pp. 2797–2811, 2015.

[10] B. Kamsu-Foguem and D. Noyes, "Graph-based reasoning in collaborative knowledge management for industrial maintenance," *Computers in Industry*, vol. 64, no. 8, pp. 998–1013, 2013.

[11] ZhiLiFang, http://baike.sogou.com/v66616234.htm.

[12] BaiduZhiXin, http://yingxiao.baidu.com/product/site/zhixin/.

[13] M. M. Song, Z. Li, B. Zhou, and C. L. Li, "Cloud computing model for big geological data processing," *Applied Mechanics and Materials*, vol. 475-476, pp. 306–311, 2014.

[14] M. Cao and L. Lu, "Nonparametric test models of geological hybrid parents based on big data," *ICIC Express Letters*, vol. 9, no. 9, pp. 2491–2498, 2015.

[15] C. Li, J. Li, H. Zhang, A. Gong, and D. Wei, "Big data application architecture and key technologies of intelligent geological survey," *Geological Bulletin of China*, vol. 34, no. 7, pp. 1288–1299, 2015.

[16] P. Vermeesch and E. Garzanti, "Making geological sense of 'big data' in sedimentary provenance analysis," *Chemical Geology*, vol. 409, pp. 20–27, 2015.

[17] G. Yan, Q. Xue, K. Xiao, J. Chen, J. Miao, and H. Yu, "An analysis of major problems in geological survey big data," *Geological Bulletin of China*, vol. 34, no. 7, pp. 1273–1279, 2015.

[18] C. L. Wu, G. Liu, X. L. Zhang, Z. W. He, and Z. T. Zhang, "Discussion on geological science big data and its applications," *Chinese Science Bulletin*, vol. 61, no. 16, pp. 1797–1807, 2016.

[19] Q. Liu, Y. Li, H. Duan, Y. Liu, and Z. Qin, "Knowledge graph construction techniques," *Jisuanji Yanjiu yu Fazhan/Computer Research and Development*, vol. 53, no. 3, pp. 582–600, 2016.

[20] H. Liu, F. Sun, B. Fang, and X. Zhang, "Robotic room-level localization using multiple sets of sonar measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 1, pp. 2–13, 2017.

[21] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, no. 99, pp. 1–13, 2016.

[22] Z. Y. Liu and C. B. Zhang, "Review of the 30-year studies of the methodology of science and technology in China-Based on the bibliometric analysis of journal articles," *Studies in Philosophy of Science and Technology*, vol. 31, no. 4, pp. 82–89, 2014.

[23] Y. Zhao and Y. Z. Sha, "The Knowledge mapping analysis on the research of information science: based on ACA," *Library Tribune*, vol. 28, no. 6, pp. 63–69, 2008.

[24] J. M. Tang, "Descriptive research of excellent scientific organizations based on bibliometric-setting national educational courses for example," *Journal of Intelligence*, vol. 29, no. 4, pp. 5–9, 2010.

[25] P. A. Hook, "Domain maps: purposes, history, parallels with cartography, and applications," in *Proceedings of the 11th International Conference Information Visualization (IV '07)*, Zurich, Switzerland, July 2007.

[26] Y. Zhang, P. J. H. Hu, S. A. Brown, and H. Chen, "Knowledge mapping for rapidly evolving domains: a design science approach," *Decision Support Systems*, vol. 50, no. 2, pp. 415–427, 2011.

[27] H. Liu, Y. Liu, and F. Sun, "Robust exemplar extraction using structured sparse coding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 8, pp. 1816–1821, 2015.

[28] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: kernel sparse coding methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

[29] A. Altman and M. Tennenholtz, "Ranking system: the PageRank axioms," in *Proceedings of the ACM Conference on Electronic Commerce*, pp. 1–8, 2005.

[30] Y. Q. Qu, Q. R. Meng, S. X. Ma, L. Li, and G. L. Wu, "Geological characteristics of unconformities in Mesoproterozoic successions in the northern margin of North China Block and their tectonic implications," *Earth Science Frontiers*, vol. 17, no. 4, pp. 112–127, 2010.

[31] Natural Language Processing and Information Retrieval (NLPIR) Sharing Platform, http://www.nlpir.org/.

Hindawi Publishing Corporation
Scientific Programming
Volume 2017, Article ID 4769710, 8 pages
http://dx.doi.org/10.1155/2017/4769710

Research Article

3D Localization Algorithm Based on Voronoi Diagram and Rank Sequence in Wireless Sensor Network

Xi Yang, Fang Yan, and Jun Liu

School of Information, Beijing Wuzi University, Beijing, China

Correspondence should be addressed to Xi Yang; yxyoung@163.com

Received 23 September 2016; Revised 29 November 2016; Accepted 26 December 2016; Published 22 January 2017

Academic Editor: Wenbing Zhao

Accurate nodes' localization is a key problem in wireless sensor network (WSN for short). This paper discusses and analyzes the effects of Voronoi diagram in 3D location space. Then it proposes Sequence Localization Correction algorithm based on 3D Voronoi diagram (SLC3V), which introduces 3D Voronoi diagram to divide the 3D location space and constructs the rank sequence tables of virtual beacon nodes. SLC3V uses RSSI method between beacon nodes as a reference to correct the measured distance and fixes the location sequence of unknown nodes. Next, it selects optimal parameter N and realizes the weighted location estimate with N valid virtual beacon nodes by normalization process of rank correlation coefficients. Compared with other sequence location algorithms, simulation experiments show that it can improve the localization accuracy for nodes in complex 3D space with less measurements and computational costs.

1. Introduction

Due to its broad applications, researchers have been concerned with wireless sensor network (WSN). As one of the key technologies of the Internet of Things, WSN has been widely applied in various fields, such as military affairs, wise industries, environmental monitor, and smart home. The location information of wireless sensor nodes is needed in many applications. And location information can be used to self-organize, manage networks, select the optimal route, determine the occurrence area of the monitoring events, track moving targets, and so on. Therefore, node localization is attractive in wireless sensor network and plays a key role in the application of wireless sensor network [1–3].

Due to the characteristics of wireless sensor network, traditional localization methods in computer network cannot be directly applied, such as Global Positioning System (GPS).

Only a few sensor nodes can be equipped with GPS module or predeployed in a specific location if the operation cost and maintenance cost are taken into consideration. These nodes can know their own coordinates and assist other nodes to realize self-localization.

Localization accuracy directly affects the application effects, so great attentions have been paid to the localization problem. And many researchers have put forward many theories and efficient solutions.

In the past, most of the researches only paid much attention to the self-localization in 2D space with beacon nodes, which know their own location information beforehand (for example, [4, 5]). And current localization algorithms in 2D can be classified into two kinds: range-based algorithm and range-free algorithm [6, 7]. For range-based algorithms, distances are commonly computed by using different parameters such as time, angles, or signal strength and the location is estimated on the basis of the distances. For the range-free algorithms, unknown nodes calculate their approximate locations by using information from a few beacon nodes.

With the mature technology and market promotion, beacon nodes may be dynamic and localization will expand from the two-dimensional (2D) space to three-dimensional (3D) space.

As we all know, wireless sensor nodes are deployed in real environment, which is three-dimensional. And many applications often need three-dimensional location information. Localization in 3D space is more difficult than in 2D space. In addition, more factors should be considered in 3D localization, such as the environmental changes, the

insufficient number of beacon nodes, and various disturbing effects in the signal transmission process.

Now, 3D localization has become a current research trend and one of the hot problems. Many researchers extend 2D localization technologies into 3D space, such as tetrahedral centroid localization algorithm and 3D DV-Hop, APIT, RSSI location, and partial filter, which have achieved good results and have been used to some extent.

Currently, localization algorithms in 3D space can include two categories: hierarchical location and nonhierarchical location. In hierarchical location, nodes are mostly deployed in the interior of the monitoring buildings. In this condition, users or networks only need the floor information of unknown node replacing the specific Z coordinate value. Some of the proposals [8, 9] in this category are discussed in greater detail. However, nodes may be deployed underwater or in hills in nonhierarchical location. Users often need to estimate the specific Z coordinate value. Such approaches are depicted in [10–12].

In this paper, we introduce Voronoi diagram into localization algorithm and propose a new Sequence Localization Correction algorithm based on Voronoi diagram, which can be used in 3D space. The new algorithm divides location space with 3D Voronoi diagram, corrects the measured distance, and fixes location sequence of unknown nodes, which reduces the space partition complexity and raises localization accuracy. In order to reduce the effects of the number of real beacon nodes, SL3CV selects optimal parameter N to determine the number of valid virtual beacon nodes in the last localization estimation, which also improves the localization accuracy. The localization estimation of unknown nodes can be calculated through the weights based on the optimal location sequence table of virtual beacon nodes.

This paper includes 5 sections. The concepts of 3D Voronoi diagram and location sequences are described in Section 2. Section 3 describes and analyzes the localization procedures of SLC3V algorithm in 3D space. This paper analyzes and compares the proposed localization algorithm with other algorithms through an exhaustive systematic performance study and simulation experiments in Section 4. Finally, this paper concludes in Section 5.

2. Related Techniques

2.1. 3D Voronoi. There are many space partition methods. Voronoi diagram is a kind of partition method, which divides the space into a number of subregions. Now Voronoi diagram is widely used in various fields, such as geographical information system, information system, and meteorology. Many researchers use Voronoi diagram to study the coverage problem in WSN.

Voronoi diagram divides the plane into many regions. We are given a finite set of points $\{p_1, \ldots, p_n\}$, $2 \leqslant n < \infty$, in the Euclidean plane. Let $X(x, y)$ be any point in the space. Let $d(p_i, X)$ denote the Euclidean distance with $X(x, y)$ and $p_i(x_i, y_i)$, $i \in I = \{1, 2, \ldots, n\}$. Then

$$d(p_i, X) = \sqrt{(x_i - x)^2 + (y_i - y)^2}. \quad (1)$$

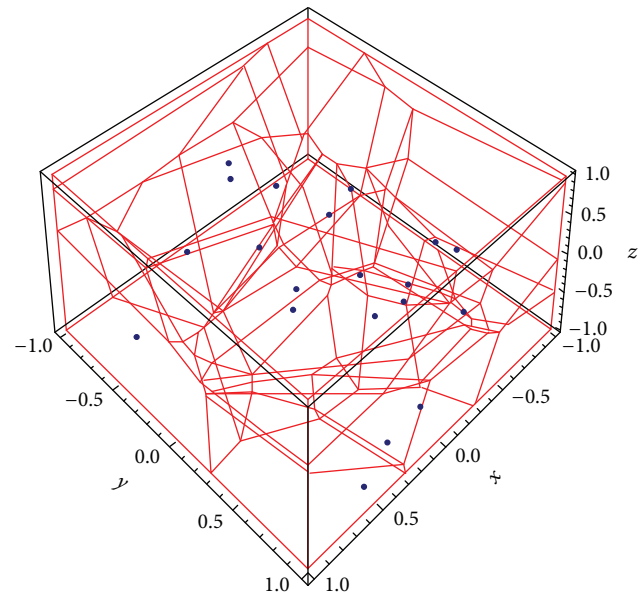


Figure 1: Voronoi diagram in 3D space.

Voronoi diagram divides the plane into n regions (Voronoi region) around each generator point p_i, which makes any point X in Voronoi region $V(p_i)$ satisfy the following condition:

$$V(p_i, X) = \left\{x \mid d(p_i, X) \leqslant d\left(p_j, X\right),\ \forall i \neq j \wedge i, j \in I\right\}. \quad (2)$$

Then $V(P)$ is the sets of Voronoi region $V(p_i)$.

We find that 2D Voronoi diagram includes many continuous polygons, which are composed of a set of perpendicular bisectors. And all perpendicular bisectors are vertical to the connection lines between two adjacent points. With the increase of dimension, the formation of Voronoi cells has changed into high dimensional polyhedron. Therefore a Voronoi cell in 3D space is a 3D polyhedron.

After divided, any point in any given Voronoi cell is closer to the corresponding Voronoi site than other Voronoi sites. And all Voronoi cells combine together without overlapping and seams in [13, 14].

Because of the practicability of 3D Voronoi diagram, there are many computational methods developed to divide 3D discrete point set. In order to decrease the computational complexity, many researchers proposed various fast generation methods of 3D Voronoi diagram.

Figure 1 shows the 3D Voronoi diagram of 20 scattered points in the 3D closed space, in which the dots are 20 discrete points randomly deployed.

2.2. Sequence-Based Localization. Recently, some researches propose sequence-based localization method, which is efficient by combining range-based algorithm and range-free algorithm, but it is used in almost 2D space [15, 16]. The procedure of sequence-based localization in 2D space is given as follows:

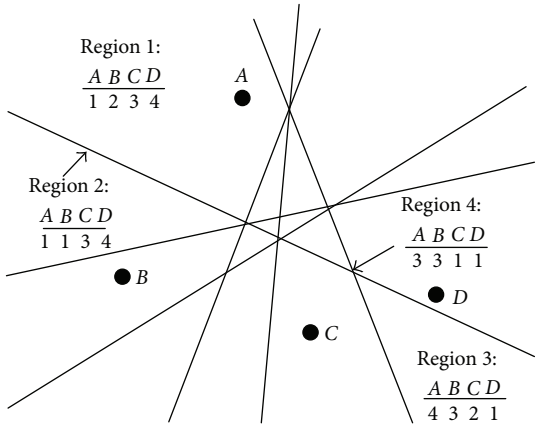


Figure 2: Example of sequence for four reference nodes.

(1) Construct the boundary of 2D space.

(2) Divide 2D space into three kinds of region, such as vertices, edges, and faces by a set of perpendicular bisectors between all beacon nodes.

(3) Compute the centroids of every region, which are called virtual beacon nodes, and the distance between virtual beacon nodes and real beacon nodes.

(4) Determine all location sequences of virtual beacon nodes in the localization space and rank them in a location sequence table.

(5) Calculate the rank correlation coefficient between the location sequence of unknown node and the other sequence of virtual beacon nodes in the location sequence table and search for the maximum value as the nearest location sequence to the unknown node location sequence. The centroid mapped to by that sequence is the localization estimate of the unknown node.

Figure 2 shows an example of sequence for four reference nodes. In the example, the four predefined nodes are scattered in the plane. Region 3 is a face, and its location sequence is 4321. Similarly, Region 2 is an edge, and its location sequence is 1134. Similarly, Region 4 is a point, and its location sequence is 3311.

Compared with other RSSI localization methods, sequence-based localization algorithm does not use RSSI to compute distance. RSSI is used to determine the location sequence of the unknown node.

In three-dimensional space, we extend the 2D sequence-based algorithm and divide 3D space into three kinds of region, such as edges, faces, and bodies by a set of min-vertical planes.

3. SLC3V Algorithm

Based on 2D sequence-based localization algorithm, 3D sequence-based orthocenter localization algorithm (3DSBO for short) is proposed in [17]. However, we find that the space division methods are not optimal and localization errors are rather large with the decrease in the number of real beacon nodes while they are high-accuracy localization technique. And the errors will increase if we do not consider the environmental influences on the estimation of the location sequences. And 3DSBO only use three virtual beacon nodes to estimate the location of unknown nodes.

Voronoi diagram has high order mode, so we consider the 3D application scenarios. While 3D Voronoi diagram has been introduced into sequence localization [18], SL3V cannot correct location sequence of the unknown node according to the environment. Furthermore, SL3V chooses the optimal polyhedron and calculates the coordinate with the virtual beacon nodes in the optimal polyhedron, which increases the localization error.

This paper extends Sequence Localization Correction algorithm based on 3D Voronoi diagram (SLC3V), which considers dividing the 3D space by using 3D Voronoi diagram. After partition, SLC3V determines the centroids of three kinds of regions in the polyhedrons, which are called virtual beacon nodes. Then it computes the distance between virtual beacon nodes and real beacon nodes. According to the distance rank, we can form the location sequence table of all beacon nodes, including real beacon nodes and virtual beacon nodes. Then we compute and correct the location sequence table of the unknown nodes by using RSSI method. Therefore the location sequence table of the unknown nodes will be closer to the truth.

By estimating the coordinate of real beacon nodes with Kendall coefficients, it selects the optimal parameter N, which decides the valid number of virtual beacon nodes used in localization estimation. Finally, it produces the optimal localization sequences of N valid virtual beacon nodes and estimates the coordinate of the unknown node with Kendall coefficients between the sequences in optimal localization sequence table and the unknown node sequence.

3.1. Division of 3D Space. In the 3D space, we construct the cube boundary according to the location environment. The real beacon nodes are scattered in the 3D space randomly. The real beacon nodes are the real nodes deployed by users.

Next we use the 3D Voronoi diagram to divide the 3D space. Then the Voronoi polyhedrons are produced. Each real beacon node is interior of the corresponding Voronoi polyhedron. In order to reduce the complexity of partition, we use the fast method in [13] to generate 3D Voronoi diagram.

After partition, 3D space includes three kinds of regions, such as edges, faces, and bodies by a set of min-vertical planes.

We present an example to illustrate the above ideas. As shown in Figure 3, there are 3 real beacon nodes defined as A, B, and C, which are marked by the solid dotes.

3.2. Location Sequence Calculation. Firstly, we should determine all virtual beacon nodes based on the partition method. This paper marks all the regions and calculates the centroid of every region, which is defined as virtual beacon nodes. Then we sort all virtual beacon nodes. Because the location space is 3D, the kinds of regions change from vertices, edges, and faces into edges, faces, and bodies.

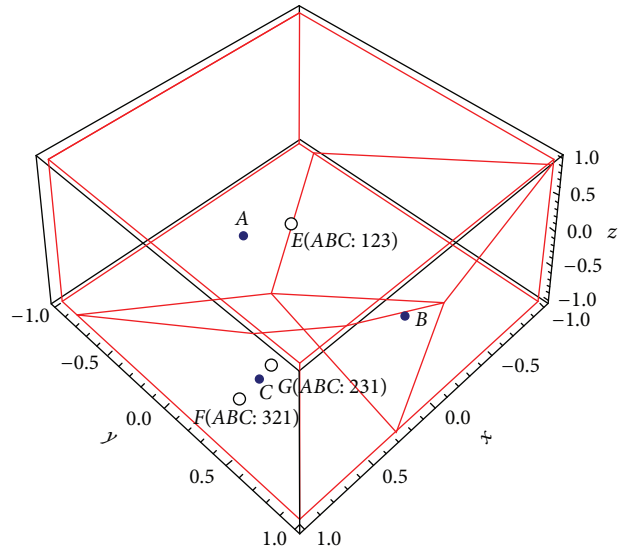


Figure 3: Real beacon nodes in 3D space.

For simplifying the calculation, we define the centroid of every region as follows. The centroid of any edge is its midpoint. The centroid of any face is calculated according to [15]. The centroid of any body is midpoint of the line between the two farthest points in the corresponding polyhedron.

Then we can rank the location sequences of virtual beacon nodes according to distance between the virtual beacon nodes and real beacon nodes.

In the 3D space, the positions of real beacon nodes are known, and all virtual beacon nodes are determined after partition. So we can calculate the distance between any virtual beacon nodes and all real beacon nodes. Then we can assign the location sequence table values based on distance, which is the indication of how near or far the virtual beacon node is from all real beacon nodes.

In the example of Figure 3, hollow dotes, such as E, F, and G, represent virtual beacon nodes. And each virtual beacon node has its own location sequence based on the distance rank, which is marked beside it in Figure 3. Node E is the centroid of edges, and its location sequence is 123, since the distance rank of A from it is 1 (A is the closest), and the respective distance rank of C is 3 (C is the farthest). Similarly, node G is the centroid of bodies, and its location sequence is 131. Node F is the centroid of faces, and its location sequence is 321.

After this step, the location sequence table of all virtual beacon nodes can be calculated in any 3D space with real beacon nodes.

3.3. Location Sequence Correction of Unknown Nodes. In order to obtain the distance rank of unknown nodes, we assume that the real beacon nodes send data package to its neighbor nodes. And all unknown nodes are in the transmission area of real beacon nodes. So the unknown nodes measure the distance to real beacon nodes by using the Received Signal Strength Indicator (RSSI) method. According to the distance measured, any unknown node can determine its own location sequence. If location sequences of unknown nodes are more accurate, the localization estimate of unknown nodes has better accuracy. So we will use the distance and RSSI between real beacon nodes as a reference to correct the measured distance and fix location sequence of unknown nodes.

RSSI_i is the average RSSI (dbm) of the signal, which unknown node X receives from beacon node B_i. P_i is the average power (mW) of the signal which unknown node X receives from beacon node B_i. So the transformational relation between RSSI_i and P_i is shown in

$$P_i = 10^{\text{RSSI}_i/10}. \tag{3}$$

RSSI_{ij} is the average Received Signal Strength Indicator (RSSI) (dbm) of the signal, which real beacon node B_i receives from real beacon node B_j. P_{ij} is the average power of the signal which real beacon node B_i receives from real beacon node B_j. Similarly, the transformational relation between RSSI_{ij} and P_{ij} is shown in

$$P_{ij} = 10^{\text{RSSI}_{ij}/10}. \tag{4}$$

d_{Bij} is the distance between real beacon node B_i and real beacon node B_j, which is deduced according to (1). d_i^j is the distance between unknown node X and real beacon node B_i. Then, we can deduce the relation equation as follows:

$$d_i^j = \frac{P_{ij}^{1/\eta} \times d_{Bij}}{P_i^{1/\eta}}. \tag{5}$$

So, d_i represents the average value of d_i^j, which is the distance between unknown nodes X and real beacon node B_i. R is the number of real beacon nodes.

$$d_i = \frac{\sum_{j=1}^{R} d_i^j}{R}. \tag{6}$$

From the above definition, location sequence of unknown node is deduced and corrected as follows:

(1) All real beacon nodes broadcast information in the same power periodically. Then unknown nodes will receive RSSI value of each real beacon node and calculate respective average value RSSI_i. Real beacon node B_i also can calculate the corresponding RSSI_{ij} of real beacon node B_j.

(2) All real beacon nodes broadcast P_{ij} and d_{Bij}, which are calculated by using (4). Then unknown nodes will receive information and calculate d_i. Finally, location sequence of unknown nodes can be decided according to the fixed d_i.

Because of the RSSI correction, unknown nodes can correct their measured distance according to the environment. Thus the algorithm in this paper can fix location sequence of unknown node and ensure location sequence consistent with the actual situation.

3.4. Optimal Parameter N Selection. In this step, we calculate the order correlation coefficients between the vectors in sequence table of virtual beacon nodes and the unknown node sequence, respectively.

Statistics offers two metrics to calculate the rank correlation coefficient between two location sequences. They are Spearman's Coefficient and Kendall's Tau, respectively. According to the analysis [15] and our simulation results, Kendall's Tau can obtain better performance in localization algorithm. So we describe it simply. Kendall's Tau is often used to calculate the correlation between the two sequences. Given two location sequences $U = \{u_i\}$ and $V = \{v_i\}$, $1 \le i \le n$, where u_i and v_i represent the ranks of different beacon nodes, Kendall's Tau compares all the $n(n-1)/2$ possible pairs of ranks (u_i, v_i) and (u_j, v_j), respectively, to determine the numbers of matching pairs and nonmatching pairs. A pair is matching or concordant if $u_i > u_j \Rightarrow v_i > v_j$ or $u_i < u_j \Rightarrow v_i < v_j$ and nonmatching or discordant if $u_i > u_j \Rightarrow v_i < v_j$ or $u_i < u_j \Rightarrow v_i > v_j$. The correlation between two sequences is calculated as follows:

$$\tau = \frac{(n_c - n_d)}{\sqrt{n_c + n_d + n_{tu}}\sqrt{n_c + n_d + n_{tv}}}, \tag{7}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, n_{tu} is the number of ties in U, and n_{tv} is the number of ties in V. The range of τ is $[-1, 1]$.

At the same time, we should calculate the Kendall coefficients between the vectors in sequence table of virtual beacon nodes and the real beacon nodes sequence, respectively.

Compared with other algorithm, this paper uses parameter N to determine the optimal number of virtual beacon nodes that are used in the weighted estimation. The suitable parameter N can decrease the localization error. We define all real beacon nodes as the sample set S and select the optimal parameter N according to localization estimation of real beacon nodes. We calculate the optimal parameter N_j value of jth real beacon node. Then the final optimal parameter N is the average of N_j.

For example, given the location sequence U of any real beacon node $Y_j(x, y, z)$, $Y_j \in S$, SLC3V obtains the location sequence table of virtual beacon nodes $T = \{T_1, T_2, \ldots\}$.

According to (7), we can calculate the correlation coefficients $\tau(T_k)$ of location sequence table U and T_k. $\tau(T_k)$ reflects real beacon node Y_j's proximity to the sequence T_k. Then we can normalize $\tau(T_k)$ by altering its range from $[-1, 1]$ to $[0, 1]$. The normalized $\tau(T_k)$ is used as weights. Then we can estimate the coordinate of real beacon node Y_j, respectively.

Therefore, the Mth estimation coordinate of real beacon node Y_j is calculated by selecting M largest coefficients as follows:

$$\widehat{Y}_{jM}\left(\widehat{x}_{jM}, \widehat{y}_{jM}, \widehat{z}_{jM}\right) = \frac{1}{\sum_{k=1}^{M}\left(\left(1 + \tau\left(T_k\right)\right)/2\right)} \sum_{k=1}^{M}\left(\frac{1 + \tau\left(T_k\right)}{2} C_k\right), \tag{8}$$

where C_k is the coordinate of the kth location sequence T_k. M is any integer values in the range of $[1, N_{\max}]$. $N_{\max}$ is the constant predefined by users, such as 10 or the numbers of virtual beacon nodes. So we have $N_{\max}$ estimation coordinates of real beacon node Y_j.

Then we calculate the localization error of any estimation coordinates of real beacon node Y_j as follows:

$$E\left(Y_{jM}\right) = \left(x_j - \widehat{x}_{jM}\right)^2 + \left(y_j - \widehat{y}_{jM}\right)^2 + \left(z_j - \widehat{z}_{jM}\right)^2. \tag{9}$$

According to the same steps, we calculate $\sum_j E(Y_{jM})$ of all real beacon nodes. Then N is the optimal parameters of all real beacon nodes in $[1, N_{\max}]$, which makes the localization error $\sum_j E(Y_{jM})$ minimum. Once N is calculated, other unknown nodes are estimated with the same N.

$$N = \arg \min_{M \in [1, N_{\max}]} \sum_j E\left(Y_{jM}\right). \tag{10}$$

3.5. Weighted Estimation. After determining the optimal parameters N, we produce optimal location sequence table, which has N largest Kendall coefficients between all virtual beacon nodes and the unknown node sequence.

We use the normalized Kendall coefficients between the sequences in optimal location sequence table and the unknown node location sequence as the weights and estimate the coordinate of the unknown node.

For example, given the localization sequence W of unknown node $X(x, y, z)$, SLC3V algorithm constructs an optimal location sequence table $T = \{T_1, T_2, \ldots, T_N\}$. Then we can calculate the coefficient $\tau(T_i)$, which is the one of each sequence in location sequence W and T_i.

Therefore, unknown node X's coordinate is estimated as follows:

$$X\left(\widehat{x}, \widehat{y}, \widehat{z}\right) = \frac{1}{\sum_{i=1}^{N}\left(\left(1 + \tau\left(T_i\right)\right)/2\right)} \sum_{i=1}^{N}\left(\frac{1 + \tau\left(T_i\right)}{2} C_i\right). \tag{11}$$

3.6. Periodic Update of Network Topology. By using the above procedures, all unknown nodes can be estimated. When the network topology changes, the algorithm will regularly update and delete the invalid beacon node. The localization calculation restarts from space division.

3.7. Complexity and Analysis. The main advantage of SLC3V algorithm is that better partition method reduces the complexity and introduction of valid virtual beacon nodes raises the location accuracy.

The step of space division mainly decides Voronoi polyhedrons, which takes $O(n^2)$ time and space each. Then the steps of localization sequence construction require $O(n \log n)$ operations. Location sequence of unknown nodes costs some time and energy, but it only has the relation with the number of beacon nodes. For calculating Kendall's coefficients between two sequences, there are $O(N^2)$ operations. And searching through it takes $O(N^2 n^2 \log n)$ time and $O(N^2 n^3)$ space. Since the step of weighted estimation includes the normalization and estimation, it takes $O(N^2)$ time and $O(N)$ space.

Therefore the worst case time requirement in SLC3V is $O(N^2n^4)$ and the worst case space requirement is $O(N^2n^3)$.

However, the 3D sequence-based localization in [17] takes $O(n^6)$ time and $O(n^5)$ space in the worst case. SLC3V algorithm has improved in the algorithm complexity. SLC3V and SL3V have almost the same time and space complexity.

4. Simulation and Evaluation

4.1. Simulation Model. In wireless channels, there is the log-normal shadowing model that is most common simulation model. The log-normal shadowing model generates RSS samples as a function of distance.

$$P_R(d) = P_T - \mathrm{PL}(d_0) - 10\eta \lg \frac{d}{d_0 + X_\sigma}, \quad (12)$$

where P_R is the power of received signal, which defines Received Signal Strength Indication (RSSI). P_T is the power of transmitted signal, and $\mathrm{PL}(d_0)$ is the path loss for a reference distance of d_0. η is the path loss exponent. And we define its random variation as a Gaussian random variable of zero mean and σ^2 variance $X_\sigma = N(0, \sigma^2)$. The unit of power is dBm, and the one of distances is meter.

However, this model shows how signals attenuate ideally. In fact, there are many nonideal factors taken into account, such as various obstructions and electromagnetic interferences. If the situations mentioned above are to be considered, then (12) should be modified by introducing extra parameters in the right-hand side.

4.2. Simulation Environment. For simplicity, we simulate SLC3V algorithm and analyze its performance in Matlab. We assume that whether beacon nodes or unknown nodes are in the mutual radio range of other nodes, which means that all nodes can receive the message sent by other nodes. And node communication radius is 8 m. A 48-bit arithmetic linear congruential pseudorandom number generator was used, and results were averaged over 100 random trials. We choose $N_{\max}$ is 10.

P_e represents the localization error in this paper and is defined as follows:

$$P_e = \frac{|\text{estimate location} - \text{real location}|}{\text{Wireless communication radius}} \times 100\% \quad (13)$$

In order to compare and evaluate SLC3V algorithm, this paper analyses the two performances, such as localization error and computation time. Then we compare our algorithm with other sequence-based localization algorithms. Firstly, we implement 3DSBO [17] and SL3V [18] in the same environment.

4.3. Simulation Results of SLC3V. Average computation time in 3 algorithms above is in consistency with our time complexity analysis in Section 3.7, as described in Table 2.

When the number of real beacon nodes raises to 32, SLC3V algorithm has obvious advantages in computation time, which comes about because of its partition method.

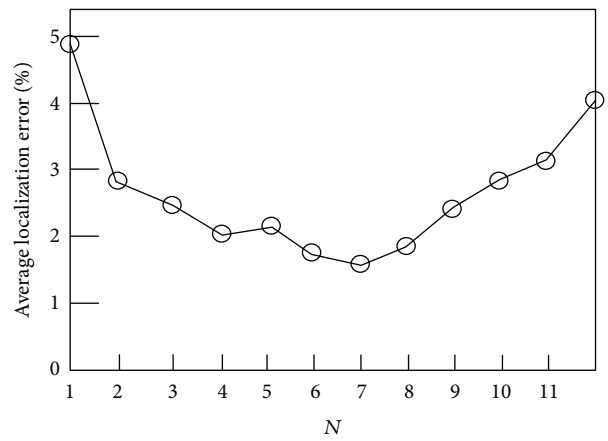


Figure 4: Average localization error in different N.

Table 1: Average localization error in 3D.

Number of beacon nodes	3DSBO (%)	SL3V (%)	SLC3V (%)
8	18.9	18.3	12.0
16	9.6	9.4	8.9
24	4.9	4.8	4.3
32	2.5	2.4	1.9

Table 2: Average computation time in 4 algorithms.

Number of beacon nodes	3DSBO (%)	SL3V (%)	SLC3V (%)
8	0.52	0.46	0.40
16	1.33	1.01	1.00
24	2.59	2.10	2.12
32	3.80	2.80	2.81

Due to the better partition method, the space complexity of SLC3V algorithm is lower. Location sequence correction and optimal parameter N are introduced to make our average localization error the lowest in the above 3 algorithms. Whether increasing or decreasing the real beacon nodes, SLC3V algorithm has higher location accuracy, as shown in Table 1. Because SLC3V selects valid virtual beacon nodes in last estimation this algorithm raises average localization error to 12.0% when it only has 8 real beacon nodes, which show that it reduces the rising rate of localization error. Compared with SL3V, SLC3V fixes the location sequence with the RSSI correction and select optimal number N. So SLC3V has higher accuracy.

In order to verify the effect of the optimal N, average localization errors are calculated in different N through extensive experiments. Figure 4 draws the curve of average localization errors with 32 real beacon nodes. The average localization error is lowest when optimal N is 7 calculated by SLC3V.

Figure 5 plots the average localization error due to 3DSBO, SL3V, and SLC3V as a function of the standard

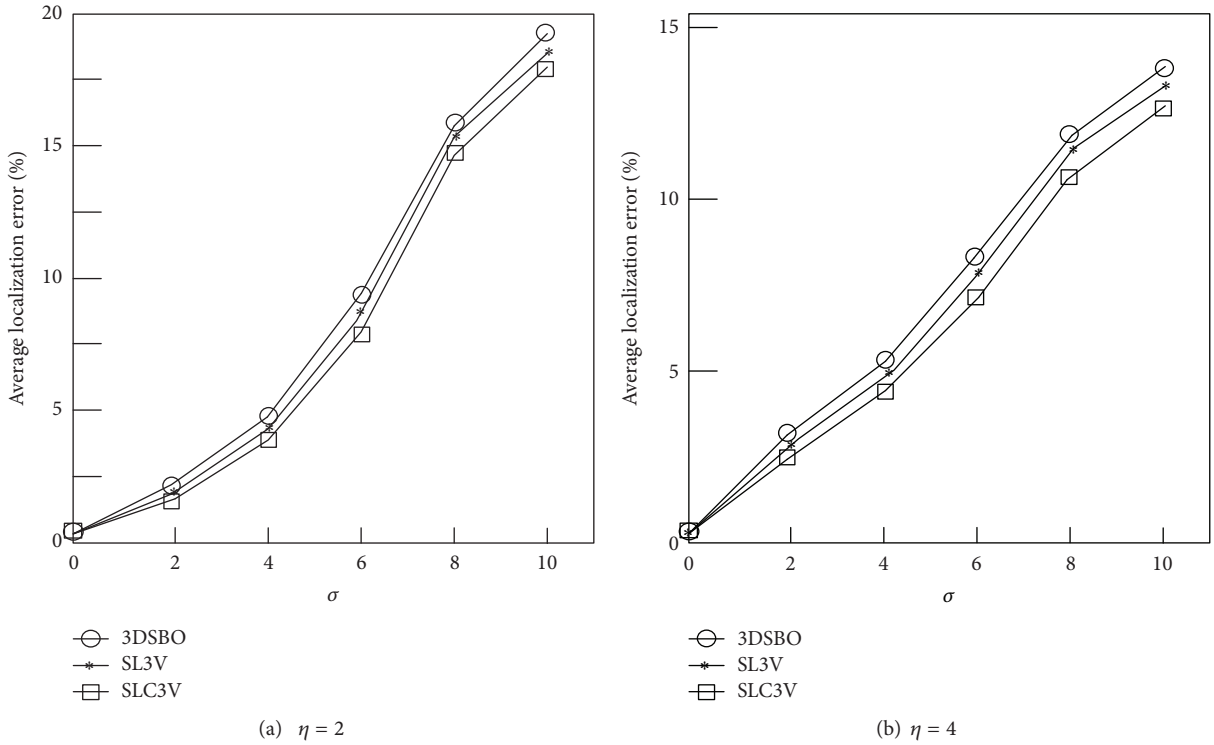


Figure 5: Average localization error curves with standard deviation σ.

deviation in RSS log-normal distribution σ for different values of path loss exponents η, where the number of beacon nodes is 8. So we find that SLC3V is better than others under changing environment.

In Figures 5(a) and 5(b), path loss exponent η is, respectively, 2 and 4. Average localization error using SLC3V is always lower than that in 3DSBL in different standard deviation σ. And For higher values of σ, the localization error due to SLC3V decreases faster than localization error due to 3DSBO. The main results show that SLC3V performs better than 3DSBO in the antinoise and anti-interference properties.

Based on the above experimental results of two performance parameters, SLC3V algorithm has better performances in the same conditions compared with other 3D sequence-based location algorithms.

5. Conclusion and Future Work

Localization in wireless sensor networks is still a critical research issue for researchers recently. Most researchers concentrate on the study in 2D space. With the wide application and proliferation of WSNs, the problem of localizing nodes in 3D space has received much attention.

Based on SL3V, this paper proposes an efficient algorithm localization algorithm called Sequence Localization Correction algorithm based on 3D Voronoi diagram (SLC3V), which combines 3D Voronoi diagram with the sequence localization in WSN.

SLC3V algorithm extends 2D sequence-based location into 3D space. SLC3V not only can realize unknown node self-localization, but also overcomes the original sequence-based localization algorithm. The shortcomings of this original sequence-based localization algorithms are the higher partition complexity, the higher location error in complicated disturbing environment, and the lower constant number of valid virtual beacon nodes. SLC3V algorithm adapts 3D Voronoi diagram to divide the space. So it effectively reduces the number and the complexity of generating the virtual beacon nodes. Considering the variability of the environment and the impact of outside interference, we introduce the RSSI correction method to adapt the location sequence of the unknown node rapidly. The selection of optimal parameter N increases valid virtual beacon nodes in location estimation compared with other algorithms, which improve the location accuracy.

In order to verify the effectiveness of SLC3V, we deploy real beacon nodes randomly in the 3D cube during simulation. Experimental results prove that the proposed algorithm has good performance in either more beacon nodes or fewer beacon nodes. It can be adaptive according to the number of beacon nodes and changes in environment and network.

In the future, we will build test-bed and deploy real nodes on it. The algorithm performance is tested on test-bed in laboratory. Then we will optimize the algorithm further and apply it in practical applications, such as the surveillance and control on food storage.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This paper is supported by Beijing Key Laboratory (no. BZ0211).

References

[1] X. Luo, D. D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, "A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems," *Future Generation Computer Systems*, vol. 61, pp. 85–96, 2016.

[2] Z. CH. Ma, Y. N. Sun, and T. Mei, "Summary for wireless sensor network," *Journal of China Institute of Communications*, vol. 25, no. 4, pp. 114–124, 2004.

[3] X. Luo, H. Luo, and X. H. Chang, "Online optimization of collaborative web service QoS prediction based on approximate dynamic programming," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 452492, 9 pages, 2015.

[4] Z.-H. Qian, D.-Y. Sun, and V. Leung, "A survey on localization model in wireless networks," *Chinese Journal of Computers*, vol. 39, no. 6, pp. 1237–1253, 2016.

[5] Y. H. Liu, Z. H. Yang, X. P. Wang, and L. R. Jian, "Location, localization, and localizability," *Journal of Computer Science and Technology*, vol. 25, no. 2, pp. 274–297, 2010.

[6] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies*, vol. 3, San Francisco, Calif, USA, April 2003.

[7] J. Cai and S. Chen, "Simple and fast convex relaxation method for cooperative localization in sensor networks using range measurements," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4532–4543, 2015.

[8] Q. Z. H. Chen, K. J. Mao, W. X. He, and X. M. Zhao, "3D localization algorithm based on degree of coplanarity and layered structure for wireless sensor networks," *Journal of Electronic Measurement and Instrument*, vol. 26, no. 8, pp. 673–681, 2012.

[9] A. Aziz, R. Kumar, I. Joe, and Y. Choi, "A cooperative scheme of 3D localization with flying wireless sensor nodes for disaster situations," *Journal of Next Generation Information Technology*, vol. 6, no. 2, pp. 37–45, 2015.

[10] Y. Zhang, S. Liu, and Z. Jia, "Localization using joint distance and angle information for 3D wireless sensor networks," *IEEE Communications Letters*, vol. 16, no. 6, pp. 809–811, 2012.

[11] R. J. Feng, X. L. Guo, J. W. Wan, Y. F. Wu, and N. Yu, "Multihop localisation with distance estimation bias for 3D wireless sensor networks," *Electronics Letters*, vol. 48, no. 14, pp. 884–886, 2012.

[12] E. Kim, S. Lee, C. Kim, and K. Kim, "Mobile beacon-based 3D-localization with multidimensional scaling in large sensor networks," *IEEE Communications Letters*, vol. 14, no. 7, pp. 647–649, 2010.

[13] X. N. Liu, "Implementation of Voronoi diagram algorithm for 3D point set," *Computer-Aided Engineering*, vol. 15, no. 1, pp. 1–3, 2006.

[14] J. Y. Liu and S. H. Liu, "A survey on Voronoi diagram," *Journal of Engineering Graphics*, vol. 25, no. 2, pp. 125–132, 2003.

[15] K. Yedavalli and B. Krishnamachari, "Sequence-based localization in wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 1, pp. 81–94, 2008.

[16] Z.-H. Liu, J.-X. Chen, and X.-K. Chen, "A new algorithm research of sequence-based localization technology in wireless sensor networks," *Acta Electronica Sinica*, vol. 38, no. 7, pp. 1552–1556, 2010.

[17] J. X. Chen and Z. H. H. Liu, "Sequences and centriods localization for 3D WSN," *Computer Engineering and Applications*, vol. 47, no. 2, pp. 81–83, 2011.

[18] X. Yang and J. Liu, "Sequence localization algorithm based on 3D voronoi diagram in wireless sensor network," *Applied Mechanics and Materials*, vol. 644-650, pp. 4422–4426, 2014.

Hindawi Publishing Corporation
Scientific Programming
Volume 2017, Article ID 7594763, 11 pages
http://dx.doi.org/10.1155/2017/7594763

Research Article

Development of a Wearable Device for Motion Capturing Based on Magnetic and Inertial Measurement Units

Bin Fang, Fuchun Sun, Huaping Liu, and Di Guo

State Key Lab of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Haidian District, Beijing 100083, China

Correspondence should be addressed to Bin Fang; fangbin1120@163.com

Received 22 July 2016; Revised 26 September 2016; Accepted 17 October 2016; Published 18 January 2017

Academic Editor: Michele Risi

This paper presents a novel wearable device for gesture capturing based on inertial and magnetic measurement units that are made up of micromachined gyroscopes, accelerometers, and magnetometers. The low-cost inertial and magnetic measurement unit is compact and small enough to wear and there are altogether thirty-six units integrated in the device. The device is composed of two symmetric parts, and either the right part or the left one contains eighteen units covering all the segments of the arm, palm, and fingers. The offline calibration and online calibration are proposed to improve the accuracy of sensors. Multiple quaternion-based extended Kalman filters are designed to estimate the absolute orientations, and kinematic models of the arm-hand are considered to determine the relative orientations. Furthermore, position algorithm is deduced to compute the positions of corresponding joint. Finally, several experiments are implemented to verify the effectiveness of the proposed wearable device.

1. Introduction

The gesture is a natural and efficient way for communication and plays an important role in human-human interaction to express and transmit information. Therefore, the technology of gesture recognition has become a hot research topic. Through gesture recognition, results are used to communicate with computers, assess the kinematics of the hand, control electronic equipment, and so on. It has been wildly used in several application areas, such as rehabilitation, sports, and animation industry [1, 2]. In order to identify human gestures, we need to obtain the positions, speeds, directions, and other information of the gesture by a certain gesture capture technology.

At present, there are two main kinds of motion capture technology, namely, vision based and contact based devices [3, 4]. Vision based devices capture the video streams for analysis to determine the hand motion. On the other hand, contact based devices depend on physical interaction with the user. Based on vision gesture method, users generally do not need to wear collection equipment and can move more freely. But it is more susceptible to background, like illumination, occlusion, and other environmental factors. Furthermore, the camera frame rate and placement requirements are higher [5]. In comparison, contact based devices are easy to implement. Examples of contact based devices are mobile touch screens, EMG-based devices, and data gloves. EMG-based device is portable, but EMG signal is easily influenced by acquisition placement, individual differences, physical condition, and other factors. Moreover, fine finger and hand motion is still difficult to determine [6]. Data gloves use multiple sensor such as optical fiber sensor, pressure resistance sensor, pressure sensor, magnetometer, accelerometer, and gyroscope, to perceive the movement [7]. This kind of method can well reflect the spatial motion trajectory, attitude, and time sequence information of the hand by capturing the position, direction, and angle of the finger, which are restricted to the environmental conditions. But, at present, the product price is very expensive, so the development of low-cost data gloves has become the goal of our research. Different types of sensory gloves have been developed overtime, both commercial and prototype ones. The commercial products [8] usually use expensive motion-sensing fibers and resistive-bend sensors and are too costly for the consumer market [9]. Consequently, the prototype data gloves are developed to lower the cost of such equipment [10]. The flex sensors or bending sensors are integrated into the data gloves. However, the above sensors just measure the

relative orientation of articulated segments by mounting the sensor across the joint of interest. This requires an accurate alignment of sensors with a particular joint. Moreover, recalibration is necessary to mitigate estimation errors due to the sensor displacements. General disadvantages of the data gloves are the lack of user customization for individual subject' hands and obstruction of tactile sensing from the palmar surface of the hand. Often this inherently goes with the lack of mounting space for embedding the sensors in the cloth. To overcome these shortcomings, the inertial and magnetic sensors are induced.

In recent years, the MEMS technology has developed tremendously. The microinertial sensors have so many advantages like low-cost, small size, low-power consumption, large dynamic range, and so on. It has gradually become one of the most popular sensors for human motion capturing [11]. Meanwhile, the magnetic sensors are commonly used together with inertial sensors for accurate and drift-free orientation estimation [12]. Inertial and magnetic measurement unit (IMMU) has been proved to be an accurate approach to estimating the orientations of body segments without the external cameras [13]. It is nonobtrusive, comparably cost effective, and easy to setup and use. It also demonstrates higher correlation and lower error compared with a research-used visual motion capture system when the same motions are recorded [14]. Furthermore, the wearable inertial and magnetic sensors are becoming increasingly popular for the gesture motion capturing.

Gesture motion capture device based on microinertial sensors can generally be divided into three types, which are hand-hold type [15], wrist-wearable type [16], and glove type [17]. Compared with the hand-hold and wrist-wearable types, the glove type is available for having more numbers and types of inertial sensors and provides more accurate results of gestures. Therefore, the forms of gloves are most commonly used. The KHU-l data glove [17] consists of six three-axis accelerometers, but it can only capture several kinds of gestures. In [18], the data glove is developed based on sixteen microinertia sensors, which can capture the movements of each finger and palm, but the information of the heading angle is missing. In [19], the inertial and magnetic measurement unit is used, but it only uses four inertial and magnetic measurement units, which is unable to obtain the information of each finger joint. Power Glove [20] is developed, which includes six nine-axis microinertial sensors and ten six-axis microinertial sensors. It covers each joint of the palm and fingers, and motion characteristics can be better evaluated. However, it does not make full use of nine microinertial sensors, and in some state, the heading angle solution is instable so that it may lead to the estimation errors of the joint angle. The research shows that the current gesture capture device does not take into account the motion of the arm, and, at the same time, the movement of the hand cannot be fully captured. Therefore, we use the inertial and magnetic measurement units to develop a new gesture capture device, which can fully capture the motion information of the fingers, hands, and arms.

On the other hand, the IMMU-based device should be focused on the following two main aspects: calibration and fusion algorithm. Calibration is important for improving the performance of IMMU. Generally, there are two phases that included inertial sensors calibration [21] and magnetometers calibration in field [22]. The inertial sensors are of low cost and low precision, the deterministic errors such as bias and scale factor exist inevitably, and the calibration methods should be designed to improve the accuracy. Moreover, the magnetometers in the field would be affected by the iron-based materials, which generate their own magnetic field. Therefore, the measurements of magnetometers are the combination of earth's magnetic field and the extra magnetic field caused by the environmental effects. Hence, magnetometers calibration should be implemented for lessening the disturbance and improving the azimuth angle. Fusion algorithm is another key procedure to estimate orientations by combining the signals of gyroscopes, accelerometers, and magnetometers. The Kalman filter is a useful tool for sensor fusion. The extended Kalman filter is [23] a general method for estimating orientations and has been applied in the products of AHRS [24]. Nevertheless, the EKF is not easy to choose the appropriate parameters and hard for computation. Then the complementary filter that combines two independent noisy measurements of the same signal is proposed [25]. Complementary filters are the high-pass signals provided by gyroscopes and the data from low-pass accelerometers and magnetometers which provide relatively accurate measurements at low frequencies fused to estimate the true orientation. Reference [26] proposed a nonlinear invariant observer, with respect to the symmetries of the system equations. In addition, the deterministic algorithm is another class for estimating the orientations.

The novel wearable device proposed is developed for gesture capturing based on the IMMUs. The paper is organized as follows. Section 2 presents the designs of the wearable device. Section 3 describes the calibration of the sensors and the algorithms of attitude estimation, and the position estimation is deduced. Section 4 reports the results of the calibration, the orientations, and gesture capturing experiments, which verify the effectiveness of the proposed device. Section 5 gives the conclusions of the paper.

2. Wearable Device Design

2.1. Inertial and Magnetic Measurement Unit Design. One of the major designs of the wearable device is the development of the low-cost inertial and magnetic measurement units. Commercial IMMU commonly contains processing units and transceiver modules except the MEMS inertial and magnetic sensors. This increases the weight and packaging size; hence, it is not suitable to use these IMMUs to put on the right positions of the fingers or arms. Moreover, it is difficult to add more IMMUs with small distance in order to benefit from redundant measurements or gain more accurate measurements of fingers because of the structure and size of the unit. Here, the MPU9250 [27] that deploys system in package technology and combines 9-axis inertial and magnetic sensors in a very small package is used. Hence, the low-cost, low-power, and light-weight IMMU can be designed and developed. It also enables powering of multiple

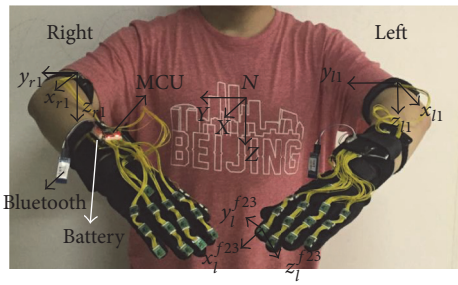


Figure 1: The wearable device.

IMMUs by microcontrol unit (MCU), which reduces the total weight of the system. Moreover, small IMMU can be fastened to the glove, which makes it more convenient and easier to use. The MPU9250 sensor is mounted on a solid PCB with dimensions of 10 × 15 × 2.6 mm and a weight of about 6 g.

Connectivity is another important issue in the design. Different types of connections between sensor units and access points on the body have been described in [28]. Wireless networking approaches for the connections are convenient, but the complexity of wireless networks is increased and the system has to trade off energy consumption for data rates. To avoid this problem, a wired approach is used in [29]. All sensor units are directly connected to a central controlling unit by cables, which lead to a very complex wiring. In the present work, a cascaded wiring approach [30] is used and developed by exploiting the master SPI bus of each IMMU. This approach simplifies wiring without any need for extra components. Since measurements reading from a string of IMMUs, the MCU need not switch to all the IMMUs to fetch the data, which leads to lower power consumption. Meanwhile, the textile cables are used to connect the IMMU to each other and to the MCU for increasing the flexibility. Here the STM32F4 microcontroller is used to develop the MCU.

2.2. Device Design. After determining the above designs, the wearable design of device can be determined. There are thirty-six IMMUs in the device, and a pair of device is separately put on the right hand and left hand. Each side has eighteen IMMUs, which covers all the segments of the arm, palm, and fingers. Each string deploys three IMMUs, and six strings are used. Five of them are used to capture the motions of the five fingers, and the other one is used to capture the motions of the palm, upper arm, and forearm. The battery and MCU are attached to the wrist. The wearable device is shown in Figure 1.

The proposed wearable device is designed based on the low-cost IMMUs, which can capture more information of the motion than traditional sensors. The traditional sensors used in the data glove such as fiber or hall-effect sensors are frail. Nevertheless, the board of inertial and magnetic sensor is an independent unit. It is more compact, more durable, and more robust. Commercial data gloves are too costly for the consumer market, but the proposed data glove in the paper is low-cost (US$ 200). Moreover, the proposed wearable device can not only capture the motion of hand but also capture the motion of arm, and the estimated results of motion are outputting real time.

3. Methods

In this section, the gesture capture algorithms are presented. First, models of the sensors are described, and the calibration method is presented to improve the measurements of the IMMUs. Then the absolute orientations filter based on quaternion-based extended Kalman filter is deduced, and the relative orientations algorithm integrated kinematics of arm-hand are proposed. Finally, the position estimation algorithm is deduced.

3.1. Models of the Sensors. Before analyzing the models of inertial and magnetic sensors, two coordinate frames that are the navigation coordinate frame N and the body frame b need to be set up. The orientations of a rigid body in the space are determined when the axis orientation of a coordinate frame attached to the body frame with respect to the navigation frame is specified. According to the frames, the sensors' models are separately built as follows.

(1) Rate Gyros. Because the MEMS rate gyros do not have enough sensitivity to measure the earth angular velocity, the model can get rid of the earth angular vector. And the output signal of a rate gyro is influenced by noise and bias; that is,

$$\omega_m = \omega + b_g + w_g, \tag{1}$$

where ω_m is measured by the rate gyros, ω is the true value, b_g is the gyro' bias, and w_g is the noise that supposed to be Gaussian with zero-means.

(2) Accelerometers. The measurements of accelerometers in the body frame b can be written as

$$\mathbf{a}_m = \mathbf{C}_n^b \mathbf{M}^a (\mathbf{a} + \mathbf{g}) + \mathbf{b}_a + \mathbf{w}_a, \tag{2}$$

where $\mathbf{a}_m$ is 3×3 matrixes measured by the accelerometers, $\mathbf{C}_n^b$ denotes the Orientation Cosine Matrix representing the rotation from the navigation frame to the body frame, $\mathbf{g}$ and $\mathbf{a} \in \mathbf{R}^3$ are the gravity vector and the inertial acceleration of the body, respectively, expressed in the navigation frame, and $g = 9.81\ \mathrm{m/s}^2$ denotes the gravitational constant; $\mathbf{M}^a$ is 3×3 matrixes that scale the accelerometers outputs; $\mathbf{b}_a$ is the vector of accelerometers' bias; $\mathbf{w}_a \in \mathbf{R}^3$ is the vector of Gaussian with zero-means.

Commonly, the absolute acceleration of the rigid body in the navigation frame is supposed to be weak $a \ll g$ or the rigid body is static. Then, the model of accelerometers can be simplified as

$$\mathbf{a}_m = \mathbf{C}_n^b \mathbf{M}^a \mathbf{g} + \mathbf{b}_a + \mathbf{w}_a. \tag{3}$$

(3) Magnetometers. The ideal magnetic vector expressed in the navigation frame is modeled by the unit vector $\mathbf{H}_h$. The measurements in the body frame b are given by

$$\mathbf{h}_m = \mathbf{C}_n^b \mathbf{M}^h \mathbf{H}_h + \mathbf{b}_h + \mathbf{w}_h, \tag{4}$$

where $\mathbf{M}^h$ is a 3×3 matrix that scales the magnetometers outputs, $\mathbf{b}_h$ denotes the disturbance vector including magnetometers' bias and magnetic effects, and $\mathbf{w}_h \in \mathbf{R}^3$ is the noises that supposed to be Gaussian with zero-means.

According to the above presentations, we can know that the models of accelerometers and magnetometers include errors, which would lead to errors in estimating orientations. Hence, the calibration should be implemented to improve the accuracy of the sensors.

3.2. Calibration. In order to improve the accuracy of the IMMU, calibration is a necessary procedure. The typical calibration method is to assign the inertial sensors to a known angular velocity and linear acceleration. This method normally needs some specific equipment, such as turntable. Here, a novel calibration method is proposed that do not need the specific equipment and is easy to implement. The calibration procedure has two steps. First, the accelerometers and magnetometers are calibrated by the offline procedure. Then the gyro is calibrated by the online procedure.

The unified mathematical models of the calibration of the accelerometers and magnetometers can be used as follows:

$$\begin{aligned} \mathbf{h}_m^b &= \mathbf{M}^h \mathbf{h}^b + \mathbf{b}^h + \mathbf{w}_m, \\ \mathbf{a}_m^b &= \mathbf{M}^a \mathbf{a}^b + \mathbf{b}^a + \mathbf{w}_a, \end{aligned} \tag{5}$$

where triaxial sensor model is written in the vector form and $\mathbf{b}^h$, $\mathbf{b}^a$ are constant offset that shift the outputs of sensors.

The calibration is the process that determines coefficients $\mathbf{b}^h$, $\mathbf{b}^a$, $\mathbf{M}^h$, $\mathbf{M}^a$ to improve the measurements of sensors. When the unit has omnidirectional rotation, the magnitudes of the true magnetic field and gravity field remain constant, and the loci of the true magnetic field measured $\mathbf{h}^b$ and $\mathbf{a}^b$ are spherical. Meanwhile, the measured $\mathbf{h}_m^b$ and $\mathbf{a}_m^b$ are ellipsoids, and they can be expressed as follows [31]:

$$\begin{aligned} \left\|\mathbf{h}^b\right\|^2 &= \left(\mathbf{h}_m^b\right)^T \mathbf{A}^h \mathbf{h}_m^b - 2\left(\mathbf{b}^h\right)^T \mathbf{A}^h \mathbf{h}_m^b + \left(\mathbf{b}^h\right)^T \mathbf{A}^h \mathbf{b}^h \\ &\quad + \widetilde{\mathbf{w}}_m \\ \left\|\mathbf{a}^b\right\|^2 &= \left(\mathbf{a}_m^b\right)^T \mathbf{A}^a \mathbf{a}_m^b - 2\left(\mathbf{b}^a\right)^T \mathbf{A}^a \mathbf{a}_m^b + \left(\mathbf{b}^a\right)^T \mathbf{A}^a \mathbf{b}^a \\ &\quad + \widetilde{\mathbf{w}}_a, \end{aligned} \tag{6}$$

where $\mathbf{A}^h = (\mathbf{G}^h)^T \mathbf{G}^h$, $\mathbf{G}^h = (\mathbf{M}^h)^{-1}$, $\mathbf{A}^a = (\mathbf{G}^a)^T \mathbf{G}^a$, $\mathbf{G}^a = (\mathbf{M}^a)^{-1}$, and $\widetilde{\mathbf{w}}_m$ and $\widetilde{\mathbf{w}}_a$ are assumed noise.

Equations (6) are the expressions of the ellipsoid. In other words, the measurements are constrained to lie on an ellipsoid. Thus, the calibration of the magnetometers and accelerometers is to seek ellipsoid-fitting methods to solve the coefficients of $\mathbf{G}^a$, $\mathbf{G}^h$, $\mathbf{b}^a$, $\mathbf{b}^h$. And the least squares algorithm is commonly used to determine the parameters. The device is rotated adequately to ensure that each unit gets enough measurements to determine the calibration parameters by least squares algorithm. Hence, the accelerometers and magnetometers of all units of the device are entirely calibrated.

After calibration, the magnetometers are calibrated to lessen the environmental magnetic effect, and the biases of the sensor outputs are compensated. Accordingly, the previous equation can be rewritten as follows:

$$\mathbf{h}_m = \mathbf{C}_n^b \mathbf{H}_h + \mathbf{w}_h. \tag{7}$$

Since the scale and bias errors from the accelerometers are calibrated, the models of the accelerometers rewritten as follows:

$$\mathbf{a}_m = \mathbf{C}_n^b \mathbf{g} + \mathbf{w}_a. \tag{8}$$

The offline calibration is implemented to determine the bias and scale of the accelerometers and magnetometers. And the online calibration is implemented to remove the gyro bias. We keep the data glove stationary for a while before use. So the readings are zeros. The bias can then be computed by the mean value of the measurements. The model is expressed as follows:

$$\boldsymbol{\omega}_m = \boldsymbol{\omega} + \mathbf{w}_g, \tag{9}$$

where $\boldsymbol{\omega}_m$ are the measurements of gyros, $\boldsymbol{\omega}$ are the true angular velocity, and $\mathbf{w}_g$ are the noise of gyros.

The calibration parameters of the sensors are compensated into the measurements so that the accuracy can be improved for further work.

3.3. Orientation Filters. Based on these kinds of sensors, two independent ways can determine the attitude and heading. One is determined by open-loop gyros. The angular rate of the rigid body is measured using gyros with respect to its body axis frame. The angles are estimated by the open-loop integration process, which has high dynamic characteristic. However, the gyro errors would cause wandering attitude angles and the gradual instability of the integration drifting. The other way is determined from open-loop accelerometers and magnetometers. The orientations can be correctly obtained from accelerometers and magnetometers in the ideal environment. However the disturbance and noises would lead to large errors and make results lack reliability. The two ways are both quite difficult to achieve acceptable performance. Sensor fusion is a great choice to attain the stable and accurate orientations. Then, the fusion algorithm of quaternion-based extended Kalman filter will be deduced.

The frames are shown in Figure 1. In the figure, the global frame is N and local reference frame is, respectively, located in each IMMU. The global reference frame z-axis is defined along the axial axis (from the head to the feet) of the subject, y-axis along the sagittal (from the left shoulder to the right shoulder) axis, and x-axis along the coronal axis (from the back to the chest). The local frame z-axis is defined along the axial axis (normal to the surface of IMMU along the downward) of the subject, y-axis along the sagittal (from the left side to the right side of the IMMU) axis, and x-axis along the coronal axis (from the back to the forward of IMMU). Meanwhile, two assumptions about data glove in use are made: (1) the body keeps static and only arm and hand are in motion; (2) the local static magnetic field is homogeneous throughout the whole arm.

3.3.1. Absolute Orientation Filter. Combining the measured angular velocity, acceleration, and magnetic field values of one single IMMU, it can stably determine the orientations with respect to a global coordinate system. The global coordinate N and each coordinate b of IMMUs are shown in Figure 1. The transformation between the representations of a 3×1 column-vector $\mathbf{x}$ between N and b is expressed as

$$\mathbf{x}^b = \mathbf{C}_n^b [\mathbf{q}] \mathbf{x}^n, \tag{10}$$

where quaternion $\mathbf{q} = [q_0; \mathbf{Q}]$, $[\mathbf{Q}]$ is the antisymmetric matrix given by

$$[\mathbf{Q}] = \begin{bmatrix} 0 & Q_3 & -Q_2 \\ -Q_3 & 0 & Q_1 \\ Q_2 & -Q_1 & 0 \end{bmatrix}. \tag{11}$$

The attitude matrix $\mathbf{C}$ is related to the quaternion by

$$\mathbf{C}(q) = \left(q_0^2 - \mathbf{Q} \cdot \mathbf{Q}\right) \mathbf{I} + 2\mathbf{Q}\mathbf{Q}^T + 2q_0 [\mathbf{Q}], \tag{12}$$

where $\mathbf{I}$ is the identity matrix.

The state vector is composed of the rotation quaternion. The state transition vector equation is

$$\begin{aligned} \mathbf{x}_{k+1} &= q_{k+1} = \Phi(T_s, \omega_k) + w_k = \exp(\Omega_k T_s) q_k + {}^q w_k, \\ {}^q w_k &= -\frac{T_s}{2} \Gamma_k {}^q \mathbf{v}_k = -\frac{T_s}{2} \begin{bmatrix} [e_k \times] + q_{4k} \mathbf{I} \\ -e_k^T \end{bmatrix} {}^g \mathbf{v}_k, \end{aligned} \tag{13}$$

where the gyro measurement noise vector ${}^g\mathbf{v}_k$ is assumed small enough that a first-order approximation of the noisy transition matrix is possible.

Then the process noise covariance matrix Q_k will have the following expression:

$$Q_k = \left(\frac{T_s}{2}\right)^2 \Gamma_k \Gamma_g \Gamma_k^T. \tag{14}$$

The measurement model is constructed by stacking the accelerometer and magnetometer measurement vectors:

$$\begin{aligned} \mathbf{z}_{k+1} &= \begin{bmatrix} \mathbf{a}_{k+1} \\ \mathbf{m}_{k+1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{C}_n^b(q_{k+1}) & 0 \\ 0 & \mathbf{C}_n^b(q_{k+1}) \end{bmatrix} \begin{bmatrix} \mathbf{g} \\ \mathbf{h} \end{bmatrix} + \begin{bmatrix} {}^a\mathbf{v}_{k+1} \\ {}^m\mathbf{v}_{k+1} \end{bmatrix}. \end{aligned} \tag{15}$$

The covariance matrix of the measurement model $\mathbf{R}_{k+1}$ is

$$\mathbf{R}_{k+1} = \begin{bmatrix} {}^a\mathbf{R}_{k+1} & 0 \\ 0 & {}^m\mathbf{R}_{k+1} \end{bmatrix}, \tag{16}$$

where the accelerometer and magnetometer measurement noise ${}^a\mathbf{v}_{k+1}$ and ${}^m\mathbf{v}_{k+1}$ are uncorrelated zero-mean white noise process and the covariance matrixes of which are ${}^a\mathbf{R}_{k+1} = \sigma_a^2 \mathbf{I}$ and ${}^m\mathbf{R}_{k+1} = \sigma_m^2 \mathbf{I}$, respectively.

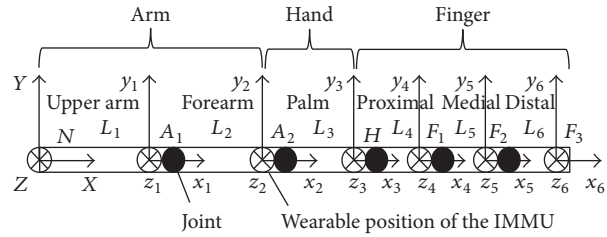


Figure 2: The model and frames of the arm-hand.

Because of the nonlinear nature of (15), the EKF approach requires that a first-order Taylor-Maclaurin expansion is carried out around the current state estimation by computing the Jacobian matrix:

$$\mathbf{H}_{k+1} = \left.\frac{\partial}{\partial \mathbf{x}_{k+1}} \mathbf{z}_{k+1}\right|_{\mathbf{x}_{k+1} = \mathbf{x}_{k+1}^-}. \tag{17}$$

Then, the orientations are estimated by the following EKF equations.

Compute the a priori state estimate:

$$\mathbf{x}_{k+1}^- = \Phi\left(\mathbf{T}_s, \boldsymbol{\omega}_k\right) \mathbf{x}_k. \tag{18}$$

Compute the a priori error covariance matrix:

$$\mathbf{P}_{k+1}^- = \Phi\left(\mathbf{T}_s, \boldsymbol{\omega}_k\right) \mathbf{P}_k \Phi\left(\mathbf{T}_s, \boldsymbol{\omega}_k\right)^T + \mathbf{Q}_k. \tag{19}$$

Compute the Kalman gain:

$$\mathbf{K}_{k+1} = \mathbf{P}_{k+1}^- \mathbf{H}_{k+1}^T \left(\mathbf{H}_{k+1} \mathbf{P}_{k+1}^- \mathbf{H}_{k+1}^T + \mathbf{R}_{k+1}\right)^{-1}. \tag{20}$$

Compute the a posteriori state estimate:

$$\mathbf{x}_{k+1} = \mathbf{x}_{k+1}^- + \mathbf{K}_{k+1} \left[\mathbf{z}_{k+1} - f\left(\mathbf{x}_{k+1}^-\right)\right]. \tag{21}$$

Compute the a posteriori error covariance matrix:

$$\mathbf{P}_{k+1} = \mathbf{P}_{k+1}^- - \mathbf{K}_{k+1} \mathbf{H}_{k+1} \mathbf{P}_{k+1}^-. \tag{22}$$

According to the above algorithm, the absolute orientations of each IMMU can be estimated. Then, the kinematics of the human arm and hand is considered.

3.3.2. Relative Orientation Filter. The kinematic frames of the arm, hand, and the forefinger are presented. There are six joints, and the coordinate frames are built including the global coordinate (N), upper-arm coordinate (A_1), forearm (A_2), palm coordinate (H), proximal coordinate (F_1), medial coordinate (F_2), and distal coordinate (F_3). And the lengths between the joints are L_1, L_2, L_3, L_4, L_5, and L_6. The frames are shown in Figure 2. Then the relative orientations between two consecutive bodies can be determined by the following:

$$\mathbf{q}_{ij}^r = \mathbf{q}_i^{-1} \cdot \mathbf{q}_j, \tag{23}$$

where $\mathbf{q}_{ij}^r$ is the quaternion of the relative orientations, $\mathbf{q}_i$ is the quaternion of the absolute orientations of the

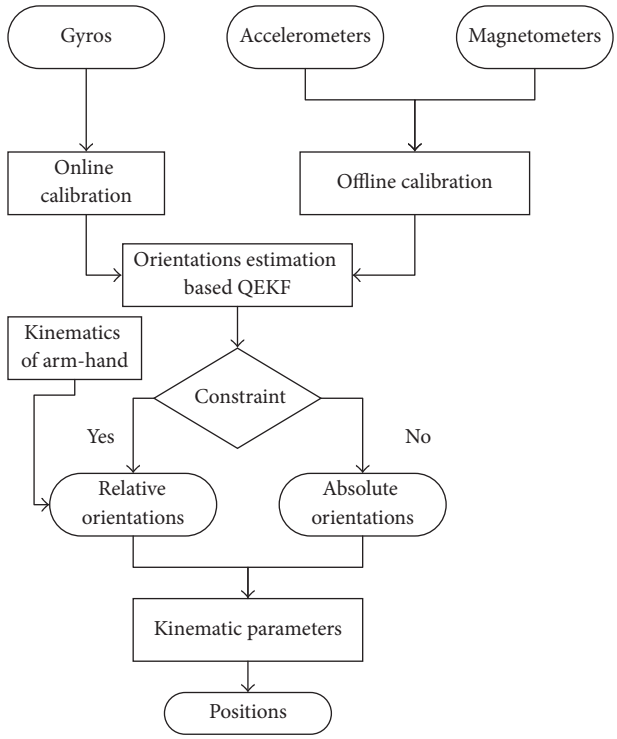


Figure 3: The diagram of the proposed method.

first coordinate, and $\mathbf{q}_j$ is the quaternion of the absolute orientations of the second coordinate.

Meanwhile, the kinematic models of the arm, hand, and fingers are considered to determine the orientations of each segment. Human arm-hand motion can be supposed to be the articulated motion of rigid body parts. These segments are upper arm (between the shoulder and elbow joints), forearm (between the elbow and wrist joints), hand (between the wrist joint and proximal joint), proximal finger (between the proximal joint and medial joint), medial finger (between the medial joint and distal joint), and distal finger (from the distal joint). Every joint has its own local frame. Shoulder can be modeled as a ball joint with three DOFs and a fixed point representing the center of the shoulder. Movements are represented as the vector between the upper arm and body. Elbow is the rotating hinge joint with two DOFs. Wrist is modeled as a rotating hinge joint with two DOFs that is calculated between the vector representing the hand and forearm. Proximal joint is modeled as rotating hinge joint with two DOFs. Medial joint and distal joint are modeled as rotating joint with one DOF. Thus, the kinematic of this model consists of ten DOFs: three in the shoulder joint, two in the elbow joint, two in the wrist joint, two in the proximal joint, one in the medial joint, and one in the distal joint. Hence, the responding constraints are used to determine the orientations of each segment.

3.4. Positions Estimation. We assume the body keeps static and the motion of arm and hand is formed by rotation of the joints. Hence, the position of the fingertip $\mathbf{p}_{F_3}^N$, expressed in the hand coordinate frame (see Figure 2) can be derived using forward kinematics:

$$\begin{aligned}\begin{bmatrix}\mathbf{p}_{F_3}^N\\ 1\end{bmatrix} &= \mathbf{T}^{NA_1}\mathbf{T}^{A_1A_2}\mathbf{T}^{A_2H}\mathbf{T}^{HF_1}\mathbf{T}^{F_1F_2}\mathbf{T}^{F_2F_3}\mathbf{p}_{F_2}^{F_3}\\ &= \mathbf{T}^{NF_3}\begin{bmatrix}\mathbf{p}_{F_2}^{F_3}\\ 1\end{bmatrix},\end{aligned} \tag{24}$$

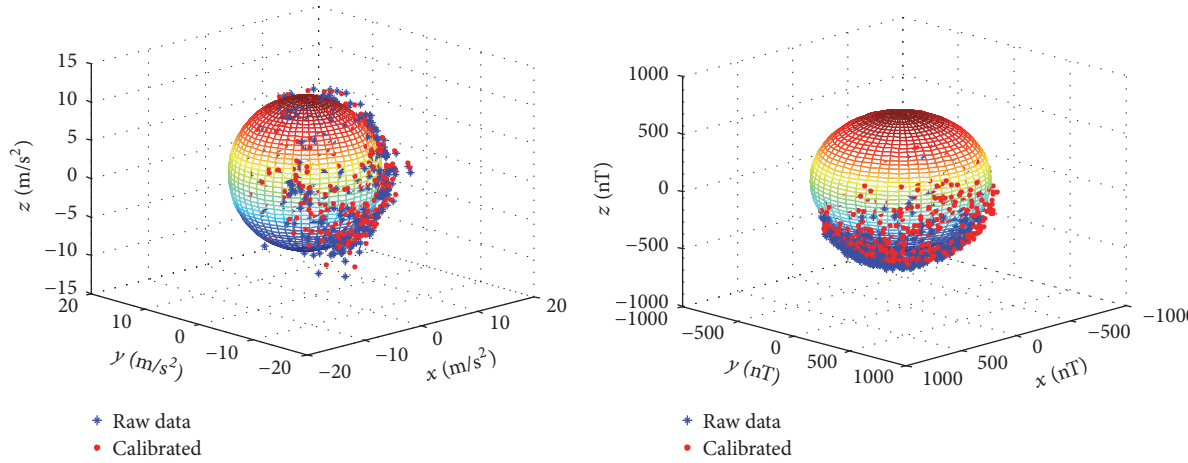
where the transformation between two consecutive bodies is expressed by $\mathbf{T}^{NA_1}$, $\mathbf{T}^{A_1A_2}$, $\mathbf{T}^{A_2H}$, $\mathbf{T}^{HF_1}$, $\mathbf{T}^{F_1F_2}$, and $\mathbf{T}^{F_2F_3}$.

The total transformation $\mathbf{T}^{NF_3}$ is given by the product of each consecutive contribution:

$$\mathbf{T}^{NF_3} = \begin{bmatrix}R\left(q^{NF_3}\right) & \mathbf{p}_{F_3}^N\\ 0_3^T & 1\end{bmatrix}, \tag{25}$$

where $R(q^{NF_3})$ is the orientation of the distal phalanx with respect to the body and $\mathbf{p}_{F_3}^N$ is the position of the distal frame expressed in the global frame.

3.5. Summary. The gestures capture method is proposed in the section. The two procedures of the calibration are firstly implemented to improve the measurements of the sensors. The inertial sensors are calibrated by the offline calibration, and the gyros are calibrated by the online calibration. Then the orientations of the IMMUs are estimated by the QEKF. And the kinematics of arm-hand is considered and the constraints are integrated to determine the relative orientations. Finally, the kinematic parameters are used to estimate the positions. A complete diagram of the method is depicted in Figure 3.



(a) The calibration results of accelerometers

(b) The calibration results of magnetometers

Figure 4: The calibration results.

Table 1: The calibration results.

	Scale	Bias
Accelerometers	$\mathbf{G}^a = \begin{bmatrix} 1 & -0.02 & 0 \\ 0 & 0.95 & 0.12 \\ 0 & 0 & 0.92 \end{bmatrix}$	$\mathbf{b}^a = \begin{bmatrix} 0.56 \\ 0.87 \\ -0.87 \end{bmatrix}$
Magnetometers	$\mathbf{G}^h = \begin{bmatrix} 1.11 & 0.02 & 0 \\ 0 & 1.10 & 0 \\ 0 & 0 & 1.05 \end{bmatrix}$	$\mathbf{b}^h = \begin{bmatrix} 5.71 \\ -39.54 \\ -82.98 \end{bmatrix}$

Table 2: The RMSE of the orientations.

	Static (°)			Dynamic (°)		
	Yaw	Pitch	Roll	Yaw	Pitch	Roll
IMMU-1	0.04	0.01	0.02	2.43	0.05	1.02
IMMU-2	0.11	0.01	0.02	1.68	0.51	1.03
IMMU-3	0.10	0.02	0.02	1.77	0.02	0.07

4. Experiments and Results

The calibration is implemented to improve the sensors accuracy, and then the comparison experiments are testified to investigate the stability and accuracy of the orientations estimation. The real-time gesture motion capture experiments prove the validness of the proposed device.

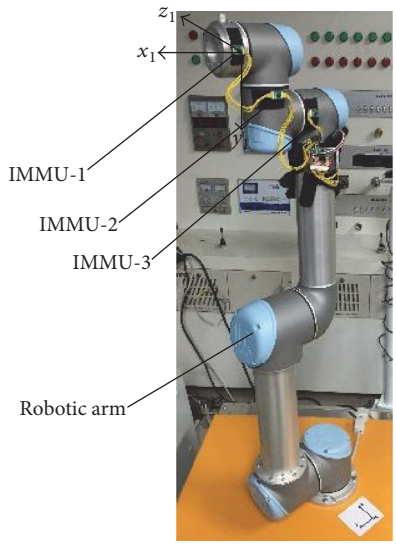
4.1. Calibration Results. The calibration results of the triaxis accelerometers and triaxis magnetometers are presented in the section. The calibration samples of IMMUs are collected by rotating in various orientations. Then, the proposed method was used to determine the correctional calibration parameters as $\mathbf{G}^a$, $\mathbf{G}^h$, $\mathbf{b}^a$, and $\mathbf{b}^h$. The outputs of the accelerometers and magnetometers are shown in Figure 4. The blue sphere in Figure 4 is the raw date of sensors and the red sphere is the calibrated data of sensors. The calibration parameters are listed in Table 1.

4.2. Evaluation Experiments. After calibrating all the sensors of the device, the evaluation experiments are designed to assess the accuracy of the orientation estimation. One of the strings of the device is attached to the robotic arm. Three IMMUs are set the same directions, shown in Figure 5. And the robotic arm is rotated from the first state to the second state. Then the estimated orientations of IMMUs compare with the true orientations of robotic arm. The results of the orientations of three IMMUs are shown in Figure 6. And the results of the root mean square error (RMSE) of the orientations are listed in Table 2.

As shown in Figure 6, during 5~8 s, the robotic arm is dynamic. And, in static case, it can be known that the results have smaller variance and higher precision than the results of the dynamic case. Moreover, the yaw angle has the lager variance compared with the roll and pitch. It should be caused by the motors of the robotic arm, because the magnetometers are disturbed. When the robotic arm is in static, the device can compute the accuracy orientations. The comparison results proved the accuracy of the orientations of the IMMUs of the device. Then the real-time gesture motion capture experiments are implemented.

4.3. Gesture Capture Experiments. The IMMUs' data is sampled, collected and computed by the MCU, and subsequently transmitted via Bluetooth to the external devices. The MCU processes the raw data, estimates the orientations of the each unit, encapsulates them into a packet, and then sends the packet to the PC by Bluetooth. The baud rate for transmitting data is 115200 bps. The frequency is 50 HZ. By using this design, the motion capture can be demonstrated by the virtual model on the PC immediately. The interface is written by C#. The wearable system is shown as Figure 7.

To further verify the effectiveness of proposed device, upper arm is swung up and down. The results of the orientations of the gestures are shown in Figure 8. And the



(a) The first state

(b) The second state

Figure 5: The comparison experiments with robotic arm.

positons of the fingertip of the right forefinger are shown in Figure 9. As shown in the figures, the wearable device can determine the realistic movements of the arm and hand. Meanwhile, the accuracy of the results is assessed by the statistics. The root mean square error (RMSE) of the orientations is listed in Table 3. The RMSE of the positons are listed in Table 4. Furthermore, the wearable device is also tested by ten healthy participants; the real-time motion gestures capture experiments are implemented to prove the validity of the proposed system.

4.4. Discussion. In this paper, a novel wearable device for gestures capturing based on magnetic and inertial measurement units is proposed. As a practically useful application, the proposed device satisfies the requirements including the accuracy, computational efficiency, and robustness. First, the calibration method is designed to improve the accuracy of the measurements of the sensors. In particular, the magnetic effects of the field are attenuated in advance. Then the

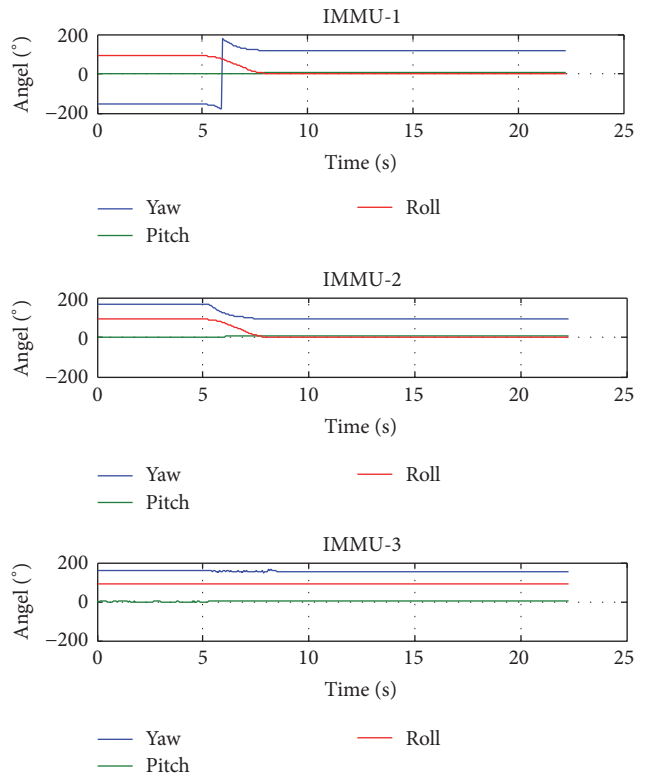


Figure 6: The results of orientations of the IMMUs.

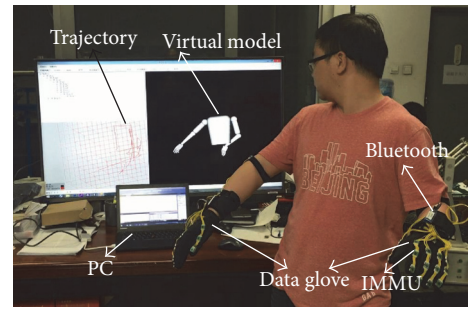


Figure 7: The wearable system.

Table 3: The RMSE of the orientations of the arm-hand.

	Angle (°)		
Upper arm	1.07	1.10	0.44
forearm	0.23	0.24	
Palm	0.19	0.18	
Proximal	0.11	0.26	
Medial		0.10	
Distal		0.08	

Table 4: The RMSE of the positions of the fingertip.

	x (cm)	y (cm)	z (cm)
Positions of the fingertip	3.45	1.35	2.34

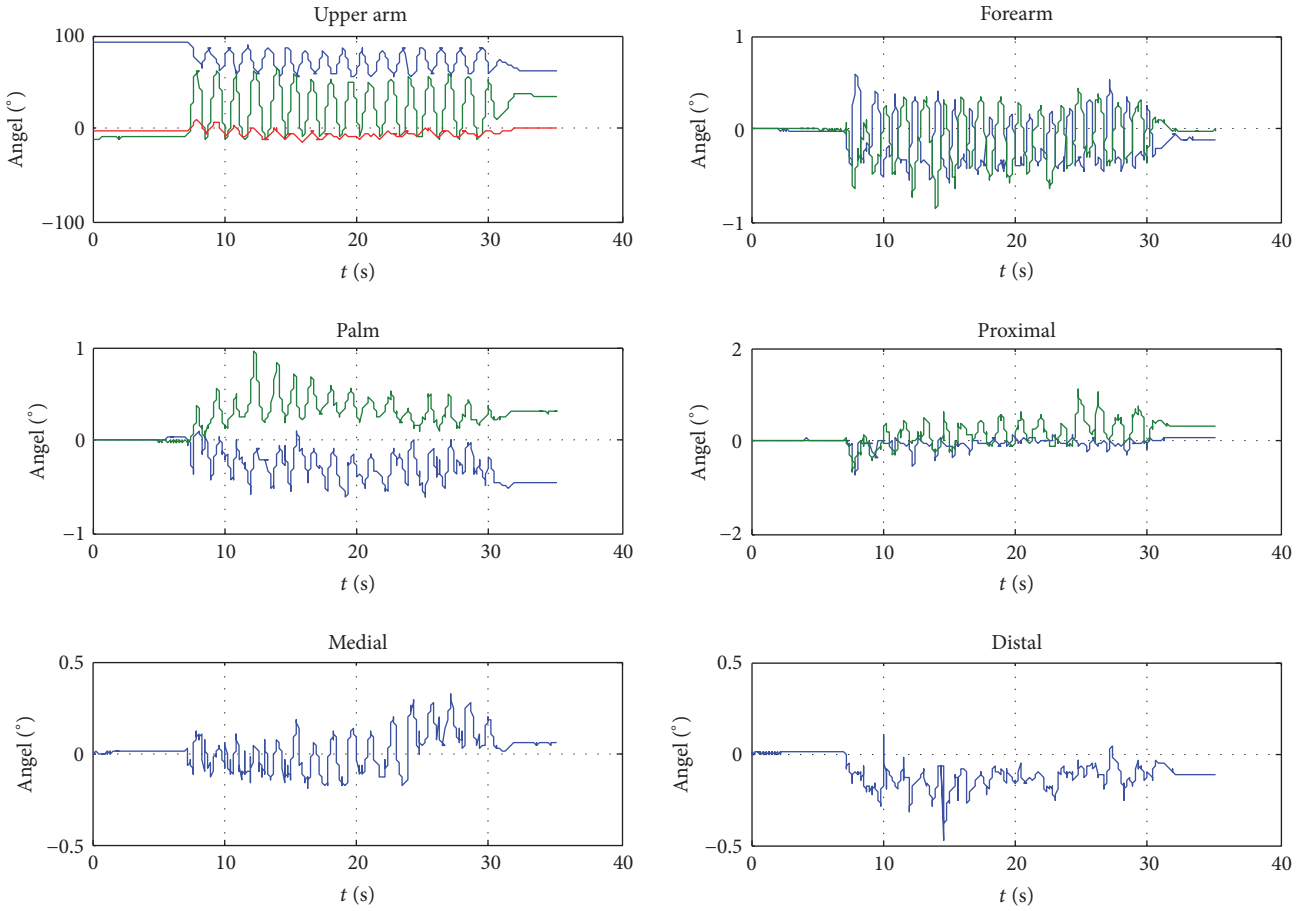


Figure 8: The orientations of gestures.

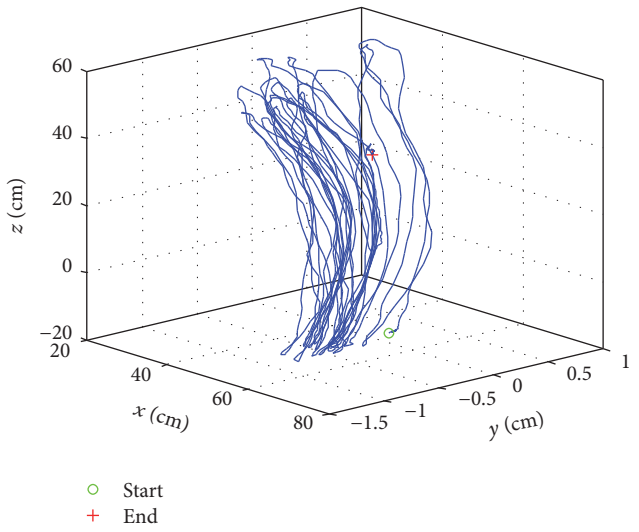


Figure 9: The positions of the fingertip of the right forefinger.

orientations of gesture are divided into absolute orientations and relative orientations. The absolute orientations are determined by the QEKFs, and the relative orientations are estimated by integrating the kinematics of the arm-hand. The positions are then easily computed. The proposed method is simple and fast. The algorithm is operated in the embedded system at 50 Hz. The results of the experiments proved the advantages of the proposed device.

5. Conclusion

This paper presents the design, implementation, and experimental results of a wearable device for gestures capturing using inertial and magnetic sensor modules containing orthogonally mounted triads of accelerometers, angular rate sensors, and magnetometers. Different from commercial motion data gloves which usually use high-cost motion-sensing fibers to acquire hand motion data, we adopted the low-cost inertial and magnetic sensor to reduce cost. Meanwhile, the performance of the low-cost, low-power, and light-weight IMMU is superior to some commercial IMMU. Furthermore, the novel device is proposed based on thirty-six IMMUs, which cover the whole segments of the two arms and hands. We designed the online and offline calibration methods to improve the accuracy of the units. We deduced the 3D arm and hand motion estimation algorithms that integrated the proposed kinematic models of the arm, hand, and fingers and the attitude of gesture and positions of fingertips can be determined. For real-time performance and

convenience, the interface with virtual model is designed. Performance evaluations verified that the proposed data glove can accurately capture the motion of gestures. The system is developed in the way that all electronic components can be integrated and easy to wear. This makes it more convenient and appealing for the user.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work has been supported by the National Science Foundation for Young Scientists of China (Grant no. 61503212) and National Natural Science Foundation of China (Grant nos. U1613212, 61327809, 61210013, and 91420302).

References

[1] B.-S. Lin, I.-J. Lee, P.-C. Hsiao, S.-Y. Yang, and W. Chou, "Data glove embedded with 6-DOF inertial sensors for hand rehabilitation," in *Proceedings of the 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '14)*, pp. 25–28, Kitakyushu, Japan, August 2014.

[2] J. M. Palacios, C. Sagués, E. Montijano, and S. Llorente, "Human-computer interaction based on hand gestures using RGB-D sensors," *Sensors*, vol. 13, no. 9, pp. 11842–11860, 2013.

[3] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2012.

[4] E. A. Arkenbout, J. C. F. de Winter, and P. Breedveld, "Robust hand motion tracking through data fusion of 5dt data glove and nimble VR kinect camera measurements," *Sensors*, vol. 15, no. 12, pp. 31644–31671, 2015.

[5] D. Regazzoni, G. De Vecchi, and C. Rizzi, "RGB cams vs RGB-D sensors: low cost motion capture technologies performances and limitations," *Journal of Manufacturing Systems*, vol. 33, no. 4, pp. 719–728, 2014.

[6] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2695–2704, 2012.

[7] L. Dipietro, A. M. Sabatini, and P. Dario, "A survey of glove-based systems and their applications," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 38, no. 4, pp. 461–482, 2008.

[8] B. Takacs, "How and why affordable virtual reality shapes the future of education," *The International Journal of Virtual Reality*, vol. 7, no. 1, pp. 53–66, 2008.

[9] G. Saggio, "A novel array of flex sensors for a goniometric glove," *Sensors and Actuators, A: Physical*, vol. 205, pp. 119–125, 2014.

[10] R. Gentner and J. Classen, "Development and evaluation of a low-cost sensor glove for assessment of human finger movements in neurophysiological settings," *Journal of Neuroscience Methods*, vol. 178, no. 1, pp. 138–147, 2009.

[11] X. L. Wang and C. L. Yang, "Constructing gyro-free inertial measurement unit from dual accelerometers for gesture detection," *Sensors & Transducers*, vol. 171, no. 5, pp. 134–140, 2014.

[12] W. Chou, B. Fang, L. Ding, X. Ma, and X. Guo, "Two-step optimal filter design for the low-cost attitude and heading reference systems," *IET Science, Measurement and Technology*, vol. 7, no. 4, pp. 240–248, 2013.

[13] D. Roetenberg, H. J. Luinge, C. T. M. Baten, and P. H. Veltink, "Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 3, pp. 395–405, 2005.

[14] J. M. Lambrecht and R. F. Kirsch, "Miniature low-power inertial sensors: promising technology for implantable motion capture systems," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 6, pp. 1138–1147, 2014.

[15] R. Xu, S. L. Zhou, and W. J. Li, "MEMS accelerometer based nonspecific-user hand gesture recognition," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1166–1173, 2012.

[16] E. Morganti, L. Angelini, A. Adami, D. Lalanne, L. Lorenzelli, and E. Mugellini, "A smart watch with embedded sensors to recognize objects, grasps and forearm gestures," *Procedia Engineering*, vol. 41, pp. 1169–1175, 2012.

[17] J.-H. Kim, N. D. Thang, and T.-S. Kim, "3-D hand motion tracking and gesture recognition using a data glove," in *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE '09)*, pp. 1013–1018, IEEE, Seoul, South Korea, July 2009.

[18] B.-S. Lin, I.-J. Lee, P.-C. Hsiao, S.-Y. Yang, and W. Chou, "Data glove embedded with 6-DOF inertial sensors for hand rehabilitation," in *Proceedings of the 10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '14)*, vol. 10, pp. 25–28, Kitakyushu, Japan, August 2014.

[19] F. Cavallo, D. Esposito, E. Rovini et al., "Preliminary evaluation of SensHand V1 in assessing motor skills performance in Parkinson disease," in *Proceedings of the 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR '13)*, pp. 1–6, Seattle, Wash, USA, June 2013.

[20] H. G. Kortier, V. I. Sluiter, D. Roetenberg, and P. H. Veltink, "Assessment of hand kinematics using inertial and magnetic sensors," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, article 70, 2014.

[21] Y. Xu, Y. Wang, Y. Su, and X. Zhu, "Research on the calibration method of micro inertial measurement unit for engineering application," *Journal of Sensors*, vol. 2016, Article ID 9108197, 11 pages, 2016.

[22] J. Keighobadi, "Fuzzy calibration of a magnetic compass for vehicular applications," *Mechanical Systems and Signal Processing*, vol. 25, no. 6, pp. 1973–1987, 2011.

[23] X. Yun and E. R. Bachmann, "Design, implementation, and experimental results of a quaternion-based Kalman filter for human body motion tracking," *IEEE Transactions on Robotics*, vol. 22, no. 6, pp. 1216–1227, 2006.

[24] Xsens Technologies, http://www.xsens.com/.

[25] T. Hamel and R. Mahony, "Attitude estimation on SO(3) based on direct inertial measurements," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '06)*, pp. 2170–2175, Orlando, Fla, USA, May 2006.

[26] S. Bonnabel, P. Martin, and P. Rouchon, "Non-linear symmetry-preserving observers on Lie groups," *IEEE Transactions on Automatic Control*, vol. 54, no. 7, pp. 1709–1713, 2009.

[27] InvenSense, MPU-9250 Nine-Axis MEMS MotionTracking Device, 2015, http://www.invensense.com/-products/motion-tracking/9-axis/mpu-9250/.

[28] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. M. Leung, "Body area networks: a survey," *Mobile Networks and Applications*, vol. 16, no. 2, pp. 171–193, 2011.

[29] MIT Media Lab, *MIThril Hardware Platform*, 2015, http://www.media.mit.edu/wearables/mithril/hardware/index.html.

[30] S. Salehi, G. Bleser, N. Schmitz, and D. Stricker, "A low-cost and light-weight motion tracking suit," in *Proceedings of the IEEE 10th International Conference on Ubiquitous Intelligence and Computing and IEEE 10th International Conference on Autonomic and Trusted Computing (UIC/ATC '13)*, December 2013.

[31] J. Fang, H. Sun, J. Cao, X. Zhang, and Y. Tao, "A novel calibration method of magnetic compass based on ellipsoid fitting," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 6, pp. 2053–2061, 2011.

Hindawi Publishing Corporation
Scientific Programming
Volume 2017, Article ID 7136702, 10 pages
http://dx.doi.org/10.1155/2017/7136702

Research Article

Underwater Matching Correction Navigation Based on Geometric Features Using Sonar Point Cloud Data

Mingjie Dong,[1] Wusheng Chou,[1] and Bin Fang[2]

[1]*Robotics Institute, Beijing University of Aeronautics and Astronautics, XueYuan Road No. 37, HaiDian District, Beijing 100191, China*
[2]*State Key Lab of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, HaiDian District, Beijing 100083, China*

Correspondence should be addressed to Mingjie Dong; dongmj@buaa.edu.cn

Received 11 September 2016; Accepted 1 December 2016; Published 4 January 2017

Academic Editor: Wenbing Zhao

In order to localize the Remotely Operated Vehicle (ROV) accurately in the reactor pool of the nuclear power plant, an underwater matching correction navigation algorithm based on geometric features using sonar point cloud data is proposed. At first, an Extended Kalman Filter (EKF) is used to compensate the motion induced distortion after the preprocessing of the sonar point cloud data. Then, the adjacent scanning point cloud data are fitted to be four different straight lines using Hough Transform and least square method. After that, the adjacent straight line is modified based on geometric features to get a standard rectangle. Since the working environment of the ROV is a rectangular shape with all dimensions known, it is used as a priori map. The matching rectangle is then used to compare with the a priori map to calculate the accurate position and orientation of the ROV. The obtained result is then applied as the measurement for the second EKF to obtain better localization accuracy. Experiments have been conducted in man-made water tank which is similar to the reactor pool of the nuclear power plant, and the results successfully verify the effectiveness of the proposed algorithm.

1. Introduction

Nuclear energy has become an important energy source worldwide and the uncertainty of human factors must be a minimum to ensure the safety. That demands high precision robots to substitute human to detect the nuclear power plant and complete the corresponding operation which was done by human workers before. Many robots for nuclear environment have been developed during the last decades. ROV is just one of them. However, ROV needs to have higher reliability as it is used in the reactor pool of the nuclear power plants and often works for more than two hours. So, it is of vital importance to have the ability to get absolute navigation with higher accuracy, especially when it works for a longer time in the nuclear environment.

However, navigation of underwater environment is challenging because GPS is not available due to the rapid attenuation of electromagnetic waves in water [1]. Active acoustic positioning systems such as long baseline (LBL), short baseline (SBL), and ultra-short baseline (USBL) are good alternatives to GPS for correcting dead reckoning errors [2]. The disadvantage of such a technique is that the deployment, calibration, recovery of the transponders, and the relatively small area of coverage consumes costly ship time and complicates the operations [3, 4]. Many researchers have used vision sensors to carry out underwater visual simultaneous localization and mapping (SLAM). The movement of underwater vehicle can be estimated from the displacement of features in the images grabbed by vision sensors and the registered images are combined to produce photomosaic of the traveled area at the same time [5, 6]. The disadvantage is that vision is limited to a few meters under the water and can be easily disturbed by turbulence, floating sediment, or lighting conditions. On the other hand, the multimodality fusion has attracted much attention from the academics and industry [7–10].

In recent years, sonar has been a very popular tool for underwater navigation. The acoustic sonar frequencies can penetrate further the water column and are not prone to turbidity; thus sonar can provide information even in bad

visibility conditions [11]. Many underwater SLAM navigation methods with respect to sonar have been proposed during the last decades. Paper [12] proposed a method that is based on the probabilistic iterative correspondence (pIC) algorithm but takes into account the distortions in the acoustic image to deal with data gathered by an underwater vehicle utilizing mechanically scanned imaging sonar (MSIS). The underwater sonar probabilistic iterative correspondence (uspIC) is proposed in [13] to deal with the significant uncertainty in the measurements or large scan time through adopting probabilistic scan matching strategy and defining a method to strongly alleviate the motion induced distortion. The result shows some improvements in the pose estimate. Meanwhile, the modified-FastSLAM algorithm is proposed and used in the navigation for an open-frame AUV research platform [14]. A novel localization algorithm for an AUV equipped with MSIS is proposed which incrementally constructs a pose graph and conducts graph optimization to correct the robot poses and, especially, the data association algorithm based on Mahalanobis distance and shape matching is deployed to determine loop closures, leading to associated scan pairs used for calculating constraints of the pose graph. The experimental results show that the algorithm outperforms traditional algorithms such as dead reckoning and uspIC in terms of both localization and mapping accuracy [15].

In this paper, we propose the underwater matching correction navigation based on geometric features using sonar point cloud data. The designed and manufactured ROV prototype is shown in Figure 1. Just as described in our previous paper [16], the ROV is radiation proof under the certain dose rate, and the basic components contain control cabinet, buoyancy module, propeller, camera, manipulator, sonar, and so on. The main body frame is made of aluminum alloys to provide sufficient strength and resist the acidic corrosion of the reactor pool of the nuclear power plant. The control cabinet which is made of thick stainless steel is installed inside the main body frame and the double seal processing had been done to protect the control system inside from water leakage. The ROV is designed neutral buoyancy and can move in the direction of surge, sway, heave, and yaw. Besides, the ROV is equipped with a variety of sensors, such as sonar, depth gauge, three-axis accelerometer, three-axis gyroscope, and three-axis magnetometer, which are used to get the position and attitude of the ROV. A fuzzy PID controller is used to realize the depth control of the ROV, and field experiment shows that the ROV can suspend at any user-specified depth under the water, which means that the ROV can work efficiently and stably in the nuclear reactor pool. In the latter paper, the ROV is assumed to be operating solely in a planar field based on the depth control.

The paper is organized as follows. Section 2 demonstrates the introduction of mechanical scanning sonar. The motion compensation based on EKF is presented in Section 3. Section 4 is a fine description of the matching correction navigation based on geometric features. Field experiment is displayed and discussed in Section 5, while conclusions are drawn and discussed in Section 6.

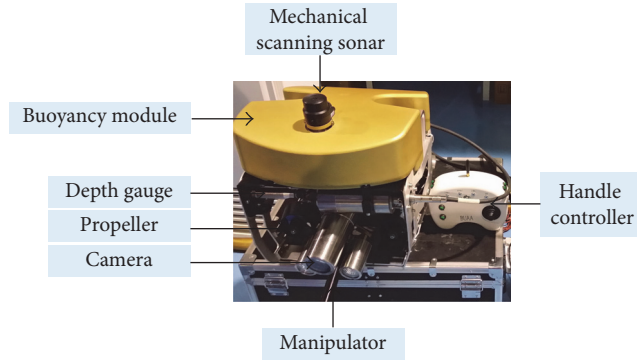


Figure 1: Prototype of the ROV.

2. Instruction of Mechanical Scanning Sonar

The Tritech Micron DST Sonar used in our ROV is a small compact mechanical scanning sonar with digital Compressed High Intensity Radar Pulse (CHIRP) system designed for underwater applications, such as obstacle avoidance and target recognition for both AUVs and ROVs. This sonar can be programmed to cover variable length sectors from a few degrees to full 360° scans. A characteristic fan-shaped beam with vertical aperture angle of 35° and narrow horizontal aperture of 3° allows a sonar image to be formed with enough information about the surrounding environment to recognize sizes, shapes, and surface reflecting characteristics of a target at distances of up to 75 meters. The sensor is mounted on the upper front part of the ROV to provide a clear view and avoid occlusions in the resulting data. Its capacity to sense the environment in which the vehicle is operated makes the DST sonar one of the most important sensors aboard the ROV.

The sonar performs scans in a horizontal 2D plane by rotating a mechanically actuated transducer head at preset angular increments [17]. For each emitted beam, an echo intensity profile is returned from the environment and discretized into a set of bins (distance versus echo-amplitude values) [18]. Since the mechanical scanning sonar needs a considerable period of time to obtain a complete scan, the ROV's motion induces a distortion in the acoustic image when ROV moves. To deal with this, it is necessary to know the ROV's pose at the beam reception time [19]. The EKF algorithm is used to estimate the position as well as its uncertainty while the sonar performs the scan and to correct the distortions induced by the motion of ROV, which will be described in the next section and the expected correction result is as in Figure 2.

However, in practical application, every ping received will contain noise and interference except for useful information, especially within the limited space like water tank or swimming pool. As shown in Figure 2, only the innermost rectangle is the boundary of the water tank. When the acoustic pulse encounters the wall, the propagation of the acoustic pulse is blocked and part of the mechanical energy is reflected back in the opposite direction depending on the nature of the obstacle. Likewise, as the reflected pulse

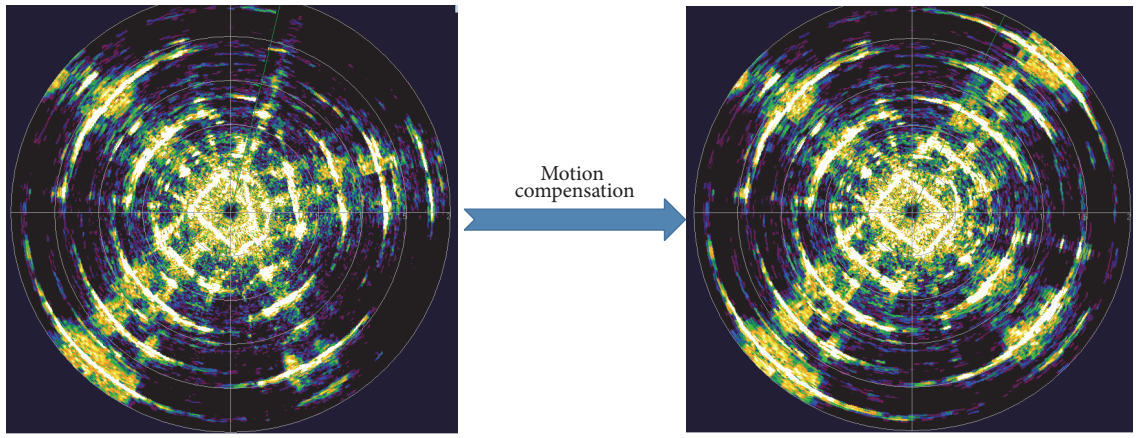


Figure 2: The distortion induced by movement of the ROV can be corrected through motion compensation.

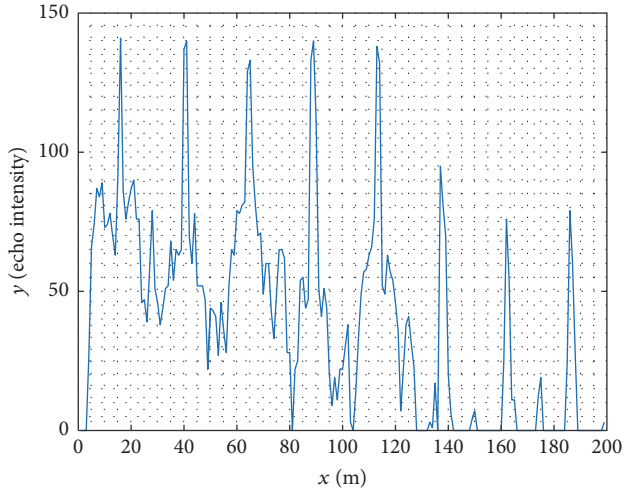


Figure 3: Sonar signal of a ping.

moves across the environment and finds other objects, part of its energy is also returned, ricocheting again with the wall and returning to the sonar head where it is interpreted as if the reflection has never taken place. In other words, the wall acts as a mirror for the acoustic pulse and, as a result, phantoms and reflections not corresponding with real objects can appear [17]. The sonar signal of one ping is as in Figure 3, where y-axis represents the echo intensity of the sonar signal, and has no unit, while x-axis represents the resulting target distance deduced by raw sonar data, from which we can see that the peak echo intensity after the first peak is the phantoms and reflections and must be excluded during the preprocessing.

To deal with this phenomenon, we need to preprocess the sonar signal before motion compensation using EKF, including threshold denoising and the sampling distance limit, as shown in our previous paper [20]. The effect of the preprocessing on the distortion in Figure 2 is as in Figure 4. Only when the sonar signal is being preprocessed and motion compensated through EKF, the matching correction navigation based on geometric features can be conducted.

3. Motion Compensation Based on EKF

The way to compensate motion induced distortion depends on the EKF algorithm, which estimates the state vector containing the position and velocity information of the vehicle. In this system, three-axis accelerometer, three-axis gyroscope, and three-axis magnetometer are used as inertial measurement unit (IMU) to estimate the position and attitude of the ROV. For the reason that the ROV can suspend at any depth under the water based the depth control algorithm and can move stably, as shown in our previous work [16], we assume that it moves only in the planar field with roll and pitch negligible. In other words, the depth value z is invariable, and roll and pitch are zero, respectively.

The different reference frames involved in the system are shown in Figure 5, where $\{E\}$, $\{N\}$, and $\{B\}$ inside the rectangle represent earth fixed reference frame, the base reference frame (the orientation of the experimental water tank), and the body reference frame, respectively.

3.1. Nonlinear Process Model. The state vector of the ROV contains the information of the position and velocity at time k.

$$\mathbf{x}_k = \begin{bmatrix} x & y & \phi & u & v & r \end{bmatrix}^T, \tag{1}$$

where the vector $[x \quad y \quad \phi]$ represents the position and orientation of the ROV in the base reference frame $\{N\}$, while $[u \quad v \quad r]$ represents the corresponding linear and angular velocities in the body reference frame $\{B\}$. The initial value of the state vector $\mathbf{x}_0$ and its covariance matrix $P(0)$ should be estimated before starting the EKF. For the reason that the

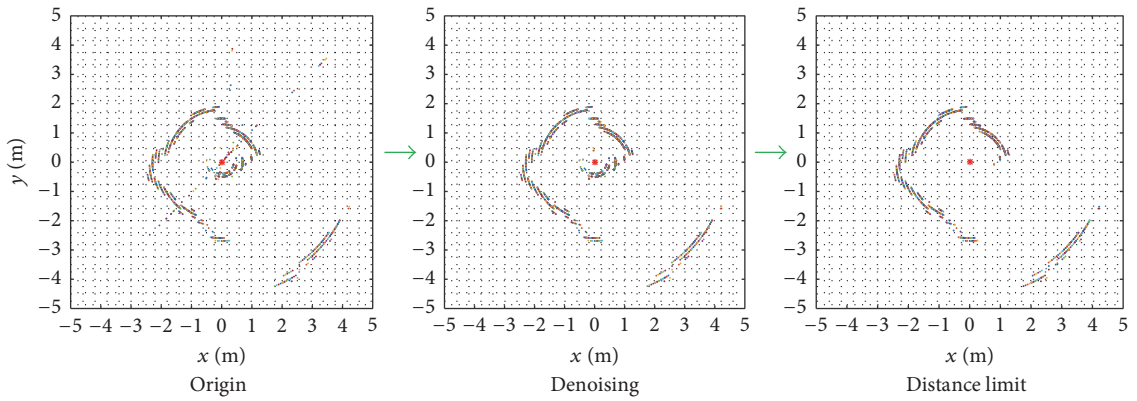


Figure 4: Preprocessing of sonar signal before motion compensation.

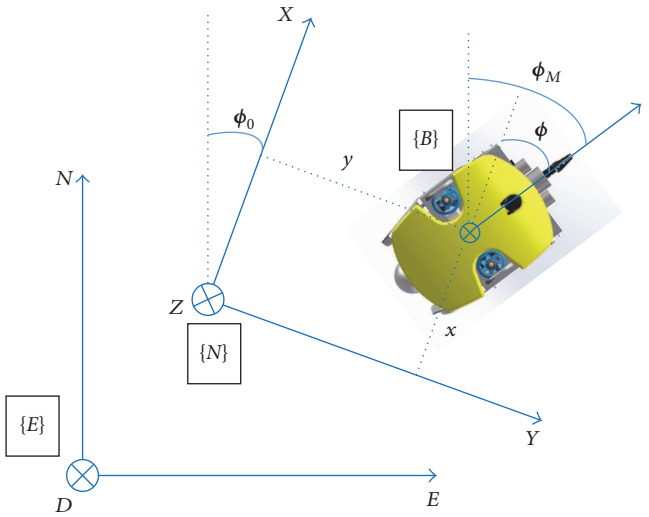


Figure 5: Different reference frame involved in the system.

ROV is set to known location with velocities set to zero at the beginning of the experiment, the state vector at time 0 is

$$\hat{\mathbf{x}}_0 = \begin{bmatrix} x_0 \\ y_0 \\ \phi \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad P(0) = \begin{bmatrix} \sigma_{x_0}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{y_0}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\phi}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

Here, a simple constant velocity kinematics model is introduced to predict the state.

$$\mathbf{x}_k = f(\mathbf{x}_{k-1}, \mathbf{a}_{k-1}),$$

$$\begin{bmatrix} x \\ y \\ \phi \\ u \\ v \\ r \end{bmatrix}_k = \begin{bmatrix} x + \left(ut + \frac{1}{2}a_u t^2\right)\cos(\phi) - \left(vt + \frac{1}{2}a_v t^2\right)\sin(\phi) \\ y + \left(ut + \frac{1}{2}a_u t^2\right)\sin(\phi) + \left(vt + \frac{1}{2}a_v t^2\right)\cos(\phi) \\ \phi + rt + \frac{1}{2}a_r t^2 \\ u + a_u t \\ v + a_v t \\ r + a_r t \end{bmatrix}, \tag{3}$$

where $\mathbf{a} = [a_u \ a_v \ a_r]^T$ is the white Gaussian acceleration noises with zero mean. The covariance of $\mathbf{a}$ is represented by the noise matrix $\mathbf{A}$.

$$E(\mathbf{a}_k) = 0,$$

$$E\left(\mathbf{a}_k \mathbf{a}_j^T\right) = \delta_{kj}\mathbf{A}, \tag{4}$$

$$\mathbf{A} = \begin{bmatrix} \sigma_{a_u}^2 & 0 & 0 \\ 0 & \sigma_{a_v}^2 & 0 \\ 0 & 0 & \sigma_{a_r}^2 \end{bmatrix}.$$

Prediction. The estimate of the state $\mathbf{x}$ is obtained as

$$\hat{\mathbf{x}}_k = f(\hat{\mathbf{x}}_{k-1}). \tag{5}$$

The covariance matrix is

$$\mathbf{P}_k = \mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^T + \mathbf{G}_k \mathbf{A}_k \mathbf{G}_k^T, \tag{6}$$

where $\mathbf{F}_k$ and $\mathbf{G}_k$ are the Jacobian matrices of partial derivatives of the function $\hat{x}_k = f(\hat{x}_{k-1})$ with respect to the state $\mathbf{x}_k$ and the noise $\mathbf{a}$, respectively.

$$\mathbf{F}(k) = \frac{\partial f}{\partial \mathbf{x}}(\hat{\mathbf{x}}_k, 0) = \begin{bmatrix} 1 & 0 & -\hat{u}t\sin(\hat{\phi}) - \hat{v}t\cos(\hat{\phi}) & t\cos(\hat{\phi}) & -t\sin(\hat{\phi}) & 0 \\ 0 & 1 & \hat{u}t\cos(\hat{\phi}) - vt\sin(\hat{\phi}) & t\sin(\hat{\phi}) & t\cos(\hat{\phi}) & 0 \\ 0 & 0 & 1 & 0 & 0 & t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad (7)$$

$$\mathbf{G}(k) = \frac{\partial f}{\partial \mathbf{a}}(\hat{\mathbf{x}}_k, 0) = \begin{bmatrix} \frac{1}{2}t^2\cos(\hat{\phi}) & -\frac{1}{2}t^2\sin(\hat{\phi}) & 0 \\ \frac{1}{2}t^2\sin(\hat{\phi}) & \frac{1}{2}t^2\cos(\hat{\phi}) & 0 \\ 0 & 0 & \frac{1}{2}t^2 \\ t & 0 & 0 \\ 0 & t & 0 \\ 0 & 0 & t \end{bmatrix}.$$

3.2. Nonlinear Measurement Model. The nonlinear measurement model is represented as

$$\mathbf{z}_k = Hx_{k,k-1} + \mathbf{m}_k, \quad (8)$$

where $\mathbf{z}$ is the measurement vector and $\mathbf{m}$ is the white Gaussian noises with zero mean with its covariance matrix $\mathbf{R}$.

$$E(\mathbf{m}_k) = 0, \quad E(\mathbf{m}_k\mathbf{m}_j^T) = \delta_{kj}\mathbf{R}_k. \quad (9)$$

Update Using Three-Axis Accelerometer. The model prediction is updated by the EKF equations each time a new measurement from the three-axis accelerometer arrives and finishes the first and second integrals.

$$\mathbf{z}_{\text{acc},k} = [x, y, u, v]^T = H_{\text{acc}}x_{k,k-1} + \mathbf{m}_k, \quad H_{\text{acc}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad (10)$$

where

$$E(\mathbf{m}_k) = 0, \quad E(\mathbf{m}_k\mathbf{m}_j^T) = \delta_{kj}\mathbf{R}_{\text{acc},k}. \quad (11)$$

Update Using Three-Axis Magnetometer. The model prediction is updated by the EKF equations each time a new measurement from the three-axis magnetometer arrives.

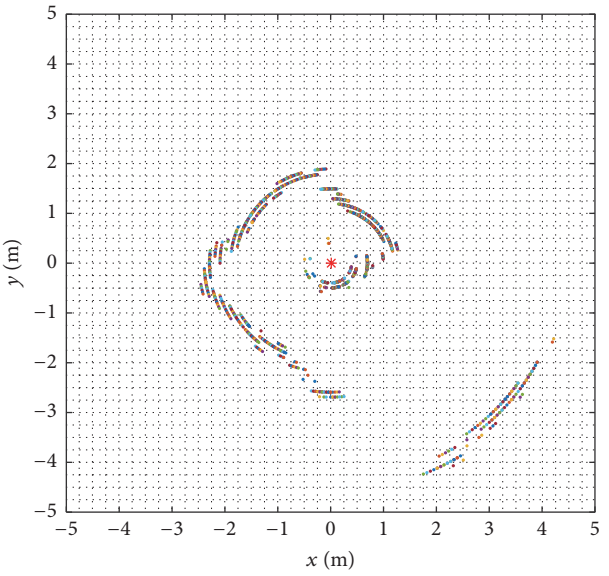


Figure 6: Sonar image after motion compensation.

What calls for special attention is that the angle from the magnetometer is ϕ_M; the yaw angle in the state vector is $\phi = \phi_M - \phi_0$.

$$\mathbf{z}_{\text{mag},k} = [\phi]^T = H_{\text{mag}}x_{k,k-1} + \mathbf{m}_k, \quad H_{\text{mag}} = [0\ 0\ 1\ 0\ 0\ 0], \quad (12)$$

where

$$E(\mathbf{m}_k) = 0, \quad E(\mathbf{m}_k\mathbf{m}_j^T) = \delta_{kj}\mathbf{R}_{\text{mag},k}. \quad (13)$$

Update Using Three-Axis Gyroscope. The model prediction is updated by the EKF equations each time a new measurement from the three-axis gyroscope arrives and finishes the first-order integrals.

$$\mathbf{z}_{\text{gyr},k} = [r]^T = H_{\text{gyr}}x_{k,k-1} + \mathbf{m}_k, \quad H_{\text{gyr}} = [0\ 0\ 0\ 0\ 0\ 1], \quad (14)$$

where

$$E(\mathbf{m}_k) = 0, \quad E(\mathbf{m}_k\mathbf{m}_j^T) = \delta_{kj}\mathbf{R}_{\text{gyr},k}. \quad (15)$$

After finishing the EKF based motion compensation, the sonar image we see is coherent and smooth; one of the experiments from the water tank (length × width × depth = 6 m × 3 m × 1.5 m) is as in Figure 6.

However, in spite of the effectiveness of the motion compensation, the boundary of the water tank is not straight enough for us to localize the ROV. To deal with this problem, we proposed the matching correction navigation algorithm based on geometric features which will be introduced in the next section in detail.

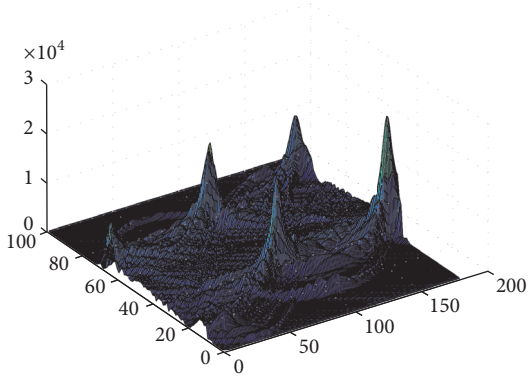


Figure 7: Four corners found using Hough Transform.

4. Matching Correction Navigation Algorithm Based on Geometric Features

The bins we obtain from the mechanical scanning sonar are a set of fixed-bearing temporal signals with different intensity. Every intensity point can be represented (ρ, θ) in polar coordinate, where ρ and θ represent range and bearing of the acoustic ping. So it is necessary to perform coordinate transformation for the measurement data. Assume that the position of point p in the scan plane is $X_p = [x_p, y_p]^T$. Then we have

$$X_p = \left[x_p, y_p\right]^T = \begin{bmatrix} \rho_p \times \cos\left(\dfrac{\theta_p \times \pi}{180}\right) \\ \rho_p \times \sin\left(\dfrac{\theta_p \times \pi}{180}\right) \end{bmatrix}. \quad (16)$$

That is what we see in Figure 6 with sonar point cloud data displayed in Cartesian coordinate, in which the origin is the center of the sonar.

4.1. Corner and Boundary Line Detection. To process the match correction navigation algorithm, we need to eliminate the influence of the noise and interference and find the four intersection points of the four water tank boundaries. The result of obtaining the four corners using Hough Transform [21] is as in Figure 7.

Then, the data sets between two adjacent corners of the four corners are used for line fitting using least square method. Let us assume that each of the four lines a, b, c, and d is

$$y = k_i x + b_i \quad (i = 1, 2, 3, 4). \quad (17)$$

With all the data between two adjacent corners substituted into (17), we can get

$$\arg\min_{k,b} \left(y_j - k_i x_j - b_i\right)^2 \quad (i = 1, 2, 3, 4). \quad (18)$$

After solving the four equations, the four straight lines we get in sonar image are as in Figure 8.

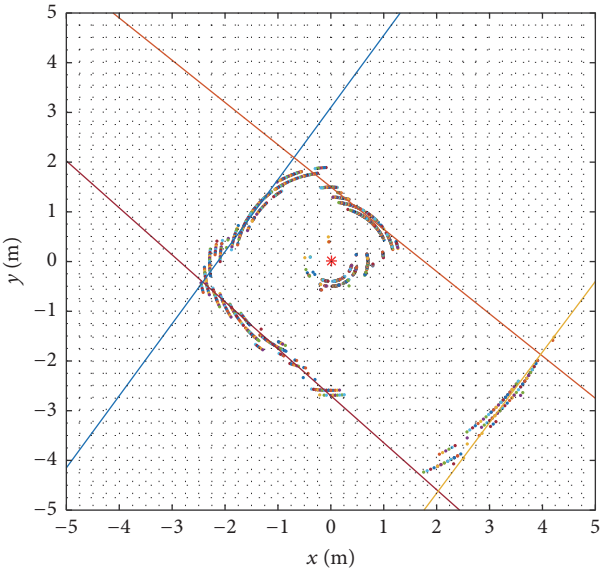


Figure 8: Sonar image after line fitting.

4.2. Rectangle Modification Using Geometric Features. However, as shown in Figure 8, the geometric figure surrounded by the four lines is not a standard rectangle. So the adjacent straight lines are modified based on geometric features to get a standard rectangle. The line with most data as we define line a is set to be the base line, which is calculated using least square method to fit the line at first, then the two adjacent lines are modified with right angle modification relative to the base line to get the slope of the line, respectively, while the fourth line is modified to be parallel to the base line. When all the other three lines' slope is known, we can get these three lines fitted using least squared method again. When the final rectangle is built, it will be compared with the a priori map to calculate the location and orientation of the ROV in the water tank with its uncertainty, respectively. The whole matching correction algorithm based on geometric features is shown in Algorithm 1 in detail.

The result after geometric modification is as in Figure 9. The read "$*$" represents the position of the sonar. The boundary of the water tank is the base reference frame as shown in Figure 5.

4.3. Key Edges Identification Based on Corner Matching. After obtaining the standard rectangle, we can calculate the position of the sonar (x, y) through measuring the distance between sonar and the straight line in the base reference frame; then the position of the ROV can be deduced according to the relative position between the sonar and the ROV's center of gravity. Meanwhile, we can also deduce the orientation of the ROV by comparing the generated rectangle with the a priori map using the slope of the line. However, it is hard for us to recognize the two edges of the water tank which is used as the base reference frame at the beginning. For this reason, we use the lines b and a as the directions x and y in the base reference frame. Every time the ROV moves and the sonar finishes a complete scan, we use Hough Transform to

Polar2Cartesian(x, y);
/* *to get the four corners of the water tank using Hough Transform** /
HoughTransform (all points);
/* *Least square method to fit lines of the most data with adjacent corners** /
$\arg\min_{k_1,b_1}(y_j - k_1 x_j - b_1)^2 \quad (j = 1, 2, \ldots, n)$
/**set the line with most data is set to the baseline* $y = k_1 x + b_1$, *the adjacent two lines to be* $y = k_2 x + b_2$ *and* $y = k_3 x + b_3$, *the fourth line to be* $y = k_4 x + b_4$* /
if $k_i = 0$ $(i = 1, 2, 3, 4)$
  then one of the lines is perpendicular to x axis
  /* *the rectangle can be easily deduced** /
else
  while $k_1 \times k_2 \neq -1$ && $k_1 \times k_3 \neq -1$, **do**
  /* b_i *and* k_1 *remains unchanged, modify* k_i* /
  $k_2 = \dfrac{-1}{k_1}$;
  $k_3 = \dfrac{-1}{k_1}$;
  end while
$k_4 = k_1$;
$\arg\min_{k_i,b_i}(y_j - k_i x_j - b_i)^2 \quad (i = 2, 3, 4)$
end if

Algorithm 1: The map matching correction based on geometric features.

find the four corners, and the adjacent corner between two scans is identified as the same corner, so the lines b and a can be identified, and then we calculate the position and the orientation of the ROV.

4.4. Position and Orientation Correction Using the Second EKF. In order to obtain the better localization accuracy of the ROV, the position and orientation deduced from the matching correction algorithm are then used as new measurement to proceed the second EKF. The whole process of the entire recursive process is as in Figure 10.

5. Field Experiment

To verify the effectiveness of the proposed navigation algorithm, many experiments have been carried out in man-made water tank which is similar to the reactor pool of the nuclear power plant. The water tank is 6 m × 3 m × 1.5 m (length × width × depth, where depth means the depth of water) as shown in Figure 11, and the ROV ran slowly.

All the data from sensors were obtained to calculate the localization and orientation of the ROV precisely, and the tick mark of the water tank is used as the ground truth to evaluate the localization accuracy of the proposed algorithm, while the orientation calculated from the magnetometer after calibration is used as the ground truth to assess the orientation accuracy of the algorithm. The value of θ in the base reference is positive when the ROV rotates in a clockwise direction. The results of the experiments are as Figures 12 and 13.

The fitting straight lines of the first and the second experiments is as in (19), in which the order is a, b, c, and d.

Table 1: Results of the experiment.

Experiments	Water tank		Pose of ROV		
	Length (m)	Width (m)	x (m)	y (m)	θ (°)
1st					
Test results	6.383	3.0963	1.7818	2.2843	83.293
Ground truth	6.0	3.0	1.80	2.30	84.52
Error	0.383	0.0963	−0.0182	−0.0157	−1.227
2nd					
Test result	6.0238	3.0528	1.9624	1.6965	43.38
Ground truth	6.0	3.0	2.0	1.70	44.56
Error	0.0238	0.0528	−0.0376	−0.0035	−0.18

$$\mathbf{y}_1 = \begin{bmatrix} -8.5 \\ 0.1176 \\ -8.5 \\ 0.1176 \end{bmatrix} \mathbf{x}_1 + \begin{bmatrix} -15.25 \\ 2.3 \\ 11.25 \\ -3.78 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} -0.945 \\ 1.0582 \\ -0.945 \\ 1.0582 \end{bmatrix} \mathbf{x}_2 + \begin{bmatrix} -2.7 \\ 2.47 \\ 1.5 \\ -6.3 \end{bmatrix}. \tag{19}$$

The localization and orientation deduced from the two experiments are as in Table 1. The localization coordinate refers to the base reference in Figure 5.

From Table 1, we can see that the error is very low and the accuracy of localization and orientation is very high. The biggest advantage compared with other underwater

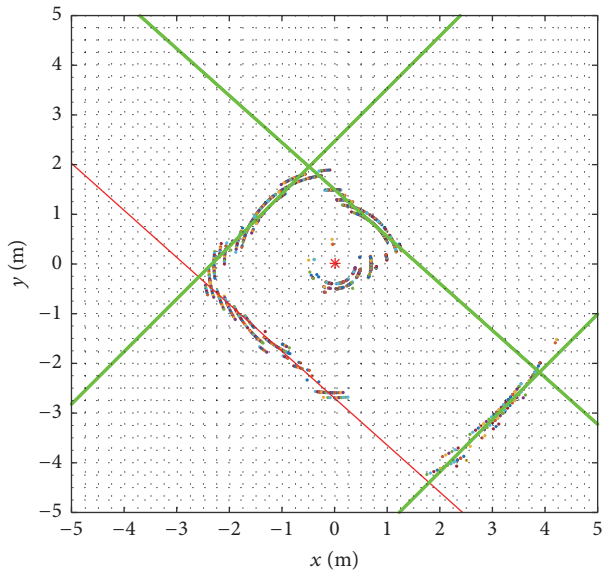


Figure 9: Rectangle after geometric modification.

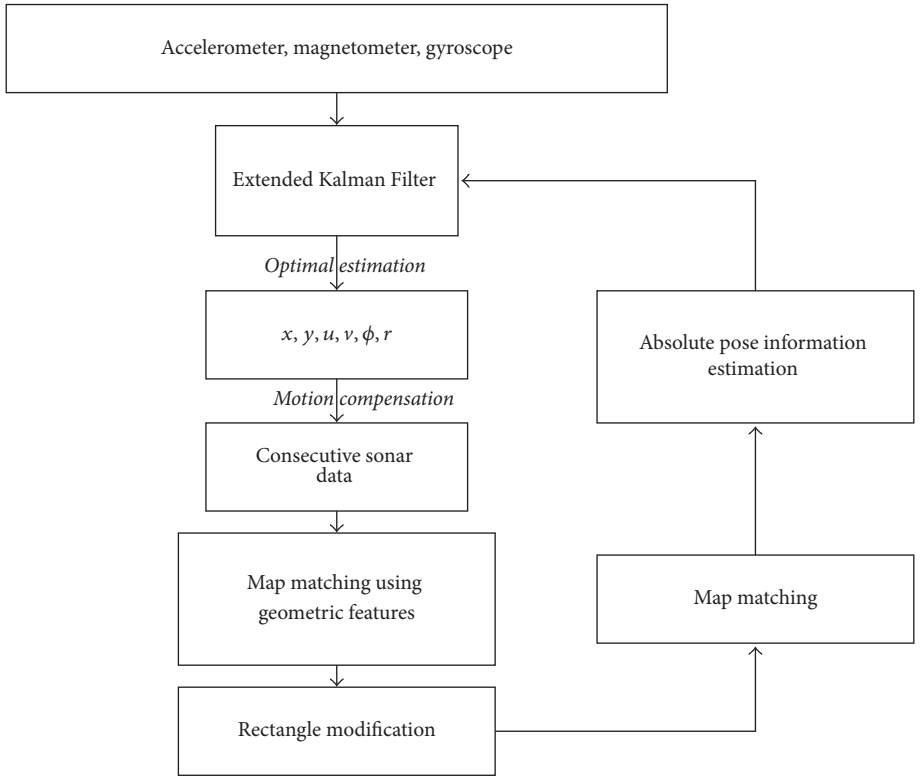


Figure 10: The entire recursive process.

navigation algorithms is that the proposed algorithm is based on the geometric features and the a priori map matching. In other words, the algorithm is not interfered with by the running time and the accuracy error will not accumulate as time goes on just like GPS works on land, which is very important in practical application.

6. Conclusion

The paper proposes a new underwater navigation algorithm based on geometric features and map matching using mechanical scanning sonar in known man-made structured environment. The motion induced distortion of the sonar point cloud data is modified and compensated for using EKF

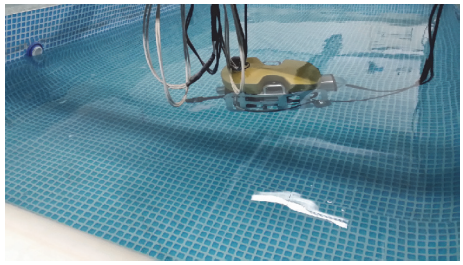


Figure 11: Experiment in water tank.

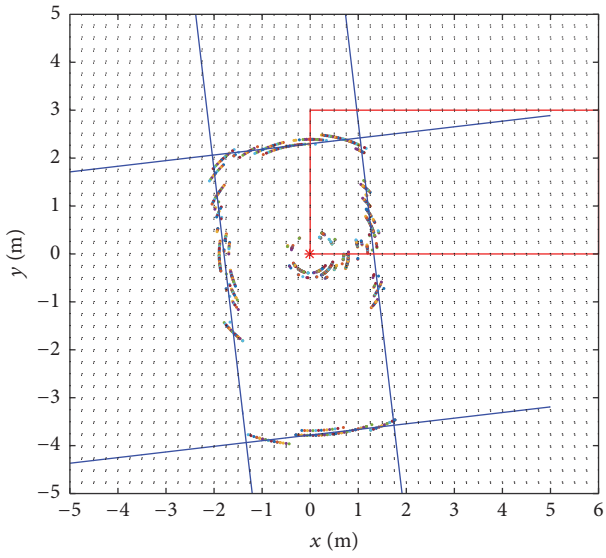


Figure 12: The first experiment.

with all the sensor data fused. Then the boundary of the water tank is carried out using Hough Transform combined with the least square fitting. To modify the error of the line fitting, the line with most sonar data is used as the base line with all the three other lines modified based on geometric features. After calculating the rectangle figure of the water tank, we can easily deduce the localization and orientation of the ROV. Water tank environment verifies the validation of the proposed algorithm.

The ROV we use as the platform is designed to help monitor underwater environment and salvage small parts like bolts and nuts in the reactor pool and other water-filled infrastructure of the nuclear power plants, and it has been experimented in the reactor simulation pool of the Daya Bay Nuclear Power Plant many times to test the leaking tightness and movement stability as we introduce in paper [12]. In the current time, this is sufficient for us to know the vehicle's accurate localization and orientation when the ROV moves slowly or even suspends under the water in specific time to guarantee the safety of the nuclear equipment. More experiments will be carried out in the reactor simulation pool of the nuclear power plant and the algorithm will get further validation.

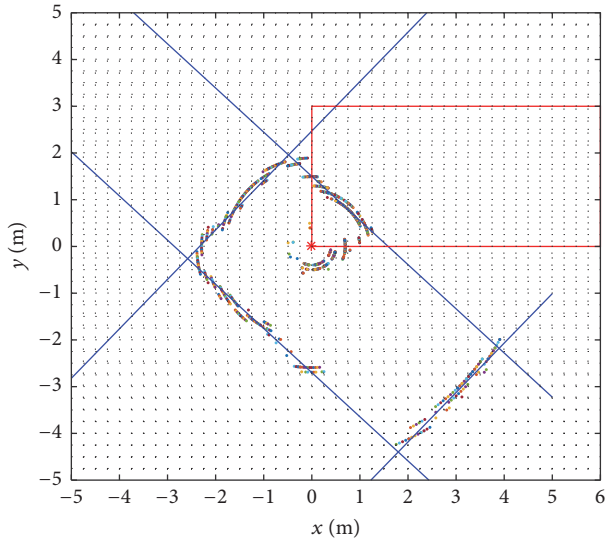


Figure 13: The second experiment.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to acknowledge the support of Natural National Key Basic Research Program of China under Grant no. 2013CB035503 and the National High Technology Research and Development Program of China under Grant no. 2011AA040201.

References

[1] Y. Shim, J. Park, and J. Kim, "Relative navigation with passive underwater acoustic sensing," in *Proceedings of the 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI '15)*, pp. 214–217, IEEE, Goyang, South Korea, October 2015.

[2] D. Ribas, P. Ridao, A. Mallios, and N. Palomeras, "Delayed state information filter for USBL-Aided AUV navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 4898–4903, May 2012.

[3] J. C. Kinsey, R. M. Eustice, and L. L. Whitcomb, "A survey of underwater vehicle navigation: recent advances and new challenges," in *Proceedings of the IFAC Conference of Manoeuvering and Control of Marine Craft*, p. 88, Lisbon, Portugal, September 2006.

[4] L. Stutters, H. Liu, C. Tiltman, and D. J. Brown, "Navigation technologies for autonomous underwater vehicles," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 38, no. 4, pp. 581–589, 2008.

[5] I. Mahon, S. B. Williams, O. Pizarro, and M. Johnson-Roberson, "Efficient view-based SLAM using visual loop closures," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1002–1014, 2008.

[6] A. Elibol, N. Gracias, and R. Garcia, "Augmented state–extended Kalman filter combined framework for topology estimation in large-area underwater mapping," *Journal of Field Robotics*, vol. 27, no. 5, pp. 656–674, 2010.

[7] H. Liu, J. Qin, F. Sun, and D. Guo, "Extreme kernel sparse learning for tactile object recognition," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–12, 2016.

[8] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-Tactile Fusion for Object Recognition," *IEEE Transactions on Automation Science and Engineering*, vol. PP, no. 99, pp. 1–13, 2016.

[9] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: kernel sparse coding methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

[10] H. Liu, Y. Liu, and F. Sun, "Robust exemplar extraction using structured sparse coding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 8, pp. 1816–1821, 2015.

[11] A. Mallios, P. Ridao, D. Ribas, and E. Hernández, "Scan matching SLAM in underwater environments," *Autonomous Robots*, vol. 36, no. 3, pp. 181–198, 2014.

[12] E. Hernández, P. Ridao, D. Ribas, and J. Batlle, "MSISpIC: a probabilistic scan matching algorithm using a mechanical scanned imaging sonar," *Journal of Physical Agents*, vol. 3, no. 1, pp. 3–12, 2009.

[13] A. Burguera, Y. González, and G. Oliver, "The UspIC: performing scan matching localization using an imaging sonar," *Sensors*, vol. 12, no. 6, pp. 7855–7885, 2012.

[14] B. He, Y. Liang, X. Feng et al., "AUV SLAM and experiments using a mechanical scanning forward-looking sonar," *Sensors*, vol. 12, no. 7, pp. 9386–9410, 2012.

[15] L. Chen, S. Wang, H. Hu, D. Gu, and L. Liao, "Improving localization accuracy for an underwater robot with a slow-sampling sonar through graph optimization," *IEEE Sensors Journal*, vol. 15, no. 9, pp. 5024–5035, 2015.

[16] M. Dong, W. Chou, B. Fang, G. Yao, and Q. Liu, "Implementation of remotely operated vehicle for direct inspection of reactor pressure vessel and other water-filled infrastructure," *Journal of Nuclear Science and Technology*, vol. 53, no. 8, pp. 1086–1096, 2016.

[17] D. Ribas, P. Ridao, and J. Neira, *Underwater SLAM for Structured Environments Using an Imaging Sonar*, Springer, Berlin, Germany, 2010.

[18] D. Ribas, P. Ridao, J. D. Tardós, and J. Neira, "Underwater SLAM in man-made structured environments," *Journal of Field Robotics*, vol. 25, no. 11-12, pp. 898–921, 2008.

[19] A. Mallios, P. Ridao, M. Carreras, and E. Hernández, "Navigating and mapping with the SPARUS AUV in a natural and unstructured underwater environment," in *Proceedings of the MTS/IEEE Kona Conference (OCEANS '11)*, pp. 1–7, Kona, Hawaii, USA, September 2011.

[20] B. Fang, W.-S. Chou, M.-J. Dong, X. Ma, and X.-Q. Guo, "Location algorithm of underwater robot based on the probabilistic iterative correspondence," *Journal of Electronics & Information Technology*, vol. 36, no. 4, pp. 993–997, 2014.

[21] W. A. Barrett and K. D. Petersen, "Houghing the hough: peak collection for detection of corners, junctions and line intersections," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 2, Kauai, Hawaii, USA, December 2001.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 8035089, 9 pages
http://dx.doi.org/10.1155/2016/8035089

Research Article

A Cost-Sensitive Sparse Representation Based Classification for Class-Imbalance Problem

Zhenbing Liu,[1] Chunyang Gao,[2] Huihua Yang,[1,3] and Qijia He[2]

[1]*School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China*
[2]*School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China*
[3]*School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Correspondence should be addressed to Chunyang Gao; 935535775@qq.com

Received 8 August 2016; Accepted 16 October 2016

Academic Editor: Kun Hua

Sparse representation has been successfully used in pattern recognition and machine learning. However, most existing sparse representation based classification (SRC) methods are to achieve the highest classification accuracy, assuming the same losses for different misclassifications. This assumption, however, may not hold in many practical applications as different types of misclassification could lead to different losses. In real-world application, much data sets are imbalanced of the class distribution. To address these problems, we propose a cost-sensitive sparse representation based classification (CSSRC) for class-imbalance problem method by using probabilistic modeling. Unlike traditional SRC methods, we predict the class label of test samples by minimizing the misclassification losses, which are obtained via computing the posterior probabilities. Experimental results on the UCI databases validate the efficacy of the proposed approach on average misclassification cost, positive class misclassification rate, and negative class misclassification rate. In addition, we sampled test samples and training samples with different imbalance ratio and use F-measure, G-mean, classification accuracy, and running time to evaluate the performance of the proposed method. The experiments show that our proposed method performs competitively compared to SRC, CSSVM, and CS4VM.

1. Introduction

As a powerful tool for statistical signal modeling, sparse representation (or sparse coding) has been successfully used in pattern recognition fields [1], such as texture classification [2] and face recognition [3, 4], in the past few years. In [3], John et al. proposed a sparse representation based classification (SRC) method when they solve the face recognition under various illuminations and occlusions, which represents an input test image as a sparse linear combination of training images and assigned the test image to the class whose training samples can best reconstruct it. In theirs work, they used L_1-regularizer rather than L_0-regularizer to regularize the objective function and then calculated the residuals between the original test sample and the reconstructed one to identify the query image's label. Such a sparse representation based classification framework has achieved a great success in face recognition and has boosted the research of sparsity related machine learning methods.

Traditional classification algorithms [5], including SRC, are designed to achieve the lowest recognition errors and assume the same losses for different types of misclassifications. However, this assumption may not be suitable for many real-world applications. For example, it may cause inconvenience to a gallery who is misclassified as an impostor and not allowed to enter the room controlled by a face recognition system but may result in a serious loss if an impostor is misclassified as a gallery and allowed entering the room. In such settings, the loss of misclassification should be taken into consideration, and "cost" information can be introduced to measure the severity of misclassification. In recent years, many cost-sensitive methods have been proposed. The typical works include the Cost-Sensitive Semisupervised Support Vector Machine (CS4VM) and Cost-Sensitive Laplacian Support Vector Machines (CSLSVM) proposed by Zhou et al. [6, 7], a cost-sensitive Naïve Bayes method from a novel perspective of inferring the order relation [8] proposed by Fang et al., and novel cost-sensitive approach proposed by

Castro and Braga to improve the performance of multilayer perceptron [9]. In [10], an instance weighting method was incorporated into various Bayesian network classifiers. The probability estimation of Bayesian network classifiers was modified by the instance weighting method, which made Bayesian network classifiers cost-sensitive. In [11], Lo et al. presented a basis expansions model for multilabel classification to handle the cost-sensitive multilabel classification problem, where a basis function is an LP classifier trained on a random k-label set. In [12], Wan et al. proposed a cost-sensitive feature selection method called Discriminative Cost-Sensitive Laplacian Score (DCSLS) for face recognition, which incorporated the idea of local discriminant analysis into Laplacian Score.

Cost-sensitive learning always coexists with class-imbalance in most applications with the goal of minimizing the total misclassification cost [13]. Class-imbalance has been considered as one of the most challenging problems in machine learning and data mining. The ratio of imbalance (the size of majority class to minority class) can be as huge as 100, even up to 10000. Much work has been done in addressing the class-imbalance problem. Cost-sensitive learning is an effective method to deal with the imbalance data classification problem. In recent year, cost-sensitive learning has been studied widely and become one of the most important topics for solving the class-imbalance problem. In [14], Zhou and Liu studied empirically the effect of sampling and threshold-moving in training cost-sensitive neural networks and revealed that threshold-moving and soft-ensemble are relatively good choices in training cost-sensitive neural networks. There are also some other cost-sensitive learning methods by improving the existed method. In [15], Sun et al. proposed a cost-sensitive boosting algorithms, which are developed by introducing cost items into the learning framework of AdaBoost. Another strategy for class-imbalance problem is based on exchanging the distribution of data sets. In [16], Jiang et al. proposed a novel Minority Cloning Technique (MCT) for class-imbalanced cost-sensitive learning. MCT alters the class distribution of training data by cloning each minority class instance according to the similarity between it and the mode of the minority class. Generally, users focus more on the minority class and consider the cost of misclassifying a minority class to be more expensive. In our study, we adopt the same strategy to address this problem.

In [17], a probabilistic cost-sensitive classifier was proposed for face recognition; they utilize the probabilistic model to estimate the posterior probability of a testing sample and calculate all the misclassification losses via the posterior probabilities. Motivated by this probabilistic model and probabilistic subspace clustering [17–19], we proposed a new method to handle misclassification cost. In sparse representation, it will play an important role for reconstruction if the value of coefficient is higher [20]. In other words, the coefficient is 1 when a query sample was represented by a dictionary with the same sample as the query one. Just like Gaussian distribution, a sample that is close to the mean vector has a higher probability. Inspired by probabilistic model, we use coefficient matrix to calculate the posterior probabilities rather than the distribution of noise (residual) in [17] and they have to estimate the distribution of noise. The main advantage of our method is to reduce the computation complexity and computation cost, and the contribution of the proposed method is obtaining the posterior probability by coefficient vector of sparse representation. After calculating all the misclassification losses via the posterior probabilities, the test sample is assigned to the class whose loss is minimal. Experimental results on UCI databases validate the effectiveness and efficiency of our methods.

This paper is organized as follows. Section 2 outlines the details of the relevant method. Section 3 presents the details of the proposed algorithm. Section 4 reports the experiments. Finally, Section 5 concludes the paper and offers suggestions for future research.

2. Related Works

In this section, we briefly introduce some related works, including sparse representation based classification and cost-sensitive learning framework.

2.1. Sparse Representation Based Classification. Sparse representation is a typically method in machine learning [3, 21, 22], which is to use labeled training samples from k distinct object classes to learn a dictionary and determine the label of an unseen new test sample correctly. We denote the data set with k_i training samples from the ith class as a matrix $A_i = [v_{i,1}, v_{i,2}, \ldots, v_{i,n_i}] \in R^{m \times k_i}$ and $n = \sum_{i=1}^{k} k_i$ is the number of all training samples, where k is the number of classes in training set. Given sufficient training samples of the ith class, any test sample $y \in R^m$ from the same class will be approximately represented linearly by the training samples of class i:

$$y = x_{i,1} v_{i,1} + x_{i,2} v_{i,2} + \cdots + x_{i,k_i} v_{i,k_i}. \tag{1}$$

Then, rewrite the above representation of y in matrix form as $y = A_i x_i$, where $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,k_i}] \in R^{k_i}$. Then, define a new matrix A for the entire training set as follows:

$$\begin{aligned} A &= [A_1, A_2, \ldots, A_k] \\ &= \left[v_{1,1}, v_{1,2}, \ldots, v_{1,n_i}, \ldots, v_{i,1}, \ldots, v_{i,n_i}, \ldots, v_{k,1}, \ldots, v_{k,n_k}\right] \\ &\in R^{m \times n}. \end{aligned} \tag{2}$$

Many method based distances are not robust in real-world applications because of various occlusions. To overcome this limitations, Wright introduced the sparse representation based classification method to represent the query image. Then, the linear representation of y can be rewritten in terms of all training samples as

$$y = A x_0, \tag{3}$$

where $x_0 = [0, \ldots, 0, x_{i,1}, x_{i,2}, \ldots, x_{i,k_i}, 0, \ldots, 0]^{\mathrm{T}} \in R^n$, whose entries are zero except those associated with the ith class. This motivates us to seek the sparsest solution by solving the following optimization problem:

$$\widehat{x}_0 = \arg\min \|x\|_0 \quad \text{s.t. } Ax = y, \tag{4}$$

where $\|\cdot\|_0$ denotes the L_0-norm, which counts the number of nonzero entries in a vector. However, the above problem of finding the sparsest solution (L_0-norm minimization problem) is nonconvex and actually NP-hard. Generally, if the solution sought is sparse enough, the solution of the L_0-minimization problem is equal to the solution of the following L_1-minimization problem [4, 22, 23]:

$$\widehat{x}_1 = \arg\min \|x\|_1 \quad \text{s.t. } Ax = y. \tag{5}$$

The real data are noisy; it may not represent the test sample exactly. To deal with the noises, John et al. extended the L_1-norm minimization problem to the following formulation:

$$y = Ax_0 + z, \tag{6}$$

where z is a noise term with bounded energy $\|z\|_2 < \varepsilon$. The sparse solution x_0 can still be obtained by solving the following stable L_1-minimization problem:

$$\widehat{x}_1 = \arg\min \|x\|_1 \quad \text{s.t. } \|y - Ax\|_2 \le \varepsilon. \tag{7}$$

To better harness such linear structure, they instead classify y based on how well the coefficients associated with all training samples of each object reproduce y. Let $\widehat{x}$ be the solution of (7), for each class i, let δ_i be the characteristic function that selects the coefficients associated with the ith class. Using the coefficients, one can approximate the given test sample y as $\widehat{y}_i = A\delta_i(\widehat{x})$, where $\delta_i(\widehat{x}) = [0, \ldots, x_{i,1}, x_{i,2}, \ldots, x_{i,k_i}, \ldots, 0]^{\mathrm{T}}$. They then compute the residual (Euclidean distance) $r_i(y)$ between y and $\widehat{y}$:

$$r_i(y) = \|y - A\delta_i(\widehat{x})\|_2. \tag{8}$$

The label of the test sample y can be identified by minimizing $r_i(y)$ as follows:

$$L(y) = \arg\min_i r_i(y). \tag{9}$$

2.2. Cost-Sensitive Function. In multiclass cost-sensitive learning, considering c gallery subjects with their class labels $G = \{G_i\}_{i=1,2,\ldots,c}$, many impostors, whose labels are I. In [7], Zhang and Zhou categorized the costs into three types: cost of false acceptance C_{IG}, cost of false rejection C_{GI}, and cost of false identification C_{GG}. Empirically, it is evident that C_{IG}, C_{GI}, and C_{GG} are unequal. Give a cost setting according to the users and reassign $C_{IG} = C_{IG}/C_{GG}$, $C_{GI} = C_{GI}/C_{GG}$, and $C_{GG} = 1$. Here, for the ease of understanding, we still preserve the original formulation. We can construct a multiclass cost matrix C as shown in

	G_1	$\cdots$	G_M	I
G_1	0	$\cdots$	C_{GG}	C_{GI}
$\vdots$	$\vdots$	0	$\vdots$	$\vdots$
G_M	C_{GG}	$\cdots$	0	C_{GI}
I	C_{IG}	$\cdots$	C_{IG}	0

(10)

where C_{ij} indicates the cost of misclassifying a sample of the ith class as the jth class. The diagonal elements of C are all zero since there is no loss for correct recognition.

Cost-sensitive learning usually sets the misclassification cost as objective function and identifies the label by minimizing loss function. Given a test sample y and its predicted class label as $\phi(y)$, respectively, the label is obtained by minimizing the objective function:

$$L(y) = \arg\min_{\phi(y)\in\{G_1,\ldots,G_c,I\}} \operatorname{loss}(y, \phi(y)), \tag{11}$$

where

$$\operatorname{loss}(y, \phi(y)) = \begin{cases} \sum_{i=1}^{c} P(G_i \mid y) C_{GI} & \text{if } \phi(y) = I \\ \sum_{\substack{i=1 \\ i\neq\tau}}^{c} P(G_i \mid y) C_{GG} + P(I \mid y) C_{IG}, & \text{if } \phi(y) = G_\tau, \end{cases} \tag{12}$$

where $\widehat{\phi}(y)$ is the optimal prediction of y and c represents the gallery subjects in classification problem.

3. Cost-Sensitive SRC

In [5], Alpaydın calculated the residuals to identify the class label of a test sample y, which is the Euclidean distance between reconstructed sample and the original test sample y. In cost-sensitive learning, the loss function (see (7)) is regarded as an objective function to identify the label of a test sample. In binary classification problem, there are two misclassification costs, and we denote the cost that misclassifies positive class as negative class by C_{10} and the cost by C_{01} conversely. Then a cost matrix can be constructed as shown in

	G_0	G_1
G_0	0	C_{01}
G_1	C_{10}	0

(13)

where G_1 and G_0 represents the label of minority class and majority class, respectively.

It is well known that the loss function can be related to the posterior probability $P(\phi(y) \mid y) \approx P(\delta_i(x_i) \mid y)$. Then the loss function can be rewritten as follows:

$$\operatorname{loss}(y, \phi(y)) = \begin{cases} \sum_{i=G_1} P(\delta_i \mid y) C_{10} & \text{if } \phi(y) = G_0 \\ \sum_{j=G_0} P(\delta_j \mid y) C_{01} & \text{if } \phi(y) = G_1. \end{cases} \tag{14}$$

The test sample y belongs to the class with higher probability. Now, we will estimate $P(\delta_i(x) \mid y)$, $(i = 0, 1)$.

In coefficient matrix, the larger the element value is, the more important the role it will play for reconstructing a test sample. In other words, it is best to represent the test sample by training samples and they have the same class label, and there are no samples from different class in this linear combination. The posterior probability can be related to the coefficient matrix. Accordingly, we rewrite the solution of (7) as $\widehat{x} = [x^+ \; x^-]^{\mathrm{T}}$, where $x^+ = [x_1^+, x_2^+, \ldots, x_{n^+}^+] \in R^{n^+}$ and

$x^- = [x_1^-, x_2^-, \ldots, x_{n^-}^-] \in R^{n^-}$ represent the positive class coefficient and negative class coefficient, respectively. Here, n^+ is the number of positive samples and n^- is the number of negative samples in dictionary. Then, we can obtain the posterior probabilities:

$$P(\delta_i \mid y) = \sum_{i=1}^{n^+} \frac{x_i^+}{X} \quad \text{if } i = G_1,$$
$$P(\delta_j \mid y) = \sum_{j=1}^{n^-} \frac{x_j^-}{X} \quad \text{if } j = G_0, \tag{15}$$

where $X = \sum_{j=1}^{n^-} x_j^- + \sum_{i=1}^{n^+} x_i^+$. Then, (14) can be written as

$$\text{loss}(y, \phi(y)) = \begin{cases} \sum_{i=1}^{n^+} \frac{x_i^+}{X} C_{10} & \text{if } \phi(y) = G_1 \\ \sum_{j=1}^{n^-} \frac{x_j^-}{X} C_{01} & \text{if } \phi(y) = G_0. \end{cases} \tag{16}$$

We can obtain the label of a test sample y by minimizing (16):

$$L(y) = \arg\min_{i \in \{0,1\}} \text{loss}(y, \phi(y)). \tag{17}$$

The whole process of CSSRC is described in Algorithm 1.

Algorithm 1 (CSSRC algorithm).

Input. Dictionary $A \in R^{m \times n}$, test sample $y \in R^m$

Output. The label $L(y)$ of test sample y

(1) Normalize the columns of A to unit L_2-norm

(2) Solve the L_1-minimization problem:

$$\widehat{x} = \arg\min \|x\|_1 \quad \text{s.t. } y = Ax \tag{18}$$

Or alternatively, solve

$$\widehat{x} = \arg\min \|x\|_1 \quad \text{s.t. } \|Ax - y\|_2 \le \varepsilon \tag{19}$$

Assume the solution is $\widehat{x} = [x^+ \;\; x^-]^{\mathrm{T}}$

(3) Calculate the loss function:

$$\text{loss}(y, \phi(y)) = \begin{cases} \sum_{i=1}^{n^+} \frac{x_i^+}{X} C_{10} & \text{if } \phi(y) = G_1 \\ \sum_{j=1}^{n^-} \frac{x_j^-}{X} C_{01} & \text{if } \phi(y) = G_0, \end{cases} \tag{20}$$

where $X = \sum_{j=1}^{n^-} x_j^- + \sum_{i=1}^{n^+} x_i^+$

(4) Obtain the label of y:

$$L(y) = \arg\min_{i \in \{0,1\}} \text{loss}(y, \phi(y)). \tag{21}$$

Table 1: Description of data sets.

Dataset	Size	Target	Ratio	min/maj
Abalone	4117	Ring = 7	9.7	391/3786
Housing	506	[20, 23]	3.8	106/400
Nursery	12960	Very-recom	38.5	328/12632
Letter	20000	A	24.3	789/19211
Pima	786	Class 1	1.7	268/500
Cmc	1473	Class 2	3.4	333/1344
Car	1728	acc	3.5	384/1344

4. Experiments

4.1. Data Sets and Experimental Setting. We test the proposed method on seven UCI data sets [24]. Detailed information about these data sets is summarized in Table 1.

In cost-sensitive learning, false positive (actual negative but predicted as positive, denoted as FP), false negative (actual positive but predicted as negative, FN), true positive (actual positive and predicted as positive, TP), and true negative (actual negative and predicted as negative, TN) can be given in a confusion matrix as follows:

	Positive Class	Negative Class
Positive Class	TP	FN
Negative Class	FP	TN

(22)

To binary classification problems, four kinds of misclassification cost are needed, which is referred to as CTP, CFP, CTN, and CFN, respectively. CTP and CTN are the costs of true positive (TP) and true negative (TN). In order to simplify the cost matrix, we set CTP = 0 and CTN = 0. CFN and CFP are the costs of false negative (FN) and false positive (FP). We always assume that the cost of misclassifying positive class instances is much higher than the cost of misclassifying negative class instances, so we set CFN $\gg$ CFP. In this paper, CFP is set to be a unit cost of 1; CFN is assigned different values: $5, 10, 15, \ldots, 50$, respectively. In our experiments, we adopt 10-folder cross-validation to get the average cost, and three evaluation criteria are adopted to evaluate the classification performance in cost-sensitive experiments: average cost (AC), error rate of false acceptance (Err(IG)), and error rate of false rejection (Err(GI)). For class-imbalance problem, we choose F-measure and G-mean to evaluate the performance. They are defined as follows [25, 26]:

$$\text{AC} = \frac{C_{10}|\text{FP}| + C_{01}|\text{FN}|}{N},$$
$$\text{Err}(GI) = \frac{|\text{FP}|}{N^+},$$
$$\text{Err}(IG) = \frac{|\text{FN}|}{N^-},$$

Table 2: Average cost of the four methods (cost ratio 1 : 10).

Methods	CSSRC	SRC	CSSVM	CS4VM
Abalone	**0.0122**	0.6585	1.5905	0.4590
Housing	**0.0610**	0.1585	1.6333	0.4190
Nursery	**0.0488**	0.3902	0.3110	0.0817
Letter	**0.1220**	0.6219	0.1774	0.3378
Pima	**0.0122**	2.4390	2.4495	0.5027
Cmc	0.5122	0.5488	1.9905	**0.5105**
Car	**0.0244**	0.0976	0.6714	0.2610

$$\begin{aligned}
\text{Recall} = \text{Acc}_+ &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\
\text{Acc}_- &= \frac{\text{TN}}{\text{TN} + \text{FP}},\\
\text{Accuracy} &= \text{Acc}_- + \text{Acc}_+,\\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\
G\text{-mean} &= \sqrt{\text{Acc}_+ \times \text{Acc}_-},\\
F\text{-measure} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},
\end{aligned} \tag{23}$$

where |FP| and |FN| represent the number of false acceptances and false rejections, respectively. N, N^+, and N^- represent the number of test samples, positive class samples, and negative class samples, and $N = N^+ + N^-$.

In order to illustrate the performance of CSSRC, sparse representation based classification (SRC), Cost-Sensitive Support Vector Machine (CSSVM), and Cost-Sensitive Semisupervised Support Vector Machine (CS4VM) are chosen to compare the performance on three experiments. The experiments are performed on Matlab 2014a and the computer with a 2.6 GHz Intel Xeon CPU.

4.2. The Effect of Cost for SRC. For data set Housing, the size is smaller than the other six data sets, so less samples are selected for train set and test set. We select 31 positive samples and 31 negative samples randomly from Housing as test samples and 41 positive samples and 41 negative samples as training samples. We select 61 positive samples and 61 negative samples as test samples from Abalone, Nursery, Letter, Pima, Cmc, and Car and 101 positive samples and 101 negative samples as training samples. We repeat sampling 100 times and get the average results.

Experiment 1. We compare the performance of these 4 methods (CSSRC, SRC, CSSVM, and CS4VM) on Abalone, Nursery, Letter, Pima, Cmc, Housing, and Car. We set cost ratio (the cost of false acceptance respect to false rejection) as 10, and the results are summarized in (22). From Table 2, we can see that the proposed cost-sensitive approach achieves lower average misclassification cost than the other three methods on Abalone, Nursery, Letter, Pima, Housing, and Car except Cmc. CSSRC's average cost is higher than CS4VM but lower than the other two methods on Cmc and lower than CS4VM on the other 6 data sets. The average cost of CSSRC is 0.5122 and CS4VM's average cost is 0.5105. They are in the same order of magnitudes. In other words, our method has better performance than SRC, CSSVM, and CS4VM.

Experiment 2. According to the results in Experiment 1, we plot two pictures from Figure 1. For either positive class or negative class, the proposed method can achieve a lower error rate on Nursery and Abalone when the cost ratio ranges from 5 to 50. Although CS4VM can obtain a lower error rate of false rejection, its error rate of false acceptance is very high, this can generate a serious total cost. From Figure 1, we can easily find that our method can achieve lower error rate of false rejection and lower error rate of false acceptance simultaneously.

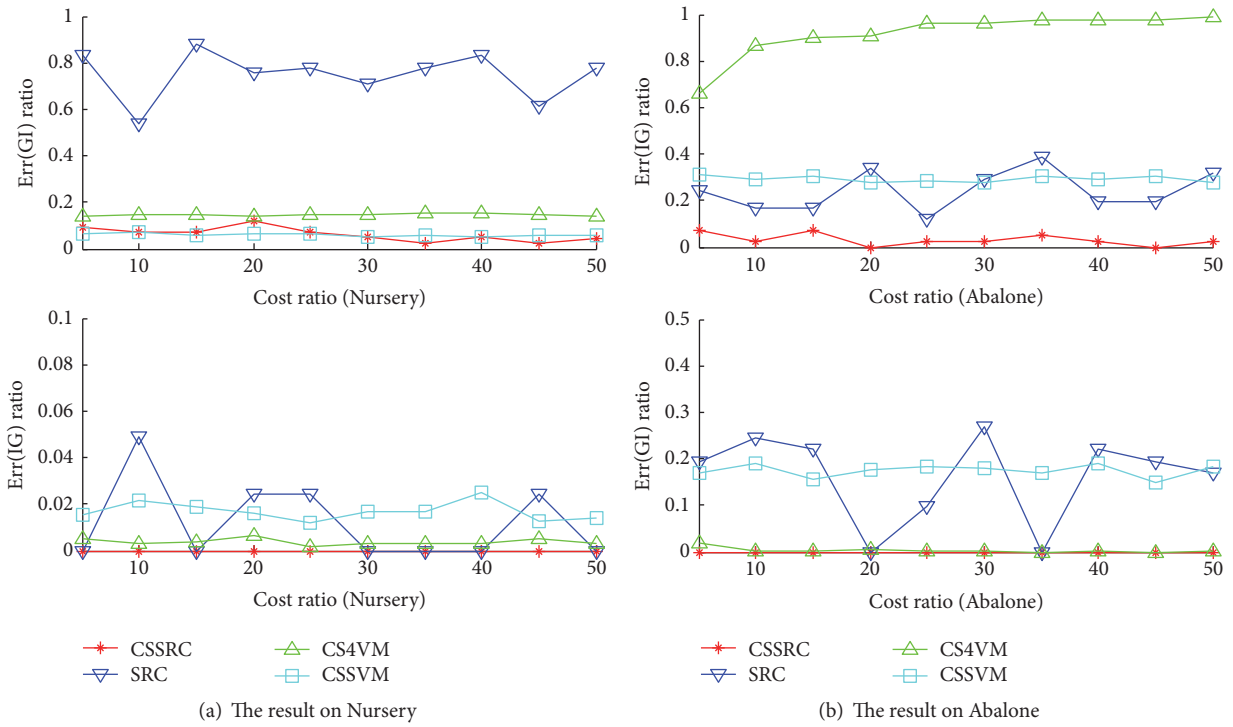
Experiment 3. In this section, we set cost ratio from 10 to 50, and the results are summarized as Table 3. The first row is coat ratios and the first two columns represent data sets and classification methods, respectively. In Experiment 2 we use merely two data sets, for proving the robust of our method; more data sets are adopted in this experiment. Our proposed cost-sensitive SRC achieved a lower average costs on four data sets. Although it is not the lowest cost on Nursery and Letter, it has the same order of magnitude as the lowest cost value.

The above there experiments have proved the effect of cost term for SRC. Particularly, the comparison of SRC and CSSRC can well validate the conclusion that the cost term can improve the performance of SRC.

4.3. Solving Class-Imbalance Problem

Experiment 1. In this section we will solve the class-imbalance problem. Table 1 has summarized the information of data sets we used, and the imbalance ratio higher 10 is Nursery and Letter. In order to set a higher imbalance ratio, we select Nursery in this experiment. Similarly, we compare the performance of these four methods (SRC, CSSVM, CS4VM, and CSSRC) on Nursery. It is difficult to reflect the performance of our method for class-imbalance problem, and F-measure, G-mean, and classification accuracy have been adopted for the class-imbalance problem. In this experiment, we take the imbalance ratio from $[1, 2, \ldots, 10]$, respectively. The size of minority class is 30 and the majority class is 30 multiplying the imbalance ratios in training set, accordingly. We select 61 positive samples and 61 negative samples as test set and run and summarize the results as in Figures 2 and 3; the sampling process has repeated 100 times and gets the average results.

Figure 2 shows the results of F-measure on Nursery, and the definition of F-measure (the harmonic mean between the classification accuracy of positive class and the classification accuracy of negative class) has been shown in Section 4.1. It is obvious that our method has achieved a higher F-measure value with respect to sparse representation based classification, Cost-Sensitive Support Vector Machine,



(a) The result on Nursery (b) The result on Abalone

Figure 1: Error rate of false acceptance and false rejection.

Table 3: Average cost of methods on five data sets with cost ratio from 10 to 50.

Cost ratio	10	15	20	25	30	35	40	45	50
Abalone									
CSSRC	**0.365**	**0.548**	**0.487**	**0.304**	**0.365**	**0.426**	**0.975**	**0.975**	**0.609**
SRC	1.621	1.646	2.256	2.512	5.487	4.365	5.853	2.280	9.146
CS4VM	4.267	7.036	9.498	12.024	14.561	17.124	19.220	21.809	24.183
CSSVM	1.615	2.213	2.746	3.795	4.175	5.569	6.238	6.542	7.643
Nursery									
CSSRC	0.365	**0.365**	0.731	**0.304**	**0.365**	**0.426**	**0.975**	**0.548**	1.829
SRC	4.268	6.585	6.341	6.500	12.804	11.536	13.670	19.207	22.560
CS4VM	0.670	1.001	1.416	1.764	2.108	2.700	2.753	3.152	3.733
CSSVM	**0.348**	0.394	**0.691**	0.836	0.895	1.204	1.204	1.531	**1.613**
Housing									
CSSRC	**0.487**	**0.365**	**0.000**	**0.000**	**0.000**	**0.426**	**0.000**	**0.000**	**0.000**
SRC	2.439	2.561	1.256	6.707	5.853	8.109	6.341	12.073	12.195
CS4VM	3.765	6.007	8.091	10.046	11.975	13.491	15.878	17.793	19.420
CSSVM	1.689	2.544	3.106	4.326	4.983	5.228	6.703	6.265	7.647
Letter									
CSSRC	**0.024**	**0.049**	**0.024**	**0.036**	**0.012**	0.061	**0.036**	0.085	**0.037**
SRC	0.036	**0.049**	0.025	0.305	0.036	**0.024**	0.049	**0.049**	1.244
CS4VM	0.606	1.150	1.146	1.483	1.891	2.282	2.405	2.760	3.983
CSSVM	0.158	0.277	0.388	0.298	0.409	0.319	0.466	0.757	0.798
Car									
CSSRC	**0.000**	**0.183**	**0.488**	**0.000**	**0.366**	**0.000**	**0.488**	**0.548**	**0.610**
SRC	1.097	1.463	2.195	3.354	1.463	5.549	4.878	5.134	5.488
CS4VM	2.500	3.849	5.307	6.556	7.939	9.365	10.605	11.514	12.854
CSSVM	1.335	1.486	1.711	2.250	2.388	2.561	2.570	2.895	2.998

Table 4: Classification accuracy on Nursery.

Accuracy	1	2	3	4	5	6	7	8	9	10
SRC	0.9298	0.9305	0.9343	0.9412	0.9420	0.9460	0.9484	0.9484	0.9530	0.9572
CS4VM	0.9186	0.9256	0.9307	0.9345	0.9335	0.9414	0.9400	0.9396	0.9369	0.9421
CSSVM	0.9279	0.9398	0.9314	0.9393	0.9345	0.9319	0.9285	0.9279	0.9285	0.9294
CSSRC	**0.9601**	**0.9623**	**0.9610**	**0.9585**	**0.9610**	**0.9600**	**0.9610**	**0.9627**	**0.9597**	**0.9616**

Table 5: Running time on Nursery.

Time/s	1	2	3	4	5	6	7	8	9	10
SRC	0.0105	0.0385	0.0664	0.0973	0.1324	0.1721	0.2158	0.2620	0.3137	0.3702
CS4VM	1.7131	1.7114	1.9854	2.0866	2.0969	2.2246	2.2572	2.3539	2.3044	2.4455
CSSVM	3.7136	8.5207	15.1722	23.3593	33.2055	45.8405	59.8414	74.8664	93.9153	115.6757
CSSRC	**0.0102**	**0.0237**	**0.0408**	**0.0594**	**0.0818**	**0.1069**	**0.1328**	**0.1594**	**0.1907**	**0.2236**

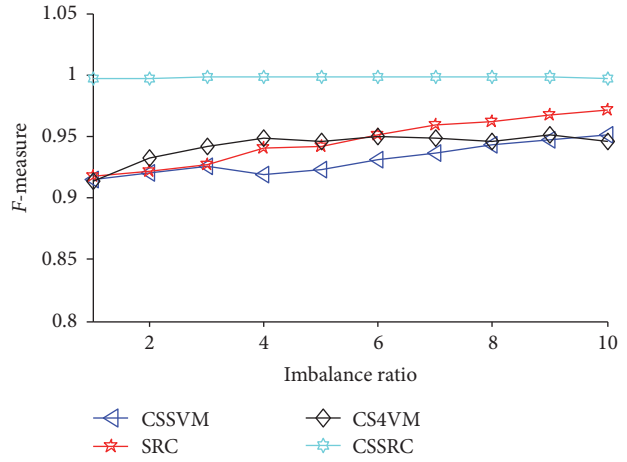


Figure 2: The result of F-measure on Nursery.

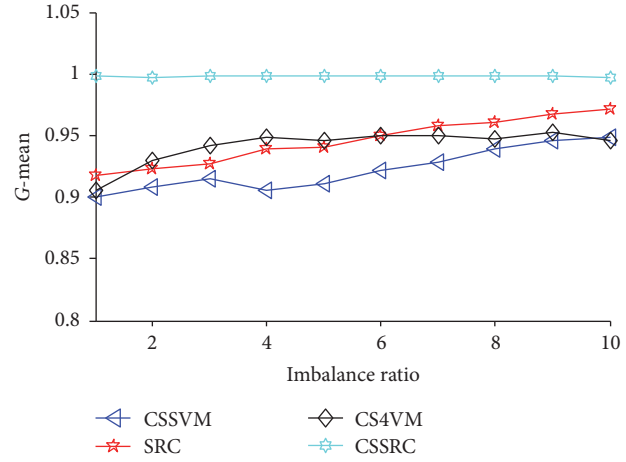


Figure 3: The result of G-mean on Nursery.

and Cost-Sensitive Semisupervised Support Vector Machine. Moreover, the method we proposed achieved a more stable performance with the increasing of imbalance ratio. Similarly, G-mean (the geometric mean between the classification accuracy of positive class and the classification accuracy of negative class) also achieved a higher value with respect the other three methods in Figure 3.

It is difficult to evaluate the performance of methods solving class-imbalance problem, but we use classification accuracy to reflect the method additionally for persuasive, and this is summarized in Table 4. On the other hand, running time represents the computation cost of a method. The result is shown in Table 5. It is obvious that our method can get the highest classification accuracy and the lowest running time on Nursery. In this paper, we use sparse representation coefficient vectors to estimate the posterior probability; this can well reduce the computing complexity and computation cost.

Experiment 2. In this experiment, we intend to validate the applicability of our method for class-imbalance problem. In Experiment 1 1 we have tested the validity of our method when the class distribution of training samples is imbalance. Now, we will select some training samples to validate our method, where the distribution of training samples is imbalance. Table 1 has summarized the information of data sets we used, and the imbalance ratio of Letter is 24.3. In order to set a higher imbalance ratio, we select Letter in this experiment. Similarly, we compare the performance of these four methods (SRC, CSSVM, CS4VM, and CSSRC) on Nursery. In this experiment, we take the imbalance ratio from $[1, 2, \ldots, 10]$, respectively. The size of minority class is 30 and the majority class is 30 multiplying the imbalance ratios in training set, accordingly. We select 61 positive samples and 61 negative samples as test set and run and summarize the results as Figures 4 and 5; the sampling process has repeated 100 times and gets the average results.

Figure 4 has shown the F-measure with imbalanced training samples; Figure 5 has shown the F-measure with imbalanced test samples. It is obvious that our method achieves a stable and higher result on Letter than the other three methods from Figures 4 and 5. Although sparse representation based classification has a similar result of F-measure with our method in Figure 5, the running time is

Table 6: Running time on Letter.

Time/s	1	2	3	4	5	6	7	8	9	10
SRC	0.0968	0.4103	0.6356	0.9178	1.2650	1.6695	2.1277	2.6364	3.3029	3.8405
CS4VM	2.0102	2.3504	2.6401	2.9405	3.2302	3.5415	3.8596	4.1482	5.4454	5.7497
CSSVM	7.5763	14.9397	22.1502	25.5690	37.0908	44.1674	51.6777	59.1158	66.6602	74.3515
CSSRC	**0.0308**	**0.1170**	**0.1914**	**0.2806**	**0.3954**	**1.3279**	**1.4672**	**1.6153**	**1.7826**	**1.9640**

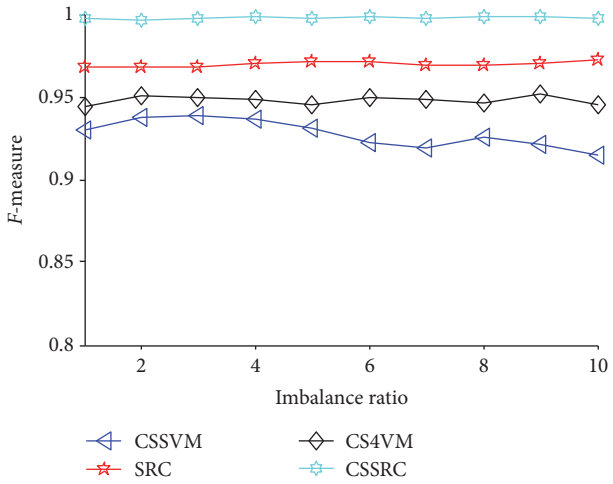


Figure 4: The result of F-measure on Letter (the distribution of training samples is imbalanced).

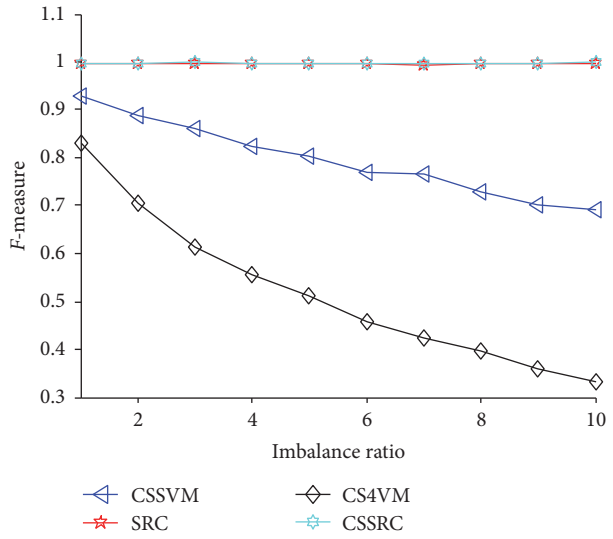


Figure 5: The result of F-measure on Letter (the distribution of test samples is imbalanced).

higher than our method in Table 6. Much experiments has been done in this section, we have compared F-measure with the imbalanced distribution of training samples and testing samples and running time, we can easily make a conclusion that our method is better than the other three methods, and we can well resolve the class-imbalance problem.

5. Conclusions and Future Works

This paper, we propose a novel cost-sensitive SRC classifier approach. The proposed approach adopted probabilistic model and sparse representation coefficient matrix to estimate the prior probability and then obtain the label of a testing sample by minimizing the misclassification losses. The experimental results show that the proposed cost-sensitive SRC has a comparable or even lower total cost with higher accuracy compare to the other three classification algorithms. Much experiment has been done and concluded that our method can well solve the class-imbalance problem. In real-world application, nearly all the data sets are class-imbalance. Our research can overcome the difficult the imbalanced distribution of data sets brought in.

In order to simplify the cost matrix, we restrict our discussion to two-class problems. So extending our current work to multiclass scenario is a main research direction for our future work.

Competing Interests

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant nos. 61562013 and 21365008), Natural Science Foundation of Guangxi (Grant no. 2013GXNSFBA019279), Innovation Project of GUET Graduate Education (no. YJCXS201558), and the Center for Collaborative Innovation in the Technology of IOT and the Industrialization (WLW20060610).

References

[1] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, 2008.

[2] J. Mairal, J. Ponce, G. Sapiro et al., "Supervised dictionary learning," in *Proceedings of the 21th Nation Conference on Advances in Neural Information Processing Systems*, vol. 1, pp. 1033–1040, Vancouver, Canada, December 2008.

[3] W. John, Y. A. Yang, G. Arvind et al., "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 44, no. 12, pp. 2368–2378, 2014.

[4] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Computer Vision—ECCV 2010: 11th European Conference on*

Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part VI, vol. 6316 of *Lecture Notes in Computer Science*, pp. 448–461, Springer, Berlin, Germany, 2010.

[5] E. Alpaydın, "Machine learning," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 3, pp. 195–203, 2011.

[6] Y. Li, J. T. Y. Kwok, and Z. H. Zhou, "Cost-sensitive semi-supervised support vector machine," in *Proceedings of the 24th National Conference on Artificial Intelligence*, vol. 1, pp. 500–505, Atlanta, Ga, USA, July 2010.

[7] Y. Zhang and Z.-H. Zhou, "Cost-sensitive face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1758–1769, 2010.

[8] X. Fang, "Inference-based naïve bayes: turning naïve bayes cost-sensitive," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 10, pp. 2302–2313, 2013.

[9] C. L. Castro and A. P. Braga, "Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 6, pp. 888–899, 2013.

[10] L. Jiang, C. Li, and S. Wang, "Cost-sensitive Bayesian network classifiers," *Pattern Recognition Letters*, vol. 45, pp. 211–216, 2014.

[11] H.-Y. Lo, S.-D. Lin, and H.-M. Wang, "Generalized k-labelsets ensemble for multi-label and cost-sensitive classification," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 7, pp. 1679–1691, 2014.

[12] J. Wan, M. Yang, and Y. Chen, "Discriminative cost sensitive laplacian score for face recognition," *Neurocomputing*, vol. 152, pp. 333–344, 2015.

[13] R. Pearson, G. Goney, and J. Shwaber, "Imbalanced clustering for microarray time-series," in *Proceedings of the Workshop on Learning from Imbalanced Dataset II (ICML '03)*, p. 3, Washington, DC, USA, 2003.

[14] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.

[15] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[16] L. Jiang, C. Qiu, and C. Li, "A novel minority cloning technique for cost-sensitive learning," *International Journal of Pattern Recognition & Artificial Intelligence*, vol. 29, no. 4, Article ID 1551004, 2015.

[17] J. Man, X. Jing, D. Zhang, and C. Lan, "Sparse cost-sensitive classifier with application to face recognition," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 1773–1776, IEEE, Brussels, Belgium, September 2011.

[18] J. Lu and Y.-P. Tan, "Cost-sensitive subspace learning for face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2661–2666, IEEE, San Francisco, Calif, USA, June 2010.

[19] J. Lu and Y.-P. Tan, "Cost-sensitive subspace analysis and extensions for face recognition," *IEEE Transactions on Information Forensics & Security*, vol. 8, no. 3, pp. 510–519, 2013.

[20] M. Z. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," in *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI '98)*, pp. 445–449, Brighton, UK, August 1998.

[21] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2368–2378, 2014.

[22] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[23] D. L. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," *Communications on Pure & Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[24] C. Blake, E. Keogh, and C. J. Merz, *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, Calif, USA, 1998.

[25] B. G. Hu and W. M. Dong, "A study on cost behaviors of binary classification measures in class-imbalanced problems," *Computer Science*, vol. 8, no. 11, Article ID e79774, 2014.

[26] Z. M. Kukar and I. Kononenko, "Cost-Sensitive Learning with Neural Networks," 445–449, 1998.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 5083074, 12 pages
http://dx.doi.org/10.1155/2016/5083074

Research Article

Monitoring Dangerous Goods in Container Yard Using the Internet of Things

Lianhong Ding, Yifan Chen, and Juntao Li

School of Information, Beijing Wuzi University, Beijing, China

Correspondence should be addressed to Lianhong Ding; lhdingbwu@sina.com

Received 6 August 2016; Revised 30 October 2016; Accepted 16 November 2016

Academic Editor: Xiong Luo

The Internet of Things (IoT), a network of objects, has been regarded as the next revolution for the global information industry after the Internet. With IoT, many intelligent applications can be accomplished or improved. This paper presents a framework for dangerous goods management in container yard using IoT technology. The framework consists of three layers: perceptual layer, transport layer, and application layer. It offers an infrastructure for management and data analysis and utilization. According to the features of dangerous goods, the framework can be enhanced for container information forecast, container gate-in and gate-out management, environment parameters monitoring, and fire control as well. In order to verify our method, a prototype system is developed, which shows good performance. With our method, safe operation of dangerous goods in container yard can be accomplished.

1. Introduction

The Internet of Things (IoT) is an emerging global Internet-based information architecture facilitating the exchange of goods and services in global supply chain networks [1]. IoT was first proposed in 1999 by Auto-ID Center [2]. The concept of IoT was widely accepted after a report from ITU released in 2005 [3]. The growth of IoT community has been encouraged by the rapid development of wireless sensor and actuator networks, identification tags such as barcode and RFID (Radio Frequency Identification), and electronic prototyping platforms such as Arduino [4].

IoT technology has been widely adopted in various fields, such as intelligent transportation and environmental protection. With the acceleration of the world economic integration, container transportation has become the most important transport mode for the world trade. Container yard is the container storage buffer in the whole operation chain for a port. The efficient and safe operation of the container yard will increase the port's relative capacity, thus improving the operation efficiency of the port. Regarding the importance of security, container yard is usually divided into two areas: the storage area for general goods and storage area for dangerous goods.

The logistics system of dangerous goods mainly consists of two parts: transportation and storage. At present, dangerous goods are mainly transported through the sea or road. The safety precautions and insurance policy for the shipping process are rather mature. Thus, dangerous goods can be protected well during their transportation. Therefore, the management and control in the storage process, especially in container yard, should be paid more attention.

A large proportion of serious accidents are caused by dangerous materials. A huge explosion happened in a container yard operated by a logistics company called Rui Hai International Logistics Co. Ltd. on 12 August 2015 in Tianjin, China. The Tianjin explosions were a series of explosions that killed over one hundred people and injured hundreds of others [5]. Rui Hai handles hazardous chemicals within the Port of Tianjin. In addition to the vast quantities of sodium cyanide and calcium carbide, 800 tons of ammonium nitrate and 500 tons of potassium nitrate were at the blast site [6]. A fire department spokesman confirmed that the firefighters had used water in combating the initial fire, which may have led to water being sprayed on calcium carbide, releasing the highly volatile gas acetylene. This may have detonated the ammonium nitrate [7]. The direct cause for this accident is that the nitrocellulose in containers spontaneously combusts

and explodes due to high temperature. This brings about the burning of the nitrocellulose and other hazardous chemicals in the adjacent containers and leads to the explosion of ammonium nitrate and other hazardous chemicals stacked in the yard [8].

To avoid similar accident, the monitoring of dangerous materials in container yard should be paid more attention. Obviously, IoT is a good solution for the monitoring system. This paper proposes a framework to manage dangerous goods in container yard. With the support of IoT, the framework helps operators obtain information about different dangerous materials, such as firefighting knowledge and temperature limitation. When emergency occurs, this system can provide related information to the firefighter. It will help the firefighters to use the right method to deal with emergencies, such as burning or explosion. With the help of wireless sensor networks, this system can also provide appropriate management strategies for relevant staffs through temperature monitoring. These strategies are generated according to the current temperature and the storage rules for dangerous materials. Similar directions can be given out according to the environment parameters such as humidity and CO_2 concentrations.

2. The Internet of Things

IoT is a novel paradigm that is rapidly gaining ground in the scenario of modern wireless telecommunications. The basic idea of IoT is the pervasive presence around us of a variety of things or objects. IoT involves many technologies including architecture, sensor/identification, coding, transmission, data processing, network, and discovery [9]. The standard, reliability, and robustness are also key concerns for IoT development.

With the changes of application requirements and the development of technologies, the concept of IoT is developed further [10, 11]. Different IoT definitions have been put forward from different perspectives such as CASAGRAS [12], CERP-IoT [13, 14], and Smart Planet [15]. Thiesse et al. found solutions based on RIFD technology or EPC mechanism [16]. Broll et al. proposed the Pervasive Service Interaction with things [17] and Vazquez et al. showed an integration solution between mobile services and smart objects [18]. Most researches focused on specific application or special function [19] such as security [20, 21], data mining model [22], and network management [23] for IoT. The Future Internet Assembly has been founded by the European Commission to support fundamental and systematic innovation in Europe for realization of the Future Internet [24].

2.1. Architecture of IoT. As a representative of the earlier scheme for IoT, EPC (Electronic Product Code) system is a vision world where all physical objects can be connected by RFID transponder through a global unique EPC code carried by the RFID tag [25]. Networked Auto-ID is an architecture proposed by the MIT AUTO-ID Center. Its target is to connect all objects by sensing devices (such as RFID and bar code) and the Internet. Corresponding architecture consists of physical tag (such as magnetic stripe encoding, barcode, two-dimensional code, and RFID), reader (such as magnetic stripe card reader, barcode reader, two-dimensional code reader, and RFID reader), network (the Internet or Intranet), object name service, and PML (Physical Markup Language) servers [26]. Japan also proposed its IoT prototype, uID IoT. It identifies real-world entities via RFIDs or barcodes, determines context information such as environment parameters from networked sensors, and adapts information services according to the data it obtains. The difference between Networked Auto-ID and uID IoT is that uID IoT collects context information such as environment parameters [27].

2.2. Technologies in IoT. IoT has several different implementation methods, such as RFID, GPS, laser sensor, infrared sensor, and other equipment. In this network, things can interact with each other without human's participation. In fact, the goal of IoT is to realize the automatic recognition and information sharing among things (or goods) through the Internet. Other associated technologies include network, database, and middleware.

RFID is a popular method to fulfill IoT. IoT can utilize RFID wireless communication to build a network of things [28]. RFID makes things "speak." The RFID tag stores rules and information [29]. There is a center system to collect the data from things through the wireless network. It recognizes objects and shares the information based on opening platform. Things can be managed by the center system.

Wireless sensor network is another key technology. Different kinds of sensors can collect context parameters according to application requirement. Generally, these parameters are transmitted by a wireless network such as GPRS. GPS technology and indoor location technique are often adopted by IoT as well.

3. Dangerous Goods Briefing

Dangerous goods are items or substances that may cause a risk to health, safety, property, or the public environment.

3.1. Classification and Identification of Dangerous Goods. The International Maritime Dangerous Goods (IMDG) code was developed as a uniform international code for the transport of dangerous goods by sea covering such matters as packing, container traffic, and stowage, with particular reference to the segregation of incompatible substances. Dangerous goods are classified into different classes according to IMDG code. General provisions for each class or division are given. Individual dangerous goods are listed in the Dangerous Goods List, with the class and any specific requirements.

In general, dangerous goods are classified into 9 classes. Each class is expressed by a single number, such as "class 1." The nine classes of dangerous goods are as follows: class 1, explosives; class 2, compressed gases and liquefied gases; class 3, flammable liquids; class 4, flammable solids; class 5, oxidizing substances and organic peroxide; class 6, toxic and

Table 1: General segregation requirements for hazardous materials.

Class	1.1, 1.2, 1.5	1.3, 1.6	1.4	2.1	2.2	2.3	3	4.1	4.2	4.3	5.1	5.2	6.1	6.2	7	8	9
Explosives 1.1, 1.2, 1.5	*	*	*	4	2	2	4	4	4	4	4	4	2	4	2	4	×
Explosives 1.3, 1.6	*	*	*	4	2	2	4	3	3	4	4	4	2	4	2	4	×
Explosives 1.4	*	*	*	2	1	1	2	2	2	2	2	2	×	4	2	2	×
Flammable gases 2.1	4	4	2	×	×	×	2	1	2	×	2	2	×	4	2	1	×
Nontoxic, nonflammable gases 2.2	2	2	1	×	×	×	1	×	1	×	×	1	×	2	1	×	×
Poisonous gases 2.3	2	2	1	×	×	×	2	×	2	×	×	2	×	2	1	×	×
Flammable liquids 3	4	4	2	2	1	2	×	×	2	1	2	2	×	3	2	×	×
Flammable solids 4.1	4	3	2	1	×	×	×	×	1	×	1	2	×	3	2	1	×
Spontaneously combustible substances 4.2	4	4	3	2	2	1	2	2	1	×	1	2	2	1	3	2	1
Substances which are dangerous when wet 4.3	4	4	2	×	×	×	1	×	1	×	2	2	×	2	2	1	×
Oxidizing substances 5.1	4	4	2	2	×	×	2	1	2	2	×	2	1	3	1	2	×
Organic peroxides 5.2	4	4	2	2	1	2	2	2	2	2	2	×	1	3	2	2	×
Poisons 6.1	2	2	×	×	×	×	×	×	1	×	1	1	×	1	×	×	×
Infectious substances 6.2	4	4	4	4	2	2	3	3	3	2	3	3	1	×	3	3	×
Radioactive materials 7	2	2	2	2	1	1	2	2	2	2	1	2	×	3	×	2	×
Corrosives 8	4	2	2	1	×	×	×	1	1	1	2	2	×	3	2	×	×
Miscellaneous dangerous substances 9	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

infectious substances; class 7, radioactive materials; class 8, corrosives; and class 9, miscellaneous dangerous substances.

Some classes, classes 1, 2, 4, 5, and 6, are subdivided into divisions. Divisions are expressed by 2 numbers. The first number identifies the class number and the second identifies the variation within that class. For example, oxidizer is class 5, division 1, which should read "division 5.1." The order in which the classes are numbered is for convenience and does not imply a relative degree of danger (i.e., class 1 is not necessarily more dangerous than class 2 or 3).

3.2. Introduction of Dangerous Goods Regulation. There are different demands for different kinds of dangerous materials. The Department of Transportation (DOT) has designated criteria in 49CFR: Title 49—Transportation, Code of Federal Regulations, Hazardous Materials Regulations (HMR), for determining what is considered hazardous for transportation. Definitions of hazardous material designations are included in 49CFR173 and DOT Hazardous Materials Table 49CFR172.101 provides descriptions required for shipping. Here, several rules are considered in our monitoring system:

(1) During the high temperature period (from June 20 to September 10 each year) from 10 a.m. to 4 p.m., it is prohibited to transport divisions 2.2, 3.1, 3.2, 4.2, 4.3, and 5.1.

(2) In the high temperature season, when the temperature exceeds 30 degrees Celsius, the dangerous goods container for spraying should be sprayed every 2 hours.

(3) According to JT 397-2007: the safety rules for handling dangerous cargo container in port, classes 1, 2, and 7 should be directly lifted down.

(4) Different kinds of dangerous goods should be stacked in different areas. Some of them should be maintained within a certain distance. The detailed requirements are given in Table 1. Numbers and symbols in Table 1 are related to the following terms:

(i) 1, away from: effectively segregated so that the incompatible materials cannot interact dangerously in the event of an accident but may be carried in the same compartment or hold or on deck provided that minimum horizontal separation of 3 m (10 feet) projected vertically is obtained.

(ii) 2, separated from: in different compartments or holds when stowed under deck. If the intervening deck is resistant to fire and liquid, vertical separation (i.e., in different compartments) is acceptable as equivalent to this segregation. For "on-deck" stowage, this segregation means separation by a distance of at least 6 m (20 feet) horizontally.

(iii) 3, separated by a complete compartment or hold from: either vertical or horizontal separation. If the intervening decks are not resistant to fire and liquid, then only longitudinal separation (i.e., by an intervening complete compartment or hold) is acceptable. For "on-deck" stowage, this segregation means separation by a distance of at least 12 m (39 feet) horizontally. The same distance must be applied if one package is

stowed "on deck" and the other one in an upper compartment.

(iv) 4, separated longitudinally by an intervening complete compartment or hold from: vertical separation alone does not meet this requirement. Between a package "under deck" and one "on deck," a minimum distance of 24 m (79 feet) including a complete compartment must be maintained longitudinally. For "on-deck" stowage, this segregation means separation by a distance of at least 24 m (79 feet) longitudinally.

(v) $\times$, the segregation, if any, is shown in detail in table of materials.

(vi) $*$, segregation among different class 1 (explosive) materials is governed by the compatibility table.

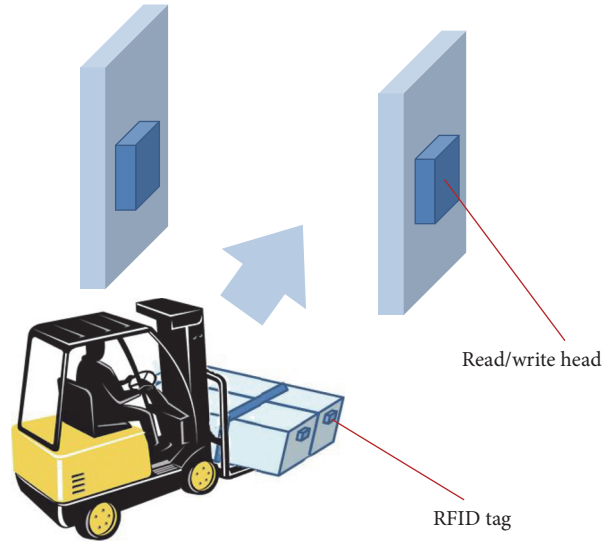


Figure 1: Gate-in and gate-out procedure with RFID.

4. Monitoring System for Dangerous Goods Using IoT

4.1. Requirement Analysis. Considering the regulations for dangerous materials introduced in Section 3.2, the monitoring system should have the following functions:

(1) *Container Information Forecast.* Owners can notice the yard and the information of arriving container in advance through the Internet or other channels. This kind of information can be container number, arrival time, source and destination, and other related information. The management systems can make work plan according to the information.

(2) *Gate-In and Gate-Out Management.* As shown in Figure 1, in the container yard entrance and exit, the reader equipment can get the container information by RFID tag on the container and upload the related information to the system database. The management system automatically records the gate-in information or gate-out information and updates the database in real time.

(3) *Environment Parameters Monitor.* This can be realized by a temperature sensor installed in the container yard. Furthermore, if sensors are deployed within containers, internal environment parameters of the container can be monitored. Different dangerous materials have different ignition points and explosive limits. Sensors inside containers make specific context-aware information for different dangerous materials possible. To a certain extent, this function depends on the deployment of the wireless sensor network. It will be introduced in Section 4.6.

(4) *Firefighting Auxiliary Function.* Firefighting methods for different dangerous materials may be different. For example, if there is a fire of sodium azide, water, foam, and carbon dioxide can be used but sand pressure cannot be used. In order to get right firefighting knowledge for different dangerous materials, two situations should be discussed. First, if the RFID tag on the container still works, the firefighter can get main information about the dangerous goods packaged in the container, including advice about firefighting, by a handhold reader directly. Second, if the tag has been destroyed, fire alert can still be obtained through the container's position information indirectly. The details will be introduced in Section 4.5.

4.2. System Architecture. In order to support the above functions, we put forward the following architecture for the monitoring system using IoT. The architecture consists of three layers, shown in Figure 2. The first layer is perceptual layer. The task of the perceptual layer is to identify physical object and collect context such as humidity and position. This layer includes RFID tag, RFID reader, sensors, GPS receivers, and handhold terminal. The second layer is transport layer. It transports information by the Internet, Intranet, or wireless network such as GPRS. The object resolution server and container yard monitoring server for dangerous goods form the third layer. The third layer can be regarded as an application layer. The object resolution server identifies the entity by the unique code and finds out the monitoring services related to the object. Container yard monitoring server gives specific context-aware information based on the information provided by object resolution server and with reference to the service rule database.

Unique code tag only records a unique code. All information about objects and position is maintained by object resolution server and container yard monitoring server. By separating the unique code and information, users can easily acquire the latest information on an entity, update that information, and obtain information on other entities related to that entity.

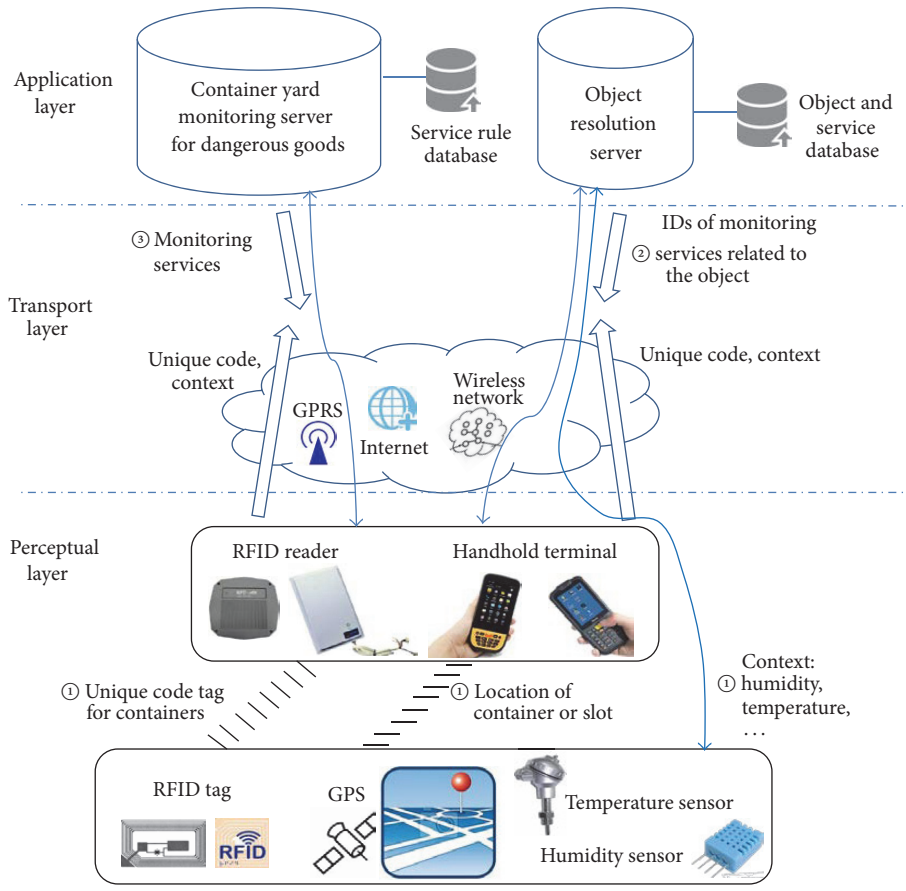


Figure 2: Architecture of the monitoring system for dangerous goods in container yard.

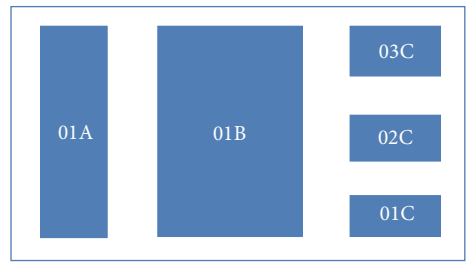


Figure 3: Exemplary layout of the container yard.

The numbers in Figure 2 indicate the order in which the information is processed:

① Obtain unique code of container from RFID tags on container by RFID reader (handhold terminal or reader head on entrance) or location information about container by handhold terminal of GPS. In this step, contexts such as humidity and temperature in containers are collected through sensors and wireless network.

② Object resolution server retrieves the context information about the object identified by unique code and the monitoring services related to object. Here, object and service database records the related information. This kind of information includes the class code of the dangerous materials in the container, the transport rule, and the ID of monitoring services related.

③ When container yard monitoring server receives monitoring service ID, it sends context-aware information about the object (such as container or dangerous goods) back to operators through handhold terminals. The details for the context-aware information are stored in the service rule database.

4.3. Layout of Container Yard and the Encoding Method for Slot Number. Figure 3 shows the overall layout of the container yard for dangerous goods. Different blocks are arranged

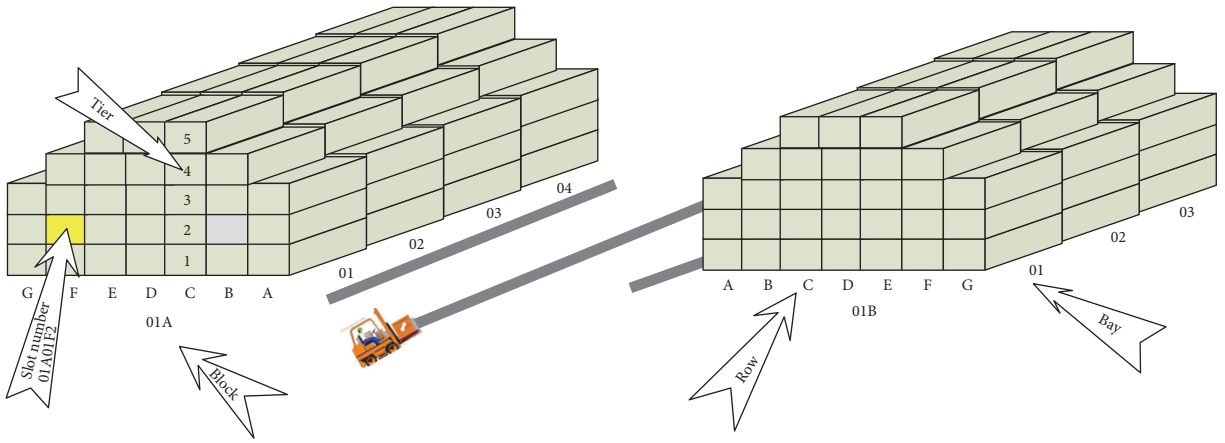


Figure 4: Layout of the container yard and slot number.

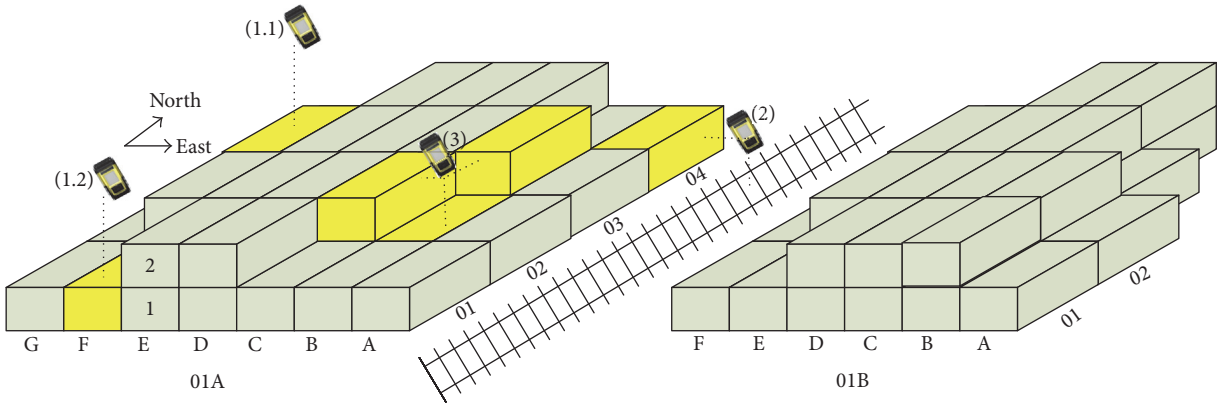


Figure 5: Identification of slot by location information.

in the rows and columns. Each block is distinguished by two numbers and one letter. Two numbers indicate the order of a block in the yard from the west to the east. The letter specifies the block order from the south to the north. According to the regulations in the recommendations on the transport of dangerous goods, different kinds of dangerous goods should be stacked in different blocks [30].

Figure 4 is an exemplary layout of a container storage facility, at either seaport terminal, rail intermodal terminal, or inland storage terminal. It illustrates the three-dimensional characteristics of the storage locations at a container yard. Each individual cube represents a container storage location, called slot, where a container can reside. Slot number is used to uniquely identify each slot in the container storage facility. Figure 4 shows a typical slot number, which uses terms such as bay, row, and tier following the block number. A bay value and a row value are used to uniquely identify a container storage location's planetary position in a block. Each bay has the width of one container's length and each row has the width of one container's depth. Containers can also be stacked on top of one another, and the height of the container storage location is represented by a tier value.

In Figure 4, the height is shown and the container cell on the second tier of location (Block 01A, Bay 01, Row F) is uniquely identified by (Block 01A, Bay 01, Row F, Tier 2). So, the slot number is 01A01F2. Such a cell-naming convention allows quick and easy identification of a storage location for containers as well as numerous other types of inventory. Other naming conventions can also be used, and they all reflect uniformity throughout the storage facility in representing the 3-dimensional storage cell locations.

4.4. Gate-In and Gate-Out Operation. As illustrated in Figure 5, each container owns a unique number written in its RFID tag. The container number follows the ISO 6346(1995): freight containers coding identification and marking. It provides a system for general application for the identification and presentation of information about freight containers. It specifies an identification system with mandatory marks for visual interpretation and optional features for automatic identification and electronic data interchange and a coding system for data on container size and type. The class number introduced in Section 3.1 is stored in the tag as well.

Container yard supporting facilities include straddle carriers, shore crane, tire gantry stackers, and forklifts (or

stacker). The storage height for dangerous goods container is generally no more than 2 tiers and safety distance should be maintained between different types of dangerous goods containers [30]. These regulations make tire gantry stacker unable to fully play its advantage. At the same time, the flameproof transformation for tire gantry stackers is expensive. Forklift truck for freight containers is more suitable for container yard of dangerous goods.

As shown in Figure 1, RFID read/write heads are mounted on the entrance and exit of the container yard. When a forklift truck, carrying a container, passes through the entrance, the read/write head reads the information from the tag on the container and updates the related information in the object and service database. Similar operations are also performed when a forklift truck passes through the exit.

The electronic tag in RFID system can be active or passive according to whether it has a built-in power supply. Passive RFID tags are used for applications such as access control, file tracking, supply chain management, and smart labels. The lower price makes employing passive RFID systems economical for many enterprises. Passive tags are adopted in our system as well.

Various frequency bands can be used in a RFID system such as low frequency (125 KHz, 135 KHz), high frequency (13.56 MHz), ultrahigh frequency (400 MHz–960 MHz), and micro wave (2.45 GHz). According to the demands of transmission distance and speed, 433 MHz, 916 MHz, and 2.45 GHz are often used for container management. In China, 433 MHz can be used by radio amateur. At the same time, 860 MHz–960 MHz is the band belonging to GSM in China. So, radio signal whose frequency is 2.45 GHz is used in our system.

4.5. Firefighting Auxiliary Based on GPS Information. As introduced in Section 4.1, there are two ways to provide right firefighting knowledge according to the type of dangerous materials for fire crews. Firefighters can scan RFID tag on container by handhold terminal, and then the unique code of the container is transmitted to the object resolution server by transmit layer. The object resolution server identifies the container by unique code and extracts related information. The information includes the United Nations code (UN code) of the dangerous goods stored in the container and the monitoring service ID labeling the query service for the knowledge about firefighting. The information is submitted to the container yard monitoring server. The monitoring server queries the service rule database according to UN code and monitoring service ID. Then, firefighting knowledge is sent back to firefighter through handhold terminal by monitoring server.

If the tag on the container is destroyed, firefighter can report the location information of a slot to the object resolution server by the GPS function of the handhold terminal. The object resolution server deduces the slot number according to the location information first. The container stacked at that slot is identified further. The rest of the process is the same as the first method.

Table 2: Dangerous goods classified by hazardous properties.

Hazard properties	Code of division
Explosive	1.1, 1.4, 1.5, 1.6, 2.1, 5.1, 5.2
Flammable	Except for nonflammable goods
Toxic	2.3, 3, 6.1, 6.2, 9
Radioactive	7
Corrosive	8

The key issue for the second method is the way to deduce the slot number according to the location information. Location information about each slot is maintained in the object and service database. As a result, we can get a container's location range by slot number query and get slot number by location query. When the container is stored to a slot, the bond between the container unique code and slot number is built and recorded in the object and service database. So, we can get information about the dangerous goods in a container either by container number or by slot number.

Here, we only discuss the closed container. When we want to gain information about a certain container, we cannot put the handhold terminal into the container. The terminal can only be located near the container. As shown in Figure 5, there are three conditions for the relationship between the container and the handhold terminal:

(1) The height of the terminal is bigger than 2 tiers. According to the regulations in the recommendations on the transport of dangerous goods, height for containers of dangerous goods is no more than 2 tiers [30]. For condition (1.1), the container with slot number 01A03F02 is identified. For situation (1.2), because there is no container in tier 2, the slot number is 01A01F01.

(2) The horizontal coordinates of the terminal are out of the range of block. The container with slot number 01A04A01 is specified.

(3) For the third situation, the distances from the terminal to the containers with slot numbers 01A02C02, 01A03B02, and 01A02B01 are calculated, respectively. Suppose the distance to slot 01A02C02 is the shortest; the slot number 01A02C02 will be specified.

The deviation of open GPS supplied by USA is less than 10 meters. Researchers and manufacturers often improve the precision of GPS by difference algorithm. The application of carrier phase difference in GPS can bring about precision resolution in centimeters. The external dimensions of the most common container, 40GP, are $12192 \times 2438 \times 2591$ (mm). So, the resolution of GPS enhanced by difference algorithm can satisfy the location requirement in the container yard.

4.6. Environment Parameter Monitor and Control. Table 2 illustrates the hazardous goods classified by hazardous properties. According to the classification of dangerous goods introduced in Section 3.1 and Table 2, we can find out that explosion and flame are the main hazards of dangerous goods.

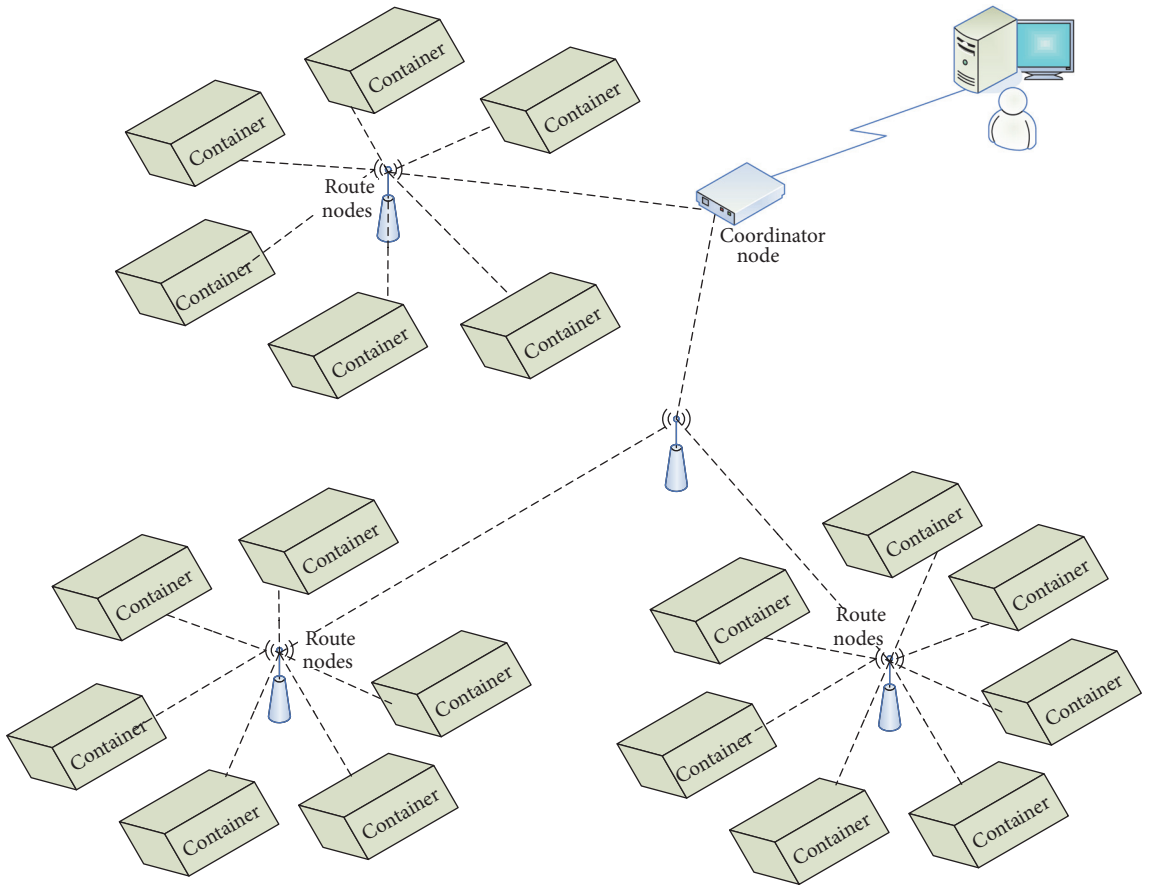


Figure 6: Schematic diagram of WSN for environment monitoring.

At the same time, there are many kinds of dangerous materials whose hazards are explosion and inflammation. Therefore, we emphasize the monitoring of explosion and inflammation in the container yard for dangerous materials. Generally, an explosion or flame needs three essential factors: combustibility, an oxidant, and temperature (ignition source). In extreme circumstances, the temperatures in containers may reach up to 50 Celsius or 70 Celsius. So the temperature is the key issue in these three elements.

As discussed before, temperature monitor and control are important for dangerous goods. This can be achieved with the support of wireless sensor networks. In order to get real-time temperature, temperature sensors should be mounted. There are two modes for the installation of temperature sensors: one temperature sensor is mounted in each container or several temperature sensors are installed at different places in the container yard. The first method needs the modification of container. It requires a large number of sensors. This method has an obvious advantage. Containers holding different kinds of dangerous goods can be monitored, respectively, according to the specific temperature requirements. This method can be enhanced to fulfill the control of humidity and CO_2 concentrations in dangerous goods containers. In order to reduce the number of sensors and simplify the installation process, compound sensor can be designed to collect information about temperature, humidity, and CO_2 concentrations.

Senor is the most important kind of monitoring node. For the first mode, monitoring node is mounted inside the container. Because the container itself has a strong shielding effect on the wireless signal, we fix antennas on the outer surface of the container door. A tree network is adopted in our system. As shown in Figure 6, the monitoring system consists of wireless sensor network and computer monitoring center. Except for sensors, there are two other kinds of monitoring nodes in the wireless sensor network. The coordinator node is responsible for the communication between the WSN and the computer monitoring center. It is the assignment for route nodes to transfer data and enlarge the coverage areas of the WSN. The monitoring nodes (sensors) collect related data, such as temperature and humidity, inside the container.

For the second way, temperature sensors are only deployed in container yard, so the outdoor temperature can be monitored with low cost. The monitoring system can also give some context-aware information. For example, when the outdoor temperature exceeds 30 degrees Celsius, it will notice the operators to spray the dangerous goods container every 2 hours.

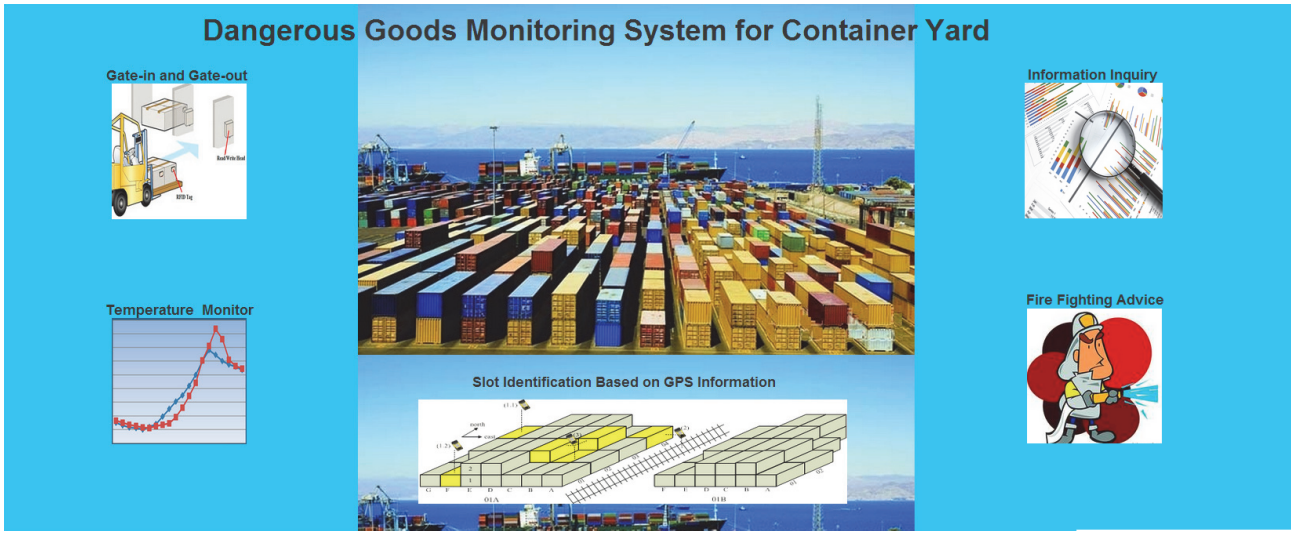


Figure 7: Main page of the prototype system.

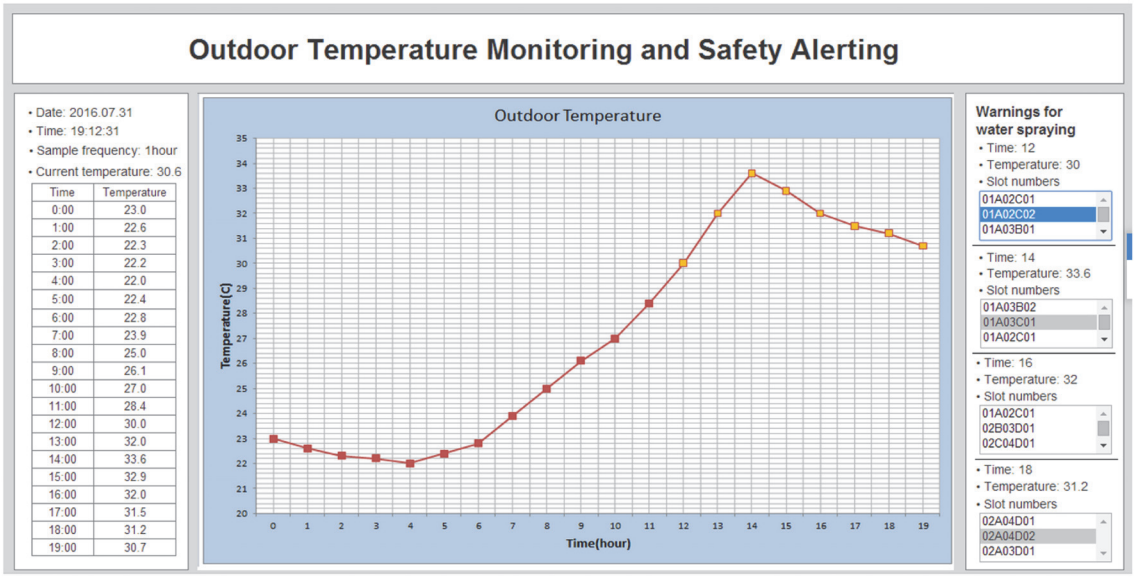


Figure 8: Function of environment parameter monitoring.

5. Prototype System

To verify the effect of our method, a prototype system is developed. Figure 7 shows the main page of the prototype system. The main functions including gate-in and gate-out management, temperature monitoring, firefighting auxiliary function, and information inquiry are implemented.

Temperature monitoring function is realized by the prototype system as the sample of environment parameters monitoring. As shown in Figure 8, outdoor temperature is collected each hour. At any time, the temperature can be browsed in table format or temperature curve. If the outdoor temperature is greater than or equal to 30 degrees Celsius, the corresponding dot in the temperature curve is painted in yellow. The right of Figure 8 lists the reminding information for spray. The reminding information is given following the rule introduced in Section 3 and is updated every 2 hours. The slot number of the containers that should be sprayed is listed as well.

Figure 9 is SHT11 intelligent sensor from Sensirion Company. It collects temperature and humidity in containers. It measures temperature with a resolution of 0.01 degrees and within ±2-degree accuracy and measures relative humidity with a resolution of 0.03% and within ±3.5% accuracy. The operating temperature range of SHT11 is from −40 to +125°C. The 2-wire serial interface and internal voltage regulation allow easy and fast system integration [31]. If electrochemical sensors are used to monitor concentration of gases such

Table 3: Field meaning of GPGGA message.

GGA	Global Positioning System Fix Data
123519	Fix taken at 12:35:19 UTC
4807.038,N	Latitude 48 deg. 07.038′ N
01131.000,E	Longitude 11 deg. 31.000′ E
1	Fix quality: 0 = invalid; 1 = GPS fix (SPS); 2 = DGPS fix; 3 = PPS fix; 4 = real-time kinematic; 5 = float RTK; 6 = estimated (dead reckoning); 7 = manual input mode; 8 = simulation mode
08	Number of satellites being tracked
0.9	Horizontal dilution of position
545.4,M	Altitude, meters, above mean sea level
46.9,M	Height of geoid (mean sea level) above WGS84 ellipsoid
(empty field)	Time in seconds since the last DGPS update
(empty field)	DGPS station ID number
*47	Checksum data, always beginning with*

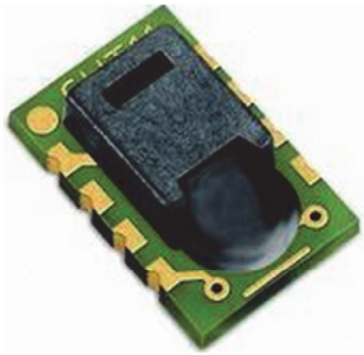


Figure 9: SHT11, temperature and humidity sensor.

as nitric oxide and hydrogen sulfide, A/D modular should be designed to convert sensors' analogue signals to digital signals.

As introduced before, the right firefighting knowledge can be obtained by RFID tag on container or location information of container. Figure 10 shows the firefighting information got by location information through user terminal based on GPS function. The left part of Figure 10 is the location information, the slot number of the container identified according to the location information, and the firefighting principles for the dangerous goods in the container identified.

NMEA 0183 is a combined electrical and data specification for communication between GPS receivers. It has been defined and controlled by the National Marine Electronics Association. The baud rate supported by NMEA 0183 is 4800 bps, and all messages of NMEA-0183 are ASCII codes. Each message begins with a dollar sign ($) and ends with a carriage return and a linefeed (<CR><LF>).

Commonly used messages include GPGGA (Global Positioning System Fix Data), GPGSA (GPS DOP and active satellite), GPGSV (GPS Satellites in View), GPRMC (Recommended Minimum Specific DPS/Transit Data Speed), GPVTG (Track Made Good and Ground), and GPGLL (Geographic Position, Latitude/Longitude). GPGGA message can provide us with information such as time and position and fix related data. "$GPGGA,123519,4807.038,N, 01131.000,E,1,08,0.9,545.4,M,46.9,M,,*47" is a GPGGA message. The meaning of each field is listed in Table 3.

The right part of Figure 10 lists the detailed information about the dangerous goods in the corresponding container. The information includes name, molecular formula, CAS number, UN number, type of hazard, and method of prevention and firefighting. It would help people to completely know the information about dangerous goods in the specific container.

6. Conclusion and Future Works

This paper proposes a framework for container yard of dangerous goods using IoT, monitoring environment parameters, and providing fire control service. It can reduce the risk of fire and explosion. When fire or explosion happens, firefighter can get right directions for firefighting through the fire control service. A prototype system is developed as well. Main functions of our method, such as gate-in and gate-out management, temperature monitoring, information inquiry, and firefighting auxiliary, are accomplished.

In the future, we will focus on the study of the storage assignment strategy for the dangerous goods in container yard. The assignment strategy must satisfy the regulations in recommendations on the transport of dangerous goods first of all [30]. We will accomplish the deployment of the system and investigate the quantitative benefits brought about by the system in our future work.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

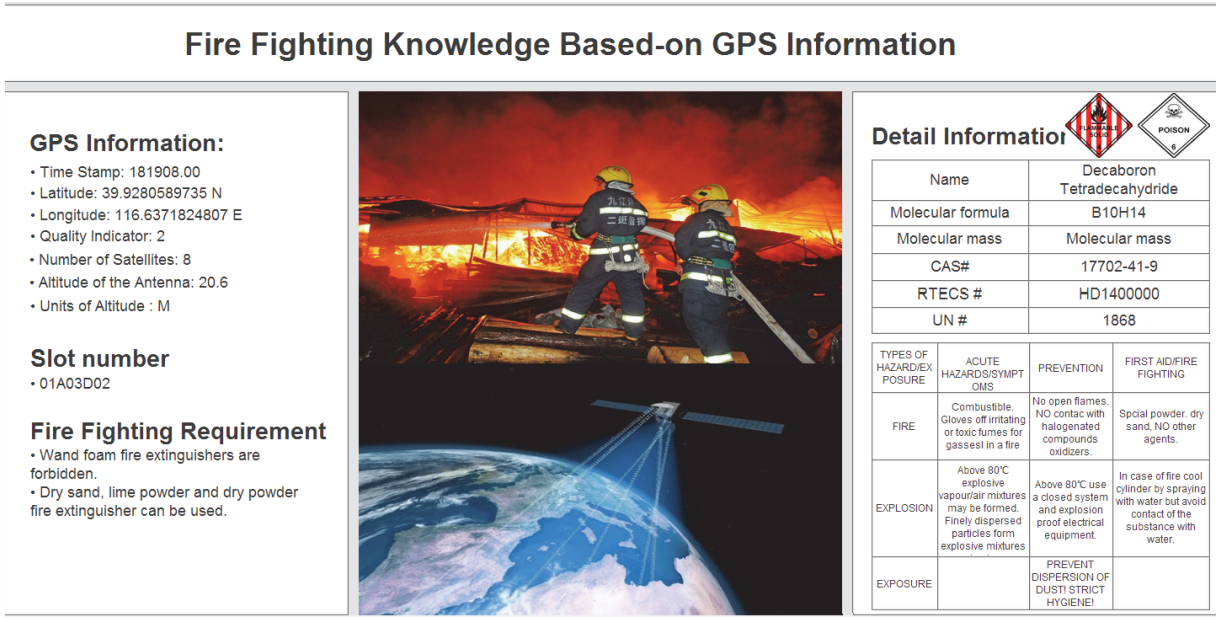


Figure 10: Firefighting knowledge obtained through location information.

Acknowledgments

This work was supported by Beijing Key Laboratory (no. BZ0211), Beijing Intelligent Logistics System Collaborative Innovation Center, and Breeding Project of BWU (Research of Logistics for Low Carbon Economy, no. GJB20162002).

References

[1] R. H. Weber, "Internet of Things—new security and privacy challenges," *Computer Law & Security Review*, vol. 26, no. 1, pp. 23–30, 2010.

[2] H. Ning and Z. Wang, "Future internet of things architecture: like mankind neural system or social organization framework?" *IEEE Communications Letters*, vol. 15, no. 4, pp. 461–463, 2011.

[3] "ITU internet report 2005: the internet of things," Tech. Rep., International Telecommunication Union, 2005.

[4] F. Pramudianto, I. R. Indra, and M. Jarke, "Model driven development for internet of things application prototyping," in *Proceedings of the International Conference on Software Engineering and Knowledge Engineering*, Boston, Mass, USA, June 2013.

[5] Wikipedia, "2015 Tianjin explosions," 2015, https://en.wikipedia.org/wiki/2015_Tianjin_explosions.

[6] R. Fergus, "Tianjin explosions: warehouse 'handled toxic chemicals without licence'—reports," The Guardian, https://www.theguardian.com/world/2015/aug/18/tianjin-blasts-warehouse-handled-toxic-chemicals-without-licence-reports.

[7] R. Iyengar, "Searching questions asked in the aftermath of the Tianjin blasts," Time, August 2015.

[8] "Report on the investigation of the explosion accident in Tianjin port," Phoenix Information, 2016.

[9] J. P. Conti, "The internet of things," *IET Communications Engineer*, vol. 4, no. 6, pp. 20–25, 2006.

[10] ITU, "The Internet of Things," ITU International Reports, 2005.

[11] INFSO D.4 Networked Enterprise & RFID INFSO G.2 Micro & Nanosystems Groups in Co-Operation with the RFID Working Group of the EPoSS, "Internet of Things in 2020", 2008.

[12] Coordination and Support Action for Global RFID-Related Activities and Standardization (CASAGRS), "RFID and the inclusive model for the internet of things," CASAGRAS Final Report, 2009.

[13] Cluster of European Research Projects on the Internet of Things (CERPIoT), *CERP-IoT Research Roadmap*, 2009.

[14] Cluster of European Research Projects on the Internet of Things (CERPIoT), *Vision and Challenges for Realising the Internet of things*, 2010.

[15] IBM Institute for Business Value, "A Smart Planet," 2009, http://www.ibm.com/smarterplanet/us/en.

[16] F. Thiesse, C. Floerkemeier, M. Harrison, F. Michahelles, and C. Roduner, "Technology, standards, and real-world deployments of the EPC network," *IEEE Internet Computing*, vol. 13, no. 2, pp. 36–43, 2009.

[17] G. Broll, M. Paolucci, M. Wagner, E. Rukzio, A. Schmidt, and H. Hussmann, "Perci: pervasive service interaction with the internet of things," *IEEE Internet Computing*, vol. 13, no. 6, pp. 74–81, 2009.

[18] J. I. Vazquez, J. Ruiz-De-Garibay, X. Eguiluz, I. Doamo, S. Rentería, and A. Ayerbe, "Communication architectures and experiences for web-connected physical smart objects," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops '10)*, pp. 684–689, Mannheim, Germany, April 2010.

[19] L. Yan, Y. Zhang, L. T. Yang, and H. S. Ning, "The internet of things: from RFID to the next-generation pervasive networked systems," *Journal of Vertebrate Paleontology*, vol. 15, no. 2, pp. 431–442, 2010.

[20] É. Renault, A. Ahmad, and M. Abid, "Toward a security model for the future network of information," in *Proceedings of the 4th*

International Conference on Ubiquitous Information Technologies and Applications (ICUT '09), pp. 1–6, IEEE, Fukuoka, Japan, December 2009.

[21] R. Roman, P. Najera, and J. Lopez, "Securing the Internet of things," *Computer*, vol. 44, no. 9, pp. 51–58, 2011.

[22] 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.

[23] H. Ning, N. Ning, S. Qu, Y. Zhang, and H. Yang, "Layered structure and management in internet of things," in *Proceedings of the International Conference on Future Generation Communication and Networking (FGCN '07)*, vol. 2, pp. 386–389, Jeju Island, Korea, December 2007.

[24] Future Internet Assembly, "Towards the future Internet: emerging trends from European research," http://www.future-internet.eu/home/future-internet-assembly.html.

[25] EPCglobal, "The EPCglobal Architecture Framework," final version 1.3, 2009.

[26] S. Sarma, D. L. Brock, and K. Ashton, "The networked physical world: proposals for engineering the next generation of computing, commerce & automatic-identification," White Paper MIT-AUTOID-WH-001, MIT Auto-ID Center, 2010.

[27] N. Koshizuka and K. Sakamura, "Ubiquitous ID: standards for ubiquitous computing and the internet of things," *IEEE Pervasive Computing*, vol. 9, no. 4, pp. 98–101, 2010.

[28] E. Welbourne, L. Battle, G. Cole et al., "Building the internet of things using RFID: the RFID ecosystem experience," *IEEE Internet Computing*, vol. 13, no. 3, pp. 48–55, 2009.

[29] R. Want, "An introduction to RFID technology," *IEEE Pervasive Computing*, vol. 5, no. 1, pp. 25–33, 2006.

[30] ONU, "Recommendations on the transport of dangerous goods: model regulations," 2005.

[31] SHT71 datasheets and info, https://www.sensirion.com/en/products/humidity-sensors/digital-humidity-sensors-for-accurate-measurements/.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 7427980, 15 pages
http://dx.doi.org/10.1155/2016/7427980

Research Article

Stationary Hand Gesture Authentication Using Edit Distance on Finger Pointing Direction Interval

Alex Ming Hui Wong[1] and Dae-Ki Kang[2]

[1]*Department of Ubiquitous IT, Graduate School, Dongseo University, 47 Jurye-Ro, Sasang-Gu, Busan 47011, Republic of Korea*
[2]*Department of Computer & Information Engineering, Dongseo University, 47 Jurye-Ro, Sasang-Gu, Busan 47011, Republic of Korea*

Correspondence should be addressed to Dae-Ki Kang; dkkang@dongseo.ac.kr

Received 12 July 2016; Revised 28 September 2016; Accepted 17 October 2016

Academic Editor: HuaPing Liu

One of the latest authentication methods is by discerning human gestures. Previous research has shown that different people can develop distinct gesture behaviours even when executing the same gesture. Hand gesture is one of the most commonly used gestures in both communication and authentication research since it requires less room to perform as compared to other bodily gestures. There are different types of hand gesture and they have been researched by many researchers, but stationary hand gesture has yet to be thoroughly explored. There are a number of disadvantages and flaws in general hand gesture authentication such as reliability, usability, and computational cost. Although stationary hand gesture is not able to solve all these problems, it still provides more benefits and advantages over other hand gesture authentication methods, such as making gesture into a motion flow instead of trivial image capturing, and requires less room to perform, less vision cue needed during performance, and so forth. In this paper, we introduce stationary hand gesture authentication by implementing edit distance on finger pointing direction interval (ED-FPDI) from hand gesture to model behaviour-based authentication system. The accuracy rate of the proposed ED-FPDI shows promising results.

1. Introduction

Hand gesture recognition has been adapted into some of our daily-use home appliances [1, 2], in electronic devices [3–5], and even in vehicles [6] as a method of input and it is said to be the interface of Internet-of-Things (IoT) in the future [7]. But with such a wide usage and implementation, security issues are bound to emerge. Due to these issues, hand gesture passwords have been researched since the past few years [8–10]. But hand gesture password is not sufficiently secure as hand gesture can be easily learned and imitated. In order to overcome this limitation, we have adopted behaviour into the hand gesture, in return enhancing hand gesture password into hand gesture authentication.

It is interesting that hand gesture recognition has been researched throughout the decades, but there has not been any clear definition or classification on it. In this paper, we classified hand gesture into two major categories: static hand gesture recognition and motion hand gesture recognition. Static hand gesture recognition is performed with captured images; it can be using only one image or more than one. Motion hand gesture recognition includes all the movement between different hand gestures; the recognition process is conducted over a flow of hand gesture from the beginning until the end of the gesture. That is, static hand gesture is analogous to still image, whereas motion hand gesture is to video. In motion hand gesture recognition, there are other subcategories such as dynamic hand gesture recognition, stationary hand gesture recognition, finger gesture recognition [1, 11], and arm gesture recognition [12–14]. Detailed information on hand gesture recognition is explained in Section 3.1.

User authentication can be implemented through various methods including biometric feature based ones [15]. There are two main approaches in biometric authentication: physiology-based and behaviour-based. Physiology-based approach includes fingerprint, iris, and voice; whereas behaviour-based or behaviometrics [16] includes human

gesture, keystroke dynamic on keyboard, mouse, and touch screen [17]. Based on [18], different people develop different behaviour even when performing the same gesture or movement. These behaviours can be evaluated through movement speed, acceleration, positioning, distance travelled, and so forth.

There have been a considerable amount of research on biometrics authentication and some of them are based on hand gesture. Keystroke dynamics [17], mouse biometrics [16], multitouch gesture-based authentication [19], digital signature with accelerometer (e.g., uWave), and other approaches [11] have shown promising results in authentication, but they still require users to have contact with the devices. Some hand gesture authentication [8] is not practical in real-life application as it only uses a single capture image of the hand gesture and can also be easily mimicked by imposters, whereas the others [12, 13] require more room to perform and have very few available gestures that can be done as compared to stationary hand gesture authentication. Stationary hand gesture authentication does not require user to come into contact with the device, nor is it a trivial image based authentication, and it does not require a lot of room to be performed. In addition, it is more reliable, user-friendly, and secure compared to other hand gesture authentications which will be discussed in detail in Section 3.1.2.

In this paper, we have applied edit distance algorithm to stationary hand gesture behaviour to authenticate user by comparing the dissimilarities between the finger's states based on the time interval. Mean and standard deviation method and acceptance rate method are set as the thresholds to accept or reject the data. As for the result evaluation and analysis, accuracy from the confusion matrix and equal error rate (EER) from receiver operating characteristic (ROC) curve have been considered. The detailed methodology is explained in Section 4.

The outline of this paper is as follows: in Section 2, we explain the related work. The following section, Section 3, discusses hand gesture, Leap Motion controller, and edit distance algorithm. Section 4 explains our proposed method ED-FPDI, followed by results and discussion in Section 5. Finally, in Section 6, we close the paper with conclusion and present our future work.

2. Related Work

Chan et al. [20] have used Leap Motion controller for authentication via hand geometry and gestures. They have implemented two types of authentication in their experiment which is the static authentication and continuous authentication. In static authentication scenario, users have been asked to draw a circle with one finger; in continuous authentication scenario, users have been asked to perform basic actions, such as key or screen tapping, scrolling, swiping, and other actions. These data are being evaluated using different features: hand and finger properties, radius of the circle drawn, acceleration of hand, pinching and grabbing strengths, and so forth. These features are then classified using random forest classifiers and the result has shown that more weights are given to the physical properties of the hand and finger. The accuracy of the experiment is 99.97% on the static authentication and 98.39% on the continuous authentication.

Fong et al. [8] have proposed a biometric authentication model using static hand gesture images. Participants have been asked to perform American Sign Language of the 26 letters and each sign language was captured using a RGB camera. Features of the hand and finger, such as the position and angle, are extracted from the image for classification. The authors have used ten machine learning algorithms which are chosen from the major types of classification, including decision trees, rule-based methods, kernel functions, and Bayes methods. The experiment has shown promising overall result with a maximum accuracy of 93.75%.

Liu et al. [11] have proposed uWave, an efficient recognition algorithm for a single three-axis accelerometer such as the Nintendo Wii Remote. This method is similar to that of a signature-based recognition where users use a handheld device and start scribbling onto a flat screen to form a pattern. The pattern is then compared with other patterns using dynamic time warping (DTW) to authenticate the users.

Lai et al. [12] have proposed a user identification method using Microsoft Kinect to detect the arm gesture of the user. The arm gesture has been recorded in the form of body silhouette and is compared with other recorded gestures using nearest neighbor (NN) algorithm to identify user. Wu et al. [13] have expanded the approach of Lai et al. [12] by adopting Kinect skeleton model instead of body silhouette and extending the method to perform user authentication instead of only identification. The DTW algorithm has been used to compare the gestures. Unlike uWave, these approaches do not require user to equip or hold any devices on them.

Sayed et al. [16] have proposed biometric authentication using mouse gesture to create pattern by moving the mouse around. The method which attempts to find the time differences between two similar patterns is learning vector quantization (LVQ) neural network.

On the other hand, Sae-Bae et al. [19] have proposed authentication on multitouch devices. By comparing the touch sequence and time interval of the sequence using DTW algorithm, they have been able to authenticate users through touchscreen devices. The authentication is not limited to only one touch sequence.

Both Chan et al. [20] and Fong et al. [8] are closely related to our proposed method. Similar to Fong et al., we have used hand gestures identical to that of sign languages, but, instead of using static images, we have recorded these gestures in motion which includes the gestures and also the changes between each gesture. Chan et al. [20], on the other hand, have implemented finger gesture authentication through Leap Motion controller with gestures such as drawing, scrolling, and tapping. But their results show that most of the weights for authentication are placed into the physical properties of the hand and finger instead of the gesture. In comparison, Fong et al.'s method [8] can be easily copied by other users since it is just static hand gesture, whereas Chan et al. method [20] focuses more on hand properties which can easily be done using hand sculpture for authentication. Fong et al.'s method [8] is undeniably accurate in authentication,

but, in some situations, it may be easy for imposters to copy static hand gesture because the method considers simple static images, whereas Chan et al.'s method [20] focuses more on hand properties which are considered as physiological biometrics instead of behavioural based biometrics, and it may also be insecure at some point as imposters can trace the genuine hand and duplicate it.

Recently, Liu et al. [21, 22] used dynamic time warping (DTW) [21] and canonical time warping (CTW) [22] to develop the kernel sparse coding method to analyze the time series to improve the performance of object recognition.

To our knowledge, we have yet to come across similar approaches to our proposed method which uses stationary hand gesture in authentication, as most of the research has been based on static hand gesture, arm gesture, and finger gesture. A comparison on these different hand gestures and the advantages of stationary hand gesture over other hand gesture is stated in Sections 3.1.1 and 3.1.2. As for other related work, due to the difference in methodology used, we have constructed a table of comparison in Table 8 and discussed in Section 5.3.2 as to why gesture-based authentication is more preferable. Since there are also different types of gesture recognition available, a comparison among different gesture recognition has been constructed in Table 9.

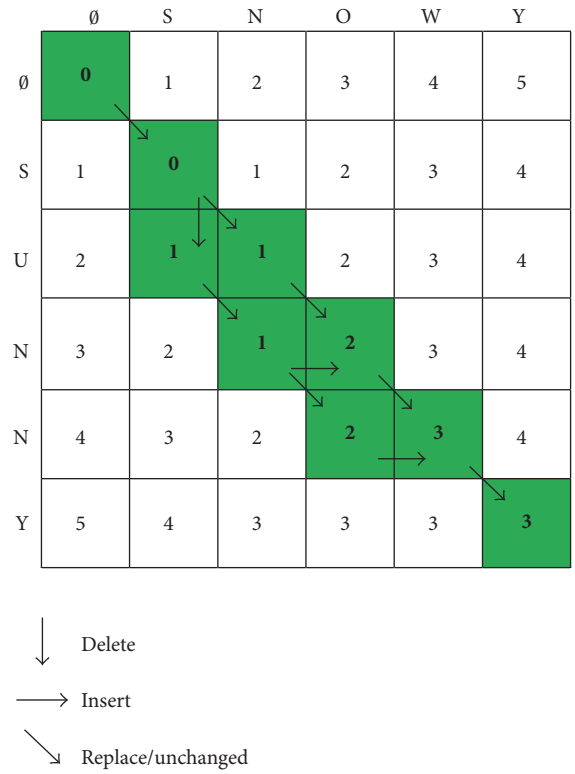


Figure 1: Edit distance matrix of string "SUNNY" to "SNOWY."

3. Background

3.1. Hand Gesture and Its Behaviour

3.1.1. Categories of Hand Gesture. Although keyboard and mouse are the traditional way of providing input to the computer, they are considered unnatural for human interaction [23, 24]. Hand gesture as an input has advantages over traditional input methods, such as being handsfree, able to communicate with the device from a distance, and more natural due to daily usage of hand gesture in human-to-human (H2H) communication[25, 26].

Hand gesture is a very vague term as there are so many different types of hand gesture available. Therefore, we have categorized hand gesture in Figure 1.

The terms used in hand gesture classification are as follows:

(i) Static means no movement or motion on the subject at all, usually in single image form.

(ii) Motion means there is movement involved.

(iii) Stationary means no movement on the particular subject, but objects that do not affect the movement of the subject can still be moving; for example, stationary hand gesture is when the hand, wrist, and arm (since movement of the wrist and arm will affect the position of the hand) are not moving or having any motion, whereas the fingers are free to move as long as this does not affect the position of the hand.

(iv) Dynamic means any part of the subject can have movement of motion even if it affects other objects; for example, dynamic hand gesture allows hand to move freely and, at the same time, fingers' position will be affected.

In some papers [12, 13], they refer to arm gesture as hand gesture which cause a lot of confusion; thus, we separate them into two different categories (Table 1). As for hand gesture, there are two major types, which are static hand gesture and motion hand gesture. We have defined two types of motion hand gesture: stationary hand gesture (hand and wrist are not moving but fingers are moving) and dynamic hand gesture (hand, fingers, and wrist are moving). When both arm and hand are considered into the recognition, we refer to it as arm with hand gesture. Lastly, as for drawing, scribbling, and swiping with just the finger, we refer to them as finger gesture. There are two subcategories in finger gesture, which are hand finger gesture (finger movement is only affected by the movement of the hand and fingers) and arm finger gesture (finger movement is affected by the movement of the arm, hand, and fingers). In our experiment, we focus on only stationary hand gesture-based authentication.

3.1.2. Advantages of Stationary Hand Gesture over Other Hand Gesture Approaches. There have been a few researches and publications on hand gesture authentication and its behaviour such as using hand gesture image (static hand gesture), dynamic hand gesture, arm gesture, and finger gesture (draw shape or signature using fingertip). But until now, there has yet to be any research on stationary hand gesture authentication. Moreover, there are reasons to why we choose stationary hand gesture over the other approaches:

Table 1: Category of hand gestures.

Gesture	Type
Arm gesture	Static arm gesture
	Motion/dynamic arm gesture
Hand gesture	Static hand gesture
	Motion hand gesture
	Dynamic hand gesture (motion on hands, fingers, and wrist)
	Stationary hand gesture (motion on fingers only)
Arm with hand gesture	Static arm with hand gesture
	Motion/dynamic arm with hand gesture
Finger gesture	Hand finger gesture (drawing or swiping with hand and finger only)
	Arm finger gesture (drawing or swiping with the arm and finger)

(i) Hand gesture image can be easily mimicked. It authenticates user by detecting the position, angle, and physical properties of hand on just one image. This image could probably be printed out and used on the system by other users or imposters.

(ii) Dynamic hand gesture will probably be the best in terms of authentication out of all the other hand gesture authentication methods, but it is difficult to compute since there are too many factors that need to be examined such as the gestures of finger, hand, palm, wrist, and arm. In addition, there has yet to be any reasonable device in accurately recognizing the entire arm, hand, and fingers together.

(iii) Arm gesture requires a larger space to be performed. It is also very difficult to conceal the gesture from outsiders. In addition, there are not many arm gestures available besides swinging up and down, left and right.

(iv) Finger gesture can be effective on simple gesture but sometimes can also be easily copied by others. Complex gesture, on the other hand, can be more secure but user may need visual cue on their gesture to prevent drawing in the wrong order or mispositioning.

Stationary hand gesture cannot be easily mimicked due to different behaviour; it does not need complicated computational algorithm or large space to perform, and it does not necessarily need any visual support while performing.

3.1.3. Stationary Hand Gesture Behaviour. Even though hand gesture can be used as password, but the issue with hand gesture is that it can be easily mimicked; however, due to the development through habits, behaviour in the gesture cannot be easily copied [27]. For example, when a hand is moving from an open to a close gesture (stretching out all fingers to retracting all fingers into the palm), different people will have distinct timing, moving angle, and position while doing the same gesture. It may not be impossible to mimic one's behaviour, but it surely takes a huge amount of effort which may take years to perfect it. Moreover, behaviour may change over time depending on gender, age, environment, and so forth. Therefore, before perfecting a particular person's behaviour, he/she may have changed or developed new behaviour.

Stationary hand gesture behaviour can be determined by, but not limited to, the finger's moving speed, acceleration, position, and moving angle. In our experiment, we have adapted finger pointing direction as the feature of hand gesture behaviour. It features the pointing direction of each finger based on the finger's state, which are open (stretch out) and close (retract) time interval. The time interval can be either long or short depending on various users. Leap Motion controller has been used to record all the data in our experiment. The data contain each finger's information including the time intervals and pointing directions. These behaviours can be easily distinguished by computer but not human, since the time interval is calculated in microseconds.

3.2. Leap Motion Controller. Leap Motion controller is a small sensor device that tracks hand, finger, or pointed object, either in static state or in motion state as input with very high accuracy and precision [28, 29]. It outperforms many other sensor devices, in terms of hand gesture detection or recognition, such as Microsoft Kinect and Creative Senz3D [30]. Moreover, the Leap Motion controller can record hand gesture at the rate of over 200 frames per second (fps) [31], but, in our experiment, the recorded data is roughly around 110 fps only. Nevertheless, it is sufficient to determine the behaviour of the gesture as the interval between each frame is denoted in microseconds.

3.3. Edit Distance. Edit distance [32] or Levenshtein distance is used to find the minimum number of operations required between two strings (or words). There are three types of operations in edit distance, insert, delete, and replace, where each transformation costs one operation. An example is to change the string "SUNNY" to "SNOWY": "S" remains, delete "U", first "N" remains, replace second "N" with "O", insert "W", and "Y" remains. This string change costs three

operations which includes delete, replace, and insert. The algorithm for edit distance is shown in

$$E(i,j) = \min \begin{cases} 1 + E(i-1,j), & \text{delete} \\ 1 + E(i,j-1), & \text{insert} \\ \text{diff}(i,j) + E(i-1,j-1), & \text{replace,} \end{cases} \quad (1)$$

where i is the length of the first string and j is the length of the second string.

Figure 1 shows the edit distance matrix where each operation costs one point. As shown in the figure, going down the matrix is the operation delete, going right is the operation insert, and going diagonally down right is the operation replace or remain unchanged. The very end of the matrix, which is located at the lower right corner of the matrix, is considered as the shortest path or minimum number of operations needed for the transformation of two strings.

In authentication, accuracy and speed are the most important aspects. We have chosen edit distance in experiment due to its speed, simplicity, and efficiency in detecting the differences in data. Edit distance saves both computational and authentication time and gives accurate prediction on either genuine user or imposter.

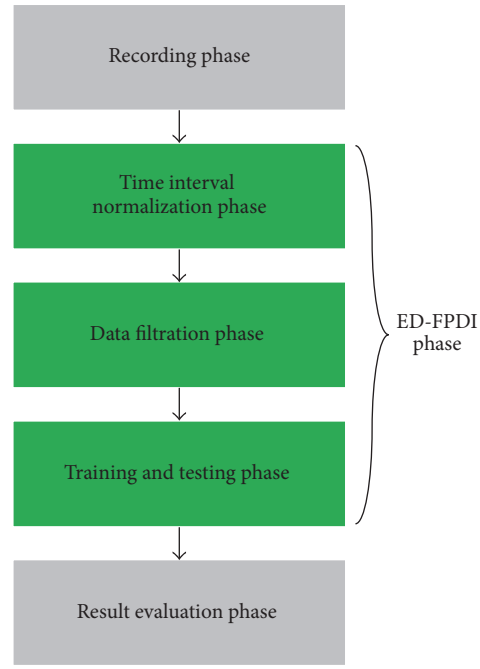


Figure 2: Flowchart of the experiment procedure.

4. ED-FPDI

In this paper, we introduce a novel approach for stationary hand gesture authentication, edit distance on finger pointing direction interval (ED-FPDI). Our experiment consists of five phases: recording phase, time interval normalization phase, data filtration phase (starting and ending point filtration), training and testing phase, and result evaluation phase. Time interval normalization phase, data filtration phase, and training and testing phase are also the core phases of ED-FPDI. The overall procedure of the experiment is illustrated in Figure 2.

4.1. Recording Phase. Vatavu and Zaiti [1] have recorded and published Leap Motion hand gesture dataset for remote control of devices. Their dataset is very useful for remote control experiments, but it does not contain stationary hand gesture data except open palm and close palm. Other gestures recorded in their dataset are mostly swinging hands in different direction and drawing letters or shapes (finger gesture). Although there were eighteen different participants data recorded, the data recorded is not used for authentication, but it is used as a remote control for television functions, such as changing channel and volume, opening menu, and showing TV guide. Therefore, their dataset is not directly suitable for our proposed stationary hand gesture authentication.

Thus, to our knowledge, there has yet to be any stationary hand gesture dataset recorded using Leap Motion controller available; thus we have to conduct our own recording data. The setup for the recording session of the experiment is demonstrated in Figure 3.

There are a total of ten participants, including seven males and three females, in their 20s, who participated in our experiment. All gestures are done with only the right hand. We have asked the participants to perform the same two sets of hand gestures at their own pace while keeping them as consistent as possible. Nine of them were asked to perform 25 times on a hand gesture which are used as the test data of the experiment, also known as the imposters, whereas one of the participants, who serves as the control group which in this case known as the genuine user, was asked to perform 125 times on the hand gesture, where 100 instances of data are used as the training data while the remaining 25 instances of data are used as the test data. In summary, a total of 350 instances of data on a single gesture were recorded; 100 instances of data from the genuine user are used for training, while 250 remaining instances of data are used for testing, where 25 instances of data are from the genuine user and the remaining 225 instances of data are from the imposters. Training data and test data from the genuine user are independent. The hand gestures recorded are known as "205" and "7631" which are shown in Figures 4 and 5, respectively. It is worth mentioning that these gestures are defined by the genuine user and are not restricted to these gestures only, and the experiment makes the assumption that all the imposters have already known the hand gestures or password gestures [12] (since the hand gestures are considered as passwords in our experiment) that the genuine user has inputted.

Figure 6 shows the axes of the Leap Motion controller. The hand will be placed on top of the controller, parallel to z-axis of the controller. In our experiment, the hand direction is always pointing towards the negative z-direction (pointing towards the monitor as shown in Figure 6) between −1.0 and −0.9 in unit vector, whereas x-direction and y-direction of

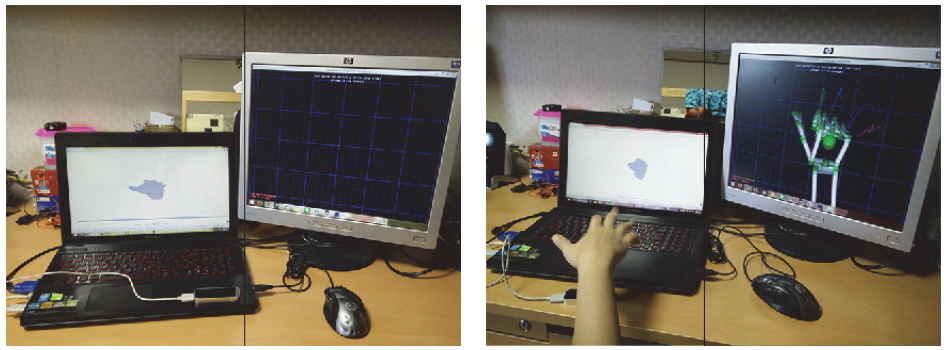


Figure 3: Setup of the hand gesture recording. Left monitor is the recorder, whereas the right monitor is the top view of the hand. The top view is necessary to detect errors and inaccuracies of hand gestures during recording.

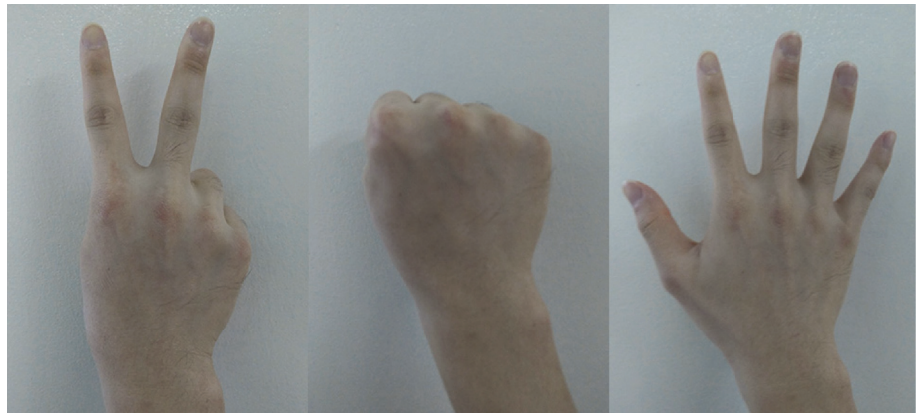


Figure 4: Hand gesture for "205."

the hand are as closed to 0 in unit vector as possible (even when performing gesture).

While recording the gestures, we have found out that gesture "7631" produces a lot of errors from the Leap Motion controller while transforming from "6" to "3" and from "3" to "1." The errors have occurred for more than half of the time during recording for each user, meaning that each user has been recording for more than 50 times just to arrive at 25 instances of data without any error. It has also worthwhile to note that the speed of each user on this gesture varies by a big margin including the genuine user and therefore it may influence the accuracy of the experiment. As mentioned before, these errors were not accidentally or purposely made by the user but by the detection inaccuracy of the Leap Motion controller or software as shown in Figure 7.

We would like to note that the hand gestures in our experiment are not static; they are motion gesture which changes from one static gesture to another static gesture. For example, it changes from "2" fingers to "0" fingers and the motion between the two gestures is considered in our experiment.

4.2. Time Interval Normalization Phase. This phase is the combination of normalizing time interval and filtering data steps. From the recorded data, the intervals between each frame of the data are nonconsistent and therefore have to be normalized before comparison. We have adopted linear interpolation to normalize the interval. For example, the intervals can be 0, 12345, 23456, and 34567 microseconds in the recorded data. After normalization, they will be 0, 10000, 20000, and 30000 microseconds. This has to be done in order for every data to have a consistent interval for ease of comparison when using edit distance algorithm.

4.3. Data Filtration Phase. Our proposed ED-FPDI considers only two timing elements from each finger which are open finger and close finger. More detailed description of the two timing elements is shown as follows:

(1) Open finger (O): stretching the finger outwards (when the pointing direction along z-axis is between -0.8 and -1.0 in vector unit)

(2) Close finger (C): retracting the finger into the palm (when the pointing direction along z-axis is larger than -0.8)

The reason that these threshold numbers are chosen as the timing elements is based on the observation that most of open finger data ("O") in the recorded dataset lie on positions between -0.8 and -1.0 in z-axis. We have applied only two segmentations due to the inaccuracy of the Leap Motion in detecting when the fingers are fully closed as shown in

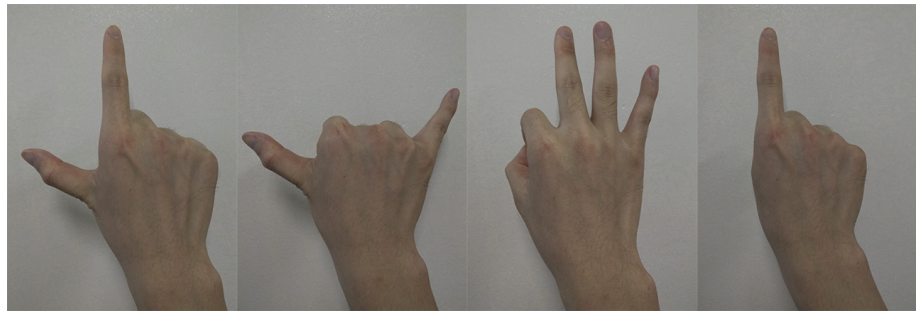


Figure 5: Hand gesture for "7631."

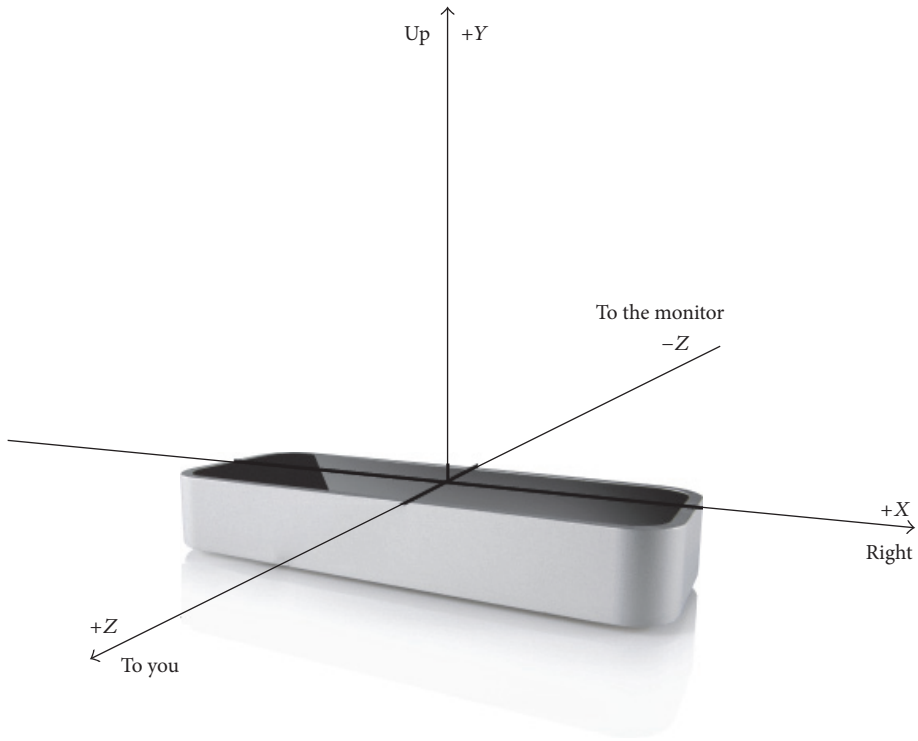


Figure 6: Leap Motion controller and its axes.

Figure 7. It can be seen that even when the fingers are completely closed, the Leap Motion controller mistakenly reports that the fingers are only halfway closed. We have observed that the motion from open to close (and vice versa) is supposed to be around −0.8 to 0.8, while the close element ("C") is between 0.8 and 1.0 in unit vector. However, again due to the errors from the Leap Motion controller, these measurements are unable to be incorporated into our system. Note that this observation is considered under the assumption when the hand pointing at z-direction is between −0.9 and −1.0 in unit vector (pointing forward with the Leap Motion controller), and the pointing y-direction is around 0 in unit vector (parallel with the Leap Motion controller). Throughout the experiment, we have also found out that only three fingers (the index, the middle, and the ring fingers) have exhibited consistent precision of pointing direction along z-axis, between −0.8 and −1.0 in vector unit. Due to the large margin differences in the pointing direction of the thumb and the little fingers varying from different people, they are not taken account in our experiment; however, the experimental results do not show significant degradation. We show two samples in Table 2 to investigate these observations.

From Table 2, we can see that Data 1 forms a string of "OOCCCCCCCOO," while Data 2 forms "OOOCCCCCO." After applying edit distance algorithm from Data 1 to Data 2, we obtain the minimal edit distance, $E(i, j) = 4$, which means that there is a minimum of four operations to change from Data 1 to Data 2.

But before the application of edit distance to the data, we need to filter out the excess starting gesture and ending gesture. Detailed visualization of the raw data and the filtered data is depicted in Figure 8. As shown in Figure 8(a), the raw data have not defined the starting and the ending point of the gesture. Therefore, we filter out the excess gesture before

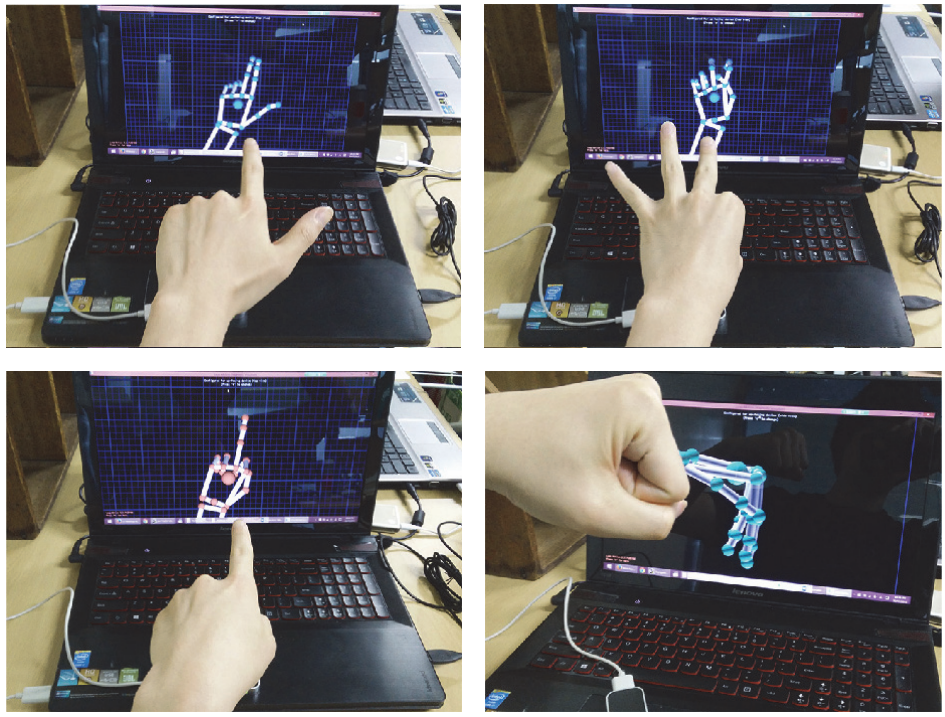


Figure 7: Hand gesture errors and inaccuracies shown by the Leap Motion controller while performing hand gestures.

Table 2: Examples of time interval for a finger from two different instances of data. Value lesser or equal to −0.8 is considered open (O), while value larger than −0.8 is considered close (C).

Interval (ms)	10	20	30	40	50	60	70	80	90	100	110
Data 1	−0.9	−0.8	−0.5	−0.1	0.4	0.8	0.3	−0.2	−0.7	−0.9	−0.9
Data 2	−0.8	−0.9	−0.8	−0.2	0.1	0.4	−0.4	−0.7	−0.8	—	—

the starting and after the ending point illustrated in Figure 8(b), by detecting the initial movement of the finger from a predefined starting hand gesture (usually hand gesture of "5"). Similarly, we detect the ending point by locating the point where the hand gesture changes from the last gesture to a predefined ending hand gesture (usually hand gesture of "5"). If the starting point or ending point of the gesture is actually "5" (such as the case "205"), then another predefined hand gesture is used. Note that this filtering detection is done manually. The purpose of filtering is to ensure the consistent starting point and the ending point for every gesture.

4.4. Training and Testing Phase

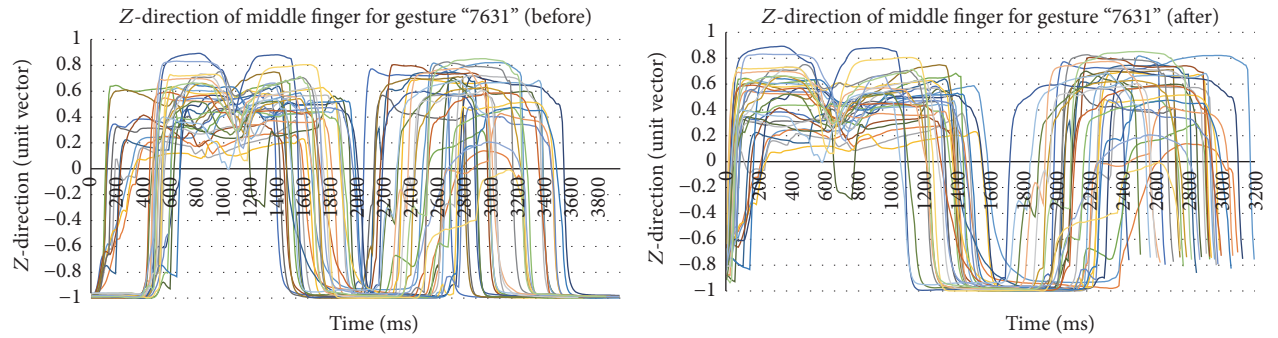
4.4.1. Mean and Standard Deviation (M&SD) Method. After the completion of filtering and segmenting the raw data, we apply edit distance algorithm to the processed dataset. First of all, we calculate pairwise edit distances $E(\mathrm{Tr}_i, \mathrm{Tr}_j)$ for the 100 training instances, where E is the edit distance and Tr is the training data. The total number of edit distance calculations for each pair of instances in the training data is 4,950, which can be easily obtained from a combination formula of $\binom{100}{2}$. From the calculated edit distance values, we find the mean and standard deviation to acquire the threshold interval. Note that we have three records (or fixed-length sequences) in one training instance of which each record corresponds to one finger. In other words, we have three threshold intervals (i.e., each of them for one finger) in one training instance. The acquired threshold intervals are used as guidelines for authentication.

After the training phase, we apply the 250 test instances for edit distance calculation with the 100 training instances. Each test instance is compared with the 100 instances of training data and return 100×3 values of $E(\mathrm{Te}, \mathrm{Tr})$, where E is the edit distance and Te is the test data. Note that we multiply by 3 because we calculate the mean of 100 $E(\mathrm{Te}, \mathrm{Tr})$ values for one finger and we repeat the same procedure for other two fingers. For each finger in the test instance, we compare its mean with the corresponding threshold interval estimated from the training data. If the mean is within the threshold interval, we mark that particular finger in the test instance as accepted. And if all three fingers in the test instance are accepted, we consider the user as a genuine user, otherwise as an imposter. One example of these procedures is shown in Table 3.

From Table 3, Data 1 will be considered as a genuine user because all of its mean values of each finger are within the threshold interval. Data 2 and Data 3 will be noted as imposters since at least one of the mean values of each finger in

Table 3: Examples of threshold intervals from the training data using the mean and standard deviation and mean distances (differences) from each test instance with training instances.

Finger	Index	Middle	Ring
Threshold	14.24~43.12	14.98~42.26	11.81~38.24
Data 1	25	26	24
Data 2	10	72	68
Data 3	50	41	37



(a) Samples of z-direction for the middle finger in gesture "7631" extracted from the training data before the filtration

(b) Samples of z-direction for the middle finger in gesture "7631" from the training data after the filtration. Note that the first 400~600 ms portion in the original data is truncated as well as the last portion which have −1 unit vector in z-direction.

Figure 8: Z-direction of middle finger for raw data and filtered data.

the data is outside of the threshold interval. Pseudocode 1 demonstrates the pseudocode of our mean and standard deviation (M&SD) method, where M is the mean and SD is the standard deviation. It can be seen that the asymptotic upper bound of M&SD method's training time lies in $O(n^2m^2)$ where $n = |T_r|$ is the number of training instances and $m = \max_i |T_{r_i}|$ is the length of the longest training instance. The asymptotic upper bound of M&SD method's test time is $O(nNm^2)$ where $n = |T_r|$ is the number of training instances, $N = |T_e|$ is the number of test instances, and $m = \max_i |T_{r_i}|$ is the length of the longest training instance.

4.4.2. Acceptance Rate (AR) Method. In addition to mean and standard deviation (M&SD) method, we have discussed in Section 4.4.1 that we propose acceptance rate (AR) method, which uses the acceptance rate as a threshold calculated from the training data. For calculating the acceptance rate, we apply mean and standard deviation obtained from the training data to the training data itself. The main idea is that, for example, if the acceptance rate of the training data is 0.5, then the acceptance rate of test data should be close to 0.5 to be accepted as that of a genuine user. In other words, if the acceptance rate of the test data is far from the acceptance rate of the training data, the chance of being accepted as that of a genuine user is low.

In order to calculate the acceptance rate of the training data (AR_{Tr}), our algorithm compares the M&SD threshold found from the previous method with all the other 4950 edit distances $E(Tr_i, Tr_j)$ in the training data. The M&SD threshold from the training data is then applied to the test data to obtain the 100 edit distances $E(Te, Tr)$. In order to obtain the acceptance rate of the test data (AR_{Te}), the 100 $E(Te, Tr)$ are compared with the same M&SD threshold from the training data. Then our algorithm uses the acceptance rate of training data (AR_{Tr}) to compare with the acceptance rate of the test data (AR_{Te}). The difference of AR_{Tr} and AR_{Te} is used to form the ROC curve. Note that the acceptance rate can be estimated in two ways: the average acceptance rate of all fingers (Average AR) and the acceptance rate of each finger (Each AR). The pseudocode for this approach is shown in Pseudocode 2.

The asymptotic upper bound of AR method's training time is $O(n^2m^2)$, where $n = |T_r|$ is the number of training instances and $m = \max_i |T_{r_i}|$ is the length of the longest training instance. And the asymptotic upper bound of AR method's test time is $O(nNm^2)$, where $n = |T_r|$ is the number of training instances, $N = |T_e|$ is the number of test instances, and $m = \max_i |T_{r_i}|$ is the length of the longest training instance. In terms of asymptotic time complexity, we can see that AR method is as fast as M&SD method.

4.5. Result Evaluation Phase. As mentioned before, we consider two evaluation methods: accuracy from confusion matrix method and EER from ROC curve method. For clear explanation of confusion matrix, we describe a few basic terminologies here:

(1) True acceptance rate (TAR): genuine user being accepted (good)

(2) True rejection rate (TRR): imposters being rejected (good)

Mean and Standard Deviation:
Input: Training dataset and test dataset
Output: Hand gesture accepted (genuine user) or rejected (imposter), and ROC curve
Training phase:
begin
(1) **for** each Tr_i compares with Tr_j, where $i \neq j$
calculate edit distance $E(\mathrm{Tr}_i, \mathrm{Tr}_j)$
(2) calculate M and SD of $\forall E(\mathrm{Tr}_i, \mathrm{Tr}_j)$
set M ± SD as threshold interval
end.
Testing phase:
Begin
(1) **for** each Te compares with ∀Tr
calculate edit distance $E(\mathrm{Te}, \mathrm{Tr})$
(2) calculate M of $E(\mathrm{Te}, \mathrm{Tr})$ in each dataset
(3) **if** $E(\mathrm{Te}, \mathrm{Tr})_{\mathrm{M}}$ of all fingers in a dataset are within threshold interval
accepted (genuine user)
else
rejected (imposter)
(4) plot ROC curve
(5) find EER from the ROC curve graph
end.

Pseudocode 1: Pseudocode of the mean and standard deviation (M&SD) method.

Acceptance Rate:
Input: Training dataset and test dataset
Output: ROC curve
Training phase:
begin
(1) **for** each Tr_i compares with Tr_j, where $i \neq j$
calculate edit distances $E(\mathrm{Tr}_i, \mathrm{Tr}_j)$
(2) calculate M and SD of $\forall E(\mathrm{Tr}_i, \mathrm{Tr}_j)$
set M ± SD as threshold
(3) **if** $E(\mathrm{Tr}_i, \mathrm{Tr}_j)$ is within threshold
$E(\mathrm{Tr}_i, \mathrm{Tr}_j)_{\mathrm{Accepted}}$
else
$E(\mathrm{Tr}_i, \mathrm{Tr}_j)_{\mathrm{Rejected}}$
(4) calculate $\mathrm{AR}_{\mathrm{Te}}$ = Total no. of $E(\mathrm{Tr}_i, \mathrm{Tr}_j)_{\mathrm{Accepted}}$/Total no. of $E(\mathrm{Tr}_i, \mathrm{Tr}_j)$
end.
Testing phase:
begin
(1) **for** each Te compares with ∀Tr
calculate edit distances $E(\mathrm{Te}, \mathrm{Tr})$
(2) **if** $E(\mathrm{Te}, \mathrm{Tr})$ is within threshold
$E(\mathrm{Te}, \mathrm{Tr})_{\mathrm{Accepted}}$
else
$E(\mathrm{Te}, \mathrm{Tr})_{\mathrm{Rejected}}$
(3) calculate $\mathrm{AR}_{\mathrm{Te}}$ = Total no. of $E(\mathrm{Te}, \mathrm{Tr})_{\mathrm{Accepted}}$/Total no. of $E(\mathrm{Te}, \mathrm{Tr})$
(4) plot ROC curve using difference between $\mathrm{AR}_{\mathrm{Te}}$ and $\mathrm{AR}_{\mathrm{Tr}}$
(5) find EER from the ROC curve graph
end.

Pseudocode 2: Pseudocode of the acceptance rate method including its evaluation.

Table 4: Results on gesture "205." User number 1 is the genuine user, while the other users are imposters.

User	Accepted	Rejected
1 (genuine)	**22**	**3**
2 (imposter)	0	25
3 (imposter)	2	23
4 (imposter)	2	23
5 (imposter)	1	24
6 (imposter)	2	23
7 (imposter)	2	23
8 (imposter)	0	25
9 (imposter)	1	24
10 (imposter)	3	22
Total imposters (225 instances of data)	**13**	**212**

(3) False rejection rate (FRR): genuine user being rejected (bad)

(4) False acceptance rate (FAR): imposters being accepted (worse)

For calculating accuracy from confusion matrix method, we calculate it by summing up the true acceptance rate (TAR) with the true rejection rate (TRR) and divide the sum with the total number of instances. However, note that this method is only applicable for mean and standard deviation (M&SD) method in our experiment.

Receiver operating characteristic (ROC) curve [33] has also been used to measure the performance of our algorithm. ROC curve is a graph that illustrates the performance of classifiers by presenting the trade-off between hit rates and false alarm rates while varying the threshold. The advantage of the ROC curve is that we can find the equal error rate (EER) [34] which has been used in most of the biometric security systems to measure the actual performance in imbalanced data. Hence, low EER indicates high accuracy. EER is obtained when both acceptance rate and rejection rate are equal. Both the M&SD method and the acceptance rate (AR) method have been evaluated using this method.

5. Experimental Results and Discussion

Tables 4 and 5 show the experimental results of the proposed algorithm on both the genuine user and imposters about the test gesture "205" and "7631," respectively. In Table 4, 22 out of 25 instances of genuine user data have been accepted by the system, whereas 13 out of 225 imposters have been accepted. In Table 5, only 20 out of 25 instances of genuine user data have been accepted by the system and 30 out of 225 imposters have been accepted. The numbers of acceptance and rejection for both genuine user and imposters are also shown individually in Tables 4 and 5.

5.1. Accuracy from Confusion Matrix. The equations of true acceptance rate (TAR), false acceptance rate (FAR), and accuracy are explained in (2), (3), and (4), respectively:

$$\text{TAR} = \frac{\text{Genuine user data accepted}}{\text{Total genuine user data}}, \tag{2}$$

Table 5: Results on gesture "7631." User number 1 is the genuine user, while the other users are imposters.

User	Accepted	Rejected
1 (genuine)	**20**	**5**
2 (imposter)	0	25
3 (imposter)	0	25
4 (imposter)	4	21
5 (imposter)	3	22
6 (imposter)	9	16
7 (imposter)	0	25
8 (imposter)	7	18
9 (imposter)	2	23
10 (imposter)	5	20
Total imposters (225 instances of data)	**30**	**195**

$$\text{FAR} = \frac{\text{Imposter user data accepted}}{\text{Total imposter data}}, \tag{3}$$

$$\text{Accuracy} = \frac{\sum \text{True acceptance} + \sum \text{True rejection}}{\sum \text{Total data}}. \tag{4}$$

From Table 4, TAR, FAR, and accuracy for gesture "205" are calculated as follows:

$$\begin{aligned} \text{TAR} &= \frac{22}{25} = 0.8800, \\ \text{FAR} &= \frac{13}{225} = 0.0578, \\ \text{Accuracy} &= \frac{22 + (225 - 13)}{250} = 0.9360. \end{aligned} \tag{5}$$

From Table 5, TAR, FAR, and accuracy for gesture "7631" are calculated as follows:

$$\begin{aligned} \text{TAR} &= \frac{20}{25} = 0.8000, \\ \text{FAR} &= \frac{30}{225} = 0.1333, \\ \text{Accuracy} &= \frac{20 + (225 - 30)}{250} = 0.8600. \end{aligned} \tag{6}$$

From the above calculations, it can be seen that gesture "205" has a higher accuracy of 0.9360, as compared to that of gesture "7631" which is only 0.8600. The reason gesture "7631" has a lower accuracy as compared to gesture "205" has been discussed in Section 4.1 and shown in Figure 7, which is due to the inaccuracy of the Leap Motion controller. Nevertheless, both gestures have shown fairly high accuracy of more than 0.8.

5.2. Equal Error Rate (EER) from Receiver Operating Characteristic (ROC) Curve. ROC curves of gesture "205" and gesture "7631" are depicted in Figures 9 and 10, respectively. The red line represents the results from the mean and standard deviation (M&SD) method, whereas the green line represents the average acceptance rate of all the fingers from

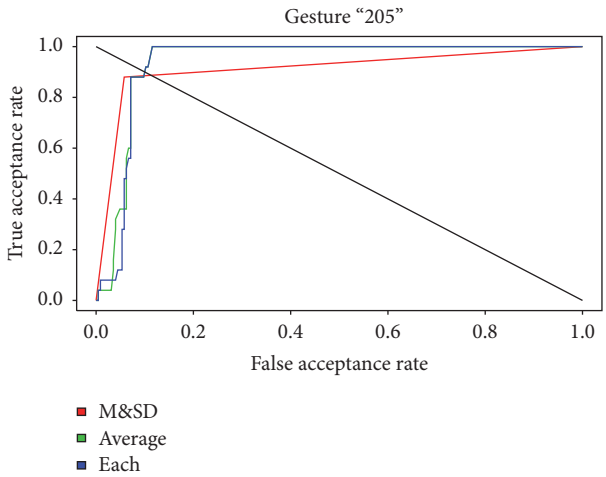


Figure 9: ROC curves and the EER line on gesture "205."

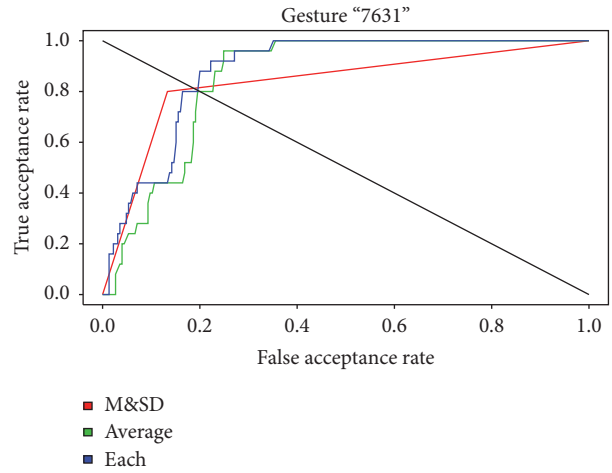


Figure 10: ROC curves and the EER line on gesture "7631."

Table 6: The summarized results of both gesture "205" and gesture "7631" using confusion matrix. High TAR, low FAR, and high accuracy indicate high performances.

Gesture	TAR	FAR	Accuracy
205	0.8800	0.0578	0.9360
7631	0.8000	0.1333	0.8600

the training data to be compared with each finger of the test data; and the blue line represents each acceptance rate of the finger from the training data to be compared with the similar finger from the test data. For example, from the training data, the index finger's acceptance rate will only be compared with the index finger from the test data. The black line that intercepts all those lines through point (0,1) to point (1,0) is the line of EER. The accuracy in ROC curve can be denoted as follows:

$$\text{Accuracy} = 1.0000 - \text{EER}. \quad (7)$$

From ROC curve of gesture "205," the M&SD method results in EER of 0.1130 and accuracy of 0.8870, whereas both average and each finger acceptance rate method produce EER of 0.1000 and accuracy of 0.9000. As for gesture "7631," the M&SD method yields EER of 0.1875 and accuracy of 0.8125 and average finger acceptance rate method exhibits EER of 0.2000 and accuracy of 0.8000, while each finger acceptance rate method shows EER of 0.1958 with 0.8042 as accuracy.

5.3. Summarized Results and Discussion

5.3.1. Summarized Results. Tables 6 and 7 show the summary of the result evaluations on both gesture "205" and gesture "7631." Both gestures are evaluated using two different methods, accuracy of confusion matrix, and equal error rate (EER) from ROC curve.

Table 6 is the result in confusion matrix which consists of the following:

(i) True acceptance rate (TAR) based on (2)

(ii) False acceptance rate (FAR) based on (3)

(iii) Accuracy based on (4)

Table 7 is the result in ROC curve which consists of the following:

(i) Evaluation methods

  (a) Mean and standard deviation M&SD

  (b) Average acceptance rate of all fingers from training data compared with test data (Average AR)

  (c) Acceptance rate of each finger from training data compared with the similar finger from test data (Each AR)

(ii) Equal error rate (EER) from the interception point of the ROC curve with the straight diagonal line shown in Figures 9 and 10

(iii) Accuracy based on (7)

In the case of accuracy, higher value indicates higher performance; however, in the case of EER, lower value signifies higher performance.

5.3.2. Discussion. As is seen from Tables 6 and 7, there is a difference between the accuracy obtained from the confusion matrix and ROC curve. This is due to the imbalance data between the genuine user and the imposters: 25 instances of genuine data and 225 instances of imposter data. If the data on both subjects are equal, for example, 100 instances of genuine data and 100 instances of imposter data, then the accuracy on both confusion matrix and ROC curve will be the same.

In Tables 6 and 7, the results on gesture "205" shows higher performance than those of gesture "7631." It is because the proposed algorithm adopts smaller acceptance threshold for the training data of gesture "205" than that for gesture "7631" in the experiment. Hence, the accuracy results for gesture "7631" with larger acceptance threshold turn out to be lower than those for gesture "205." The reason why we

Table 7: The summarized results of both gesture "205" and gesture "7631" using ROC curve. Low EER and high accuracy indicate higher performances. "Average AR" denotes the AR method using the average acceptance rate of all fingers and "Each AR" means the AR method using the acceptance rate of each finger.

Gesture	Evaluation	EER	Accuracy
205	ROC curve (M&SD)	0.1130	0.8870
	ROC curve (Average AR)	**0.1000**	**0.9000**
	ROC curve (Each AR)	**0.1000**	**0.9000**
7631	ROC curve (M&SD)	**0.1875**	**0.8125**
	ROC curve (Average AR)	0.2000	0.8000
	ROC curve (Each AR)	0.1958	0.8042

Table 8: Comparison of different biometrics approaches.

Type of biometrics	Keystroke dynamics	Mouse biometrics	Touch biometrics	Signature verification	Gesture
Naturalness	Unnatural	Unnatural	Seminatural	Seminatural	Natural
Versatility	Semiversatile (device dependent)	Not versatile	Semiversatile (device dependent)	Semiversatile (device dependent)	Versatile
Ease of use	Not easy	Not easy	Moderate	Moderate	Easy
Degree of freedom	Low	Low	Low	High	High
Error in recognition	Low	Low	Low	Moderate	High
Cost	Low	Low	Moderate	Moderate	High

adopt large acceptance threshold for gesture "7631" is due to the complications and difficulties in forming gesture "7631" as shown in Figure 7 and discussed in Section 4.1.

Table 8 shows the comparison of different biometrics approaches. As shown in the table, gesture-based biometrics has overall more advantages over other biometrics approaches. First of all, gesture-based biometrics is more natural for human to machine interaction. It is also more versatile than other approaches, because the device for gesture recognition can be implemented as small as a microchip [5]. Finally, it is easy to use, since humans usually communicate with other people using gesture, and there can be many distinct gesture patterns available.

From those systematic experiments, we have found that gesture-based biometrics has a few drawbacks in recognizing gesture such as low precision and expensive cost. We expect that as technology advances, these precision and price issues will diminish to widen a way for more gesture related applications.

As for gesture recognition, there are a few different types of gestures that can be used by humans, such as body gesture, head gesture, and hand gesture. We would like to note on the classifications of gesture because those different types of gesture and corresponding solutions often cause confusion to the readers. Table 9 shows the different type of gesture recognition that are available and their advantages and disadvantages.

Note that head gesture is not included in the table. For gesture-based control or user authentication, head gesture is impractical because many people will get nauseous while moving their heads around.

As can be seen from Table 9, stationary hand gesture has definitely more advantages over other gesture recognition in applications which need a lot of gestures. We can allegorically consider that stationary hand gesture is a motion picture while static gesture is a static image. In terms of versatility, stationary hand gesture only needs small device for recognition, whereas body gesture [35] and gait [36, 37] require larger device, for example, Leap Motion controller for hand gesture and Microsoft Kinect for body gesture. That is, the gesture space for stationary hand gesture is smaller compared to body gesture and gait which need more than a normal person space to perform. Stationary hand gesture is much more difficult to copy or mimic as it can easily be concealed from other people by just covering it with a box or your body. Finger gesture on the other hand may be difficult to conceal as it sometimes needs vision cue for the user to prevent misplacing in drawing or writing. Chances of stationary hand gesture to be implemented into other devices seem higher than other gestures as it is much more versatile and needs less space to perform. In summary, although stationary hand gesture-based user authentication has a moderate difficulty in performing because stationary hand gesture can generate more gesture patterns, this difficulty can secure the authentication from replay attack by imitation.

6. Conclusion and Future Work

Our approach, ED-FPDI in this paper has demonstrated a way to authenticate user using hand gesture with a significantly high accuracy of over 0.8 rate. This experiment is, of course, conducted under the assumption that the imposters

Table 9: Comparison among different gesture recognition.

Gesture type	Static gesture	Body gesture	Gait	Arm gesture	Finger gesture	Stationary hand gesture
Recognition	Static image	Motion	Motion	Motion	Motion	Motion
Versatility of device	Dependent on gesture	Not versatile	Not versatile	Not versatile	Versatile	Versatile
Gesture space	Dependent on gesture	Big	Big	Moderate	Moderate	Small
Difficulty in performing	Easy	Difficult	Easy	Easy	Moderate	Moderate
Ease of mimicking (due to difficulty of concealing gesture)	Very easy	Easy	Easy	Easy	Moderate	Difficult
Chances of implementation	Impractical but possible	Low	Low	Moderate	High	High

have already known the user's password gesture. It is worthwhile noting that, in an authentication system, false acceptance is very serious, so it has to be as low as possible, and our proposed algorithm has shown that the EER is as low as 0.2. It may have shown higher performance, if

(1) all fingers are taken into account;

(2) there is no hardware limitation or inaccuracy.

Hand gesture as password or authentication may not be used frequently now, but, with the upcoming technology, "smart" user interface will be implemented into most electronic devices, home appliances, vehicles, and other applications where the interfaces are mostly using gesture recognition as an input method. Our experiments indicate that hand gesture authentication shows promising future research opportunity. This is just the beginning of hand gesture authentication; therefore, more research and work have to be done before it can be used for critical authentication [11].

For our future work, we plan to record more hand gesture datasets into the experiment. In addition, we will explore new approaches to take the thumb and the little finger into account which may increase the performance. Finally, we will consider detailed analysis and improvement of Leap Motion controller as one of our future research directions.

Competing Interests

The authors declare that they have no competing interests.

References

[1] R.-D. Vatavu and I.-A. Zaiti, "Leap gestures for TV: insights from an elicitation study," in *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video (TVX '14)*, pp. 131–138, ACM, 2014.

[2] M.-J. Lee, Switch off living room light with you tv?, 2014, http://blogs.wsj.com/digits/2014/03/10/switch-off-living-room-lights-with-ya.

[3] J. Song, G. Sörös, F. Pece et al., "In-air gestures around unmodified mobile devices," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*, pp. 319–329, ACM, Honolulu, Hawaii, USA, October 2014.

[4] K. Russell, "The aria lets you control your smart watch without touching it," 2015, https://techcrunch.com/2015/05/22/the-aria-lets-you-control-your-smart-watch-without-touching-it/.

[5] GoogleATAP, Project soli, https://www.google.com/atap/project-soli/.

[6] C. Rogers, Car dashboard controls just a hand gesture away, 2015, http://www.wsj.com/articles/car-dashboard-controls-just-a-hand-gesture-away.

[7] P. Daugherty, O. Schybergson, and H. J. Wilson, "Gestures will be the next interface for the internet of things," 2015, https://hbr.org/2015/07/gestures-will-be-the-interface-for-the-internet-of-things.

[8] S. Fong, Y. Zhuang, I. Fister, and I. Fister Jr., "A biometric authentication model using hand gesture images," *BioMedical Engineering Online*, vol. 12, article 111, pp. 1–18, 2013.

[9] M. A. M. Shukran and M. S. B. Ariffin, "Kinect-based gesture password recognition," *Australian Journal of Basic and Applied Sciences*, vol. 6, no. 8, pp. 492–499, 2012.

[10] G. Ceste, A. D. Giorgio, D. Molin, and J. Turesson, "Gesture-based password recognition," Project Report, 2013.

[11] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uWave: accelerometer-based personalized gesture recognition and its applications," *Pervasive and Mobile Computing*, vol. 5, no. 6, pp. 657–675, 2009.

[12] K. Lai, J. Konrad, and P. Ishwar, "Towards gesture-based user authentication," in *Proceedings of the IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance (AVSS '12)*, pp. 282–287, September 2012.

[13] J. Wu, J. Konrad, and P. Ishwar, "Dynamic time warping for gesture-based user identification and authentication with Kinect," in *Proceedings of the 2013 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 2371–2375, IEEE, Vancouver, Canada, May 2013.

[14] F. Okumura, A. Kubota, Y. Hatori, K. Matsuo, M. Hashimoto, and A. Koike, "A study on biometric authentication based on arm sweep action with acceleration sensor," in *Proceedings of the International Symposium on Intelligent Signal Processing and Communications (ISPACS '06)*, pp. 219–222, IEEE, Yonago, Japan, December 2006.

[15] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[16] B. Sayed, I. Traore, I. Woungang, and M. S. Obaidat, "Biometric authentication using mouse gesture dynamics," *IEEE Systems Journal*, vol. 7, no. 2, pp. 262–274, 2013.

[17] J. Ho and D.-K. Kang, "Sequence alignment of dynamic intervals for keystroke dynamics based user authentication," in *Proceedings of the IEEE 15th International Symposium on Soft Computing and Intelligent Systems (SCIS), Joint 7th International Conference on and Advanced Intelligent Systems (ISIS '14)*, pp. 1433–1438, Kita-Kyushu, Japan, December 2014.

[18] E. Farella, S. O'Modhrain, L. Benini, and B. Riccó, "Gesture signature for ambient intelligence applications: a feasibility study," in *Pervasive Computing*, pp. 288–304, Springer, New York, NY, USA, 2006.

[19] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon, "Biometric-rich gestures: a novel approach to authentication on multi-touch devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 977–986, ACM, May 2012.

[20] A. Chan, T. Halevi, and N. Memon, "Leap motion controller for authentication via hand geometry and gestures," in *Human Aspects of Information Security, Privacy, and Trust*, T. Tryfonas and I. Askoxylakis, Eds., vol. 9190 of *Lecture Notes in Computer Science*, pp. 13–22, Springer, Berlin, Germany, 2015.

[21] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: kernel sparse coding methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

[22] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, 2016.

[23] Y. Fang, K. Wang, J. Cheng, and H. Lu, "A real-time hand gesture recognition method," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 995–998, Beijing, China, July 2007.

[24] J. Li, L. Zheng, Y. Chen, Y. Zhang, and P. Lu, "A real time hand gesture recognition system based on the prior facial knowledge and SVM," *Journal of Convergence Information Technology*, vol. 8, no. 11, pp. 185–193, 2013.

[25] N. Aimaiti and X. Yan, *Gestire-based interaction and implication for the future [M.S. thesis]*, Umea University, Umeå, Sweden, 2011.

[26] R. Lockton, "Hand gesture recognition using computer vision," 4th Year Project Report, 2002.

[27] BehavioSec, "The role of biometrics in it security and continuous authentication," http://behaviosec.com/wp-content/uploads/2012/01/The-Role-of-Biometrics-in-accessed.

[28] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[29] J. Guna, G. Jakus, M. Pogačnik, S. Tomažič, and J. Sodnik, "An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking," *Sensors*, vol. 14, no. 2, pp. 3702–3720, 2014.

[30] R. Balowin, "Why the leap is the best gesture-control system we've ever tested," 2012, https://www.wired.com/2012/05/why-the-leap-is-the-best-gesture-control-system-weve-ever-tested/.

[31] J. Han and N. Gold, "Lessons learned in exploring the leap motion tm sensor for gesture-based instrument design," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 371–374, London, UK, June 2014.

[32] W. J. Masek and M. S. Paterson, "A faster algorithm computing string edit distances," *Journal of Computer and System Sciences*, vol. 20, no. 1, pp. 18–31, 1980.

[33] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[34] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *Proceedings of the 2004 4th International Symposium on Chinese Spoken Language Processing (ISCSLP '04)*, pp. 285–288, IEEE, Hong Kong, December 2004.

[35] B.-W. Hwang, S. Kim, and S.-W. Lee, "A full-body gesture database for automatic gesture recognition," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR '06)*, pp. 243–248, IEEE, April 2006.

[36] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505–1518, 2003.

[37] A. Muro-de-la-Herran, B. García-Zapirain, and A. Méndez-Zorrilla, "Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 3801053, 11 pages
http://dx.doi.org/10.1155/2016/3801053

Research Article

Robust Automatic Target Recognition Algorithm for Large-Scene SAR Images and Its Adaptability Analysis on Speckle

Hongqiao Wang, Yanning Cai, Guangyuan Fu, and Shicheng Wang

Xi'an Research Institute of Hi-Tech, Xi'an 710025, China

Correspondence should be addressed to Hongqiao Wang; ep.hqwang@gmail.com

Received 17 July 2016; Revised 24 September 2016; Accepted 19 October 2016

Academic Editor: Xiong Luo

Aiming at the multiple target recognition problems in large-scene SAR image with strong speckle, a robust full-process method from target detection, feature extraction to target recognition is studied in this paper. By introducing a simple 8-neighborhood orthogonal basis, a local multiscale decomposition method from the center of gravity of the target is presented. Using this method, an image can be processed with a multilevel sampling filter and the target's multiscale features in eight directions and one low frequency filtering feature can be derived directly by the key pixels sampling. At the same time, a recognition algorithm organically integrating the local multiscale features and the multiscale wavelet kernel classifier is studied, which realizes the quick classification with robustness and high accuracy for multiclass image targets. The results of classification and adaptability analysis on speckle show that the robust algorithm is effective not only for the MSTAR (Moving and Stationary Target Automatic Recognition) target chips but also for the automatic target recognition of multiclass/multitarget in large-scene SAR image with strong speckle; meanwhile, the method has good robustness to target's rotation and scale transformation.

1. Introduction

Synthetic Aperture Radar (SAR) is an important sensor due to its all weather, day/night, high resolution imaging, and long standoff capability. Along with the development of radar technologies, as well as with increasing demands for target identification in radar applications, automatic target recognition (ATR) using SAR has become an important branch of image recognition.

Although there are many research findings in early SAR ATR field, most algorithms are based on the single target chips of the MSTAR dataset [1]. The MSTAR public dataset was provided by DARPA (Defense Advanced Research Project Agency)/AFRL (Air Force Research Laboratory). The MSTAR data is a standard dataset in the SAR ATR community, allowing researchers to fairly test and compare their ATR algorithms. The MSTAR data used in this paper consists of 128×128 pixel chips, which are 1 foot resolution, X-band, and three types of ground military vehicles, that is, BMP2, BTR70, and T72. Each chip has SAR images separated by 1° azimuth increments within an angular coverage from 0° to 360°. All the chips are taken at depression angles of 17° and 15°. As the MSTAR dataset gives the target chip directly, which are fixed in size and target position, the target segmentation and detection processes of the above-mentioned algorithms are omitted. Moreover, as there is only one target in each sample chip, the difficulty of target feature extraction is significantly reduced, which gives a low availability for practical applications. So, these algorithms are not the real sense of SAR image ATR.

In real applications, a common ATR of SAR images from the input of image to the output of recognition result can be divided into four stages, (I) image preprocessing, (II) feature processing, (III) classification, and (IV) postprocessing. In stage I, the main works include image filtering and denoising, distortion correction, image segmentation and target detection, regions of interest (ROI) discrimination, and image normalization. Stage II commonly contains the feature extraction, feature selection, feature dimension reduction, data clustering, and novelty detection. In stage III, the tasks

include the selection and design of classifier, the training of classifier, and target classification. The main work of postprocessing in stage IV is the further improvement for the early classification process; this stage critically focuses on the recognition precision and robustness, which are the chief properties of SAR ATR systems.

There are some special requirements to the SAR ATR algorithms, especially in the steps of feature extraction and the design of classifier. For example, the principle of the feature extraction method must be simple and can be easily realized with a real-time requirement; the feature has the advantages of antinoise and anticluster ability; and the feature has strong robustness for the translation, rotation, and scale transformation. On the other hand, the classifier must have high classification precision and learning efficiency.

By analyzing the literatures on SAR ATR [2], it can be found that the feature based classification methods get more and more attention than the whole image based methods. Many typical feature extraction methods have been used for SAR image classification, such as the PCA [3], the SDA [4], the shadow contour [5], multilinear subspace learning of tensor objects [6], the neighborhood geometric center scaling embedding method [7], feature selection [8], and feature sparsity [9]. In the classifier design and selection aspect, there are also some algorithms such as neural networks [10–12], support vector machine [13], and boosting [14].

For the common off-line recognition task, the above-mentioned methods maybe have some merits and have high recognition precision, but most of these methods need a large number of labeled samples to train an efficient classifier and cannot support the robust and full-process application of ATR. Whether in the realization difficulty aspect of feature extraction, or in the aspect of adaptability on speckle, there are still large gaps to the practical applications. In recent years, the event detection and event recounting to the unconstrained web videos have attracted a lot of attention; some new algorithms also have achieved state-of-the-art performances using the zero-example or limited supervision methods [15, 16]. But analyzing the image frame in the real-time videos, we can see that the targets are clear and have no noise, which is different from the SAR image ATR. The SAR images are usually fuzzy, whether the targets in the images or the backgrounds. Especially for the targets in SAR image, the inevitable speckle, which is a chaotic phenomenon that results from coherent summation of the backscattered signals, may cause great disturbance to the target features. So, the robustness of the algorithm on speckle directly decides the final recognition efficiency. As a result, the performance evaluation, especially the robustness and adaptability analysis of the algorithm, is very important work [17].

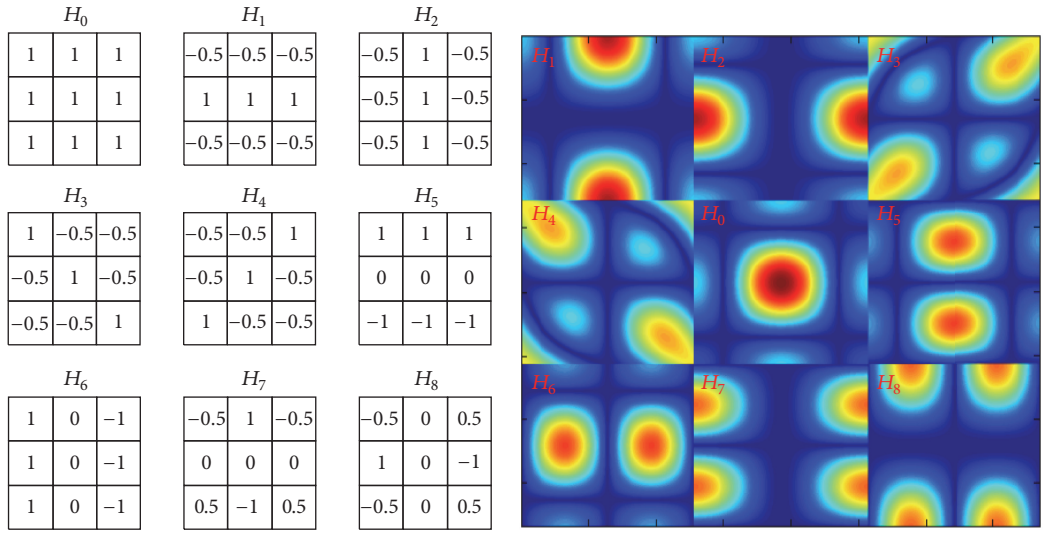
On this basis, a robust method from target detection, feature extraction to target recognition is studied, which can solve the multitarget ATR in large-scene SAR images effectively. The distributions of the method mainly include three aspects. Firstly, a robust local multiresolution analysis method for image target is presented, which brings a fast realization of feature extraction. Meanwhile, the dimension of the feature vector is relatively lower than some other methods with the similar performance. To acquire effective features, a multiscale analysis method from the center of gravity of the SAR image target is presented. By introducing a simple 8-neighborhood orthogonal basis, an image can be processed with a multilevel sampling filter, and then the multiscale features in eight directions and one low frequency filtering feature of the image can be achieved. Furthermore, the feature extraction method can be simply and rapidly realized and has good characterization performance. Secondly, aiming at the multiscale features, a multiple kernel classifier and the fusion between the multiscale features and the multiple kernels are studied, which brings higher classifier precision than the common single kernel support vector classifier. Comparing with the traditional method, the presented algorithm is far more advanced in fast detection of target and the dimension of feature vectors. Thirdly, the presented algorithm is a full-process image target recognition method, which can realize the steps from multiple targets detection, feature extraction to target recognition in large-scene SAR images, and has better practical application value. Analyzing the relevant references on SAR ATR, we can find that most target classification algorithms are only suitable for the step-by-step recognition and single target image cases, which is not the real SAR automatic target recognition. The robust algorithm is effective not only for the MSTAR target chips but also for the ATR of multiclass/multitarget in large-scene SAR image with strong speckle. Also, the method has strong robustness against the rotation and scale transformation.

In the remainder of this paper, we go along through different sections which are organized as follows: in Section 2, we introduce the local multiscale feature extraction method and design the multiscale wavelet kernel classifier. The robust target recognition method and adaptability analysis against speckle are studied in Section 3. In Section 4, several experiments are carried out to testify the effectiveness of the method proposed in this paper. Finally, we conclude in Section 5.

2. Feature Extraction and Design of Classifier

2.1. Feature Extraction. On SAR image targets (especially the vehicle targets owning some structural characters), without loss of generality, we may also take the MSTAR dataset as the object of investigation. The targets in MSTAR chips have the following characters: the sample images are the chips with the same size; there is only one target in one chip; the target lies in the center of the chip; the targets are distributed around the centers of the chips within an angular coverage between 0° and 360°; the chips have the same resolution and scale.

Most of the traditional multiresolution analysis methods are realized by filter and sampling, which can solve the generation problem of orthogonal basis effectively, but the sampling is commonly from the beginning to the end of series. These methods have disadvantages for feature extraction, because the obtained orthogonal basis may not give the similar expression for the data having the same local characteristic. Inspired by the sampling filter idea of the traditional wavelet method, a new sampling method using a local extension is adopted to solve this problem. With this method, each sampling process is extended from the local



(a) 8-neighborhood orthogonal basis (b) Frequency spectrum of the orthogonal basis

Figure 1: 8-neighborhood orthogonal basis and its frequency spectrum.

point to around, which can guarantee the generated basis pointing to the local region and can guarantee the same local characteristics having the similar projection coefficients on the basis. In the detailed implementation, we can begin from the local point of an image (e.g., the point could be a Point of Interest (POI) or a geometric center point); then several times of filtering process can be executed using the fast filter; namely, the multiscale decomposition of the image is gained, and the difference of Gaussian (DOG) like space image of the original image is also obtained. Finally, directly utilizing the key pixels sampling from the multilevel image, we can rapidly obtain the local multiscale feature of the image.

Through the above analyzing, we can firstly construct a multilevel DOG like scale space based on image. Then we design an 8-neighborhood orthogonal basis, with which the multilevel sampling filter on image can be executed. As a result, the image features in the 8 directions and a low frequency feature are derived. The structure of the 8-neighborhood orthogonal basis and its frequency spectrum are shown in Figure 1.

Aiming at the images in MSTAR dataset, the targets first can be detected using the constant false alarm rate (CFAR) method. In consideration of the targets only occupying the central location of the chips, for convenience, the central area with the size of 81×81 in each 128×128 chip is taken as the research target.

To get robust and effective features, a multiscale analysis method from the center of gravity of the SAR image target is presented. The above-mentioned 8-neighborhood orthogonal basis is also a template matrix; a convolution operation between the original image matrix and the template matrix can be executed using

$$\mathbf{g} = \mathbf{f} \otimes \mathbf{h},$$
$$\mathbf{g}(i,j) = \sum_{k,l} \mathbf{f}(i+k, j+l)\, \mathbf{h}(k,l), \tag{1}$$

where $\mathbf{g}$ is the image after sampling, $\mathbf{f}$ is the original image, and $\mathbf{h}$ is the template image. In this process, the obtained image can be seemed as a sampling filter from the original image. If the convolution is applied repeatedly, the pyramid-shaped multilevel sampling filter images is gained. Finally, the multiscale features in eight directions and one low frequency filtering feature can be achieved from the direct selection of key pixels in every level of the pyramid images. In this paper, the obtained image is processed by a 4-level local multiresolution decomposition. As for the feature extraction, we can directly choose the pixels in each level of the image as follows: in the highest level, the image size is 3×3; all the 9 pixels are chosen as the 9-dimension feature; in the second level, the 8 image blocks with the size of 3×3 corresponding to the 8 peripheral pixels in the highest level are chosen as the feature, so the feature dimension is 72; the feature extraction method in the third level is similar to the second level; the eight 3×3 image blocks corresponding to the block centers in the second level are selected; the feature dimension is also 72; in the fourth level, the peripheral 8 central pixels are directly selected, so the feature dimension is 8. So, for a given image, the total feature dimension is $9 + 72 + 72 + 8 = 161$.

2.2. Multiscale Wavelet Kernel Support Vector Classifier. Neural network based methods are widely used in hypersonic flight automatic control [18–20] and the tracking control in the Internet of Things [21], but SAR ATR is a high real-time application, and in most cases, enough SAR target images could not be collected. So, the support vector machine is chosen as the classifier, which is suitable for small samples and can be well modified for the online learning. As the common support vector classifier (SVC) has low speed owing to the solution of quadric programming, it is hard for them to apply to some real-time cases. Suykens and Vandewalle [22] proposed an improved SVM method, least squares support vector machine (LSSVM). By replacing the quadric

programming with a solution of linear equations, LSSVM has a great improvement on speed. According to the Structural Risk Minimization principle, the optimization problem of LSSVC is given as

$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + \frac{1}{2}C\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i - \boldsymbol{\omega}^{\mathrm{T}}\varphi(\mathbf{x}_i) - b = \xi_i, \quad i = 1, 2, \ldots, l, \end{aligned} \tag{2}$$

where $\boldsymbol{\omega}$ is the normal vector of the decision surface, C is the error penalty parameter, ξ_i is the error of the ith sample, $\mathbf{x}_i$ is the ith sample, y_i is the corresponding class label, and b is the bias term.

Transforming (2) to a nonrestricted optimization, the Lagrangian function can be defined as

$$L = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + \frac{1}{2}C\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i\left[\boldsymbol{\omega}^{\mathrm{T}}\varphi(\mathbf{x}_i) + b + \xi_i - y_i\right], \tag{3}$$

where α_i are Lagrangian multipliers. The optimality of upper function is as the following sets of linear equations instead of quadratic program in traditional SVC.

$$\begin{aligned} &\frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \longrightarrow \boldsymbol{\omega} = \sum_{i=1}^{l}\alpha_i\varphi(\mathbf{x}_i), \\ &\frac{\partial L}{\partial b} = 0 \longrightarrow \sum_{i=1}^{l}\alpha_i = 0, \\ &\frac{\partial L}{\partial \xi_i} = 0 \longrightarrow \alpha_i = C\xi_i, \\ &\frac{\partial L}{\partial \alpha_i} = 0 \longrightarrow \boldsymbol{\omega}^{\mathrm{T}}\varphi(\mathbf{x}_i) + b + \xi_i - y_i = 0. \end{aligned} \tag{4}$$

By eliminating variables $\boldsymbol{\omega}$ and ξ_i, we get the following equation:

$$\begin{bmatrix} 0 & \overrightarrow{\mathbf{1}}^{\mathrm{T}} \\ \overrightarrow{\mathbf{1}} & \mathbf{K} + C^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Y} \end{bmatrix}, \tag{5}$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_l)^{\mathrm{T}}$, $\overrightarrow{\mathbf{1}} = (1, 1, \ldots, 1)^{\mathrm{T}}$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l)^{\mathrm{T}}$, $\mathbf{I}$ is $l \times l$ identity matrix, $\mathbf{K}_{i,j} = \varphi(\mathbf{x}_i)^{\mathrm{T}}\varphi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$, and $k(\cdot, \cdot)$ is kernel function, $i, j = 1, 2, \ldots, l$.

Then solving the linear equations, we get the decision function of LSSVC as

$$f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i=1}^{l}\alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right]. \tag{6}$$

Set $\mathbf{H} = \mathbf{K} + C^{-1}\mathbf{I}$; we can get the classifier coefficient vector $\boldsymbol{\alpha}$ and the bias item b from (5)

$$b = \frac{\overrightarrow{\mathbf{1}}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{Y}}{\overrightarrow{\mathbf{1}}^{\mathrm{T}}\mathbf{H}^{-1}\overrightarrow{\mathbf{1}}}, \tag{7}$$

$$\boldsymbol{\alpha} = \mathbf{H}^{-1}\left(\mathbf{Y} - \frac{\overrightarrow{\mathbf{1}}\overrightarrow{\mathbf{1}}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{Y}}{\overrightarrow{\mathbf{1}}^{\mathrm{T}}\mathbf{H}^{-1}\overrightarrow{\mathbf{1}}}\right). \tag{8}$$

The fusion of kernels with multiple scales is a special situation of multiple kernel learning [23, 24]. This kernel method has better flexibility and can bring more completed scale choice than other methods, such as the multiscale kernel method. In addition, with the wavelet theory and the multiscale analysis theory continuing to mature, the multiple kernel method gains good theory background by introducing the scale space, which is a great promotion for kernel method based machine learning [25, 26].

The foundation of multiscale kernel method is seeking a set of kernel functions owning the multiscale representation capability. Among the kernel functions being widely used, the Gaussian radial basis function (RBF) (9) is the most popular one, because of its general approximation ability. Meanwhile, it is also a typical kernel and can be multiscaled

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right). \tag{9}$$

Taking the RBF kernel as the example, it can be multiscaled as (10) (suppose the generated kernels have the translation invariant)

$$\exp\left(\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma_1^2}\right), \ldots, \exp\left(\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma_m^2}\right), \tag{10}$$

where $\sigma_1 < \cdots < \sigma_m$. From (10), we can find that when σ is small, the support vector classifier (SVC) with the RBF kernel can fit the samples having drastic variation. And when σ is larger, the same classifier can well classify the samples with mild variation. So the multiscale kernels can obtain better generalization. Inspired by the scale-variant rule of wavelet transformation, the values of can be defined as

$$\sigma_i = 2^i\sigma, \quad i = 0, 1, 2, \ldots \tag{11}$$

Another typical multiscale kernel is the wavelet kernel function [27].

Theorem 1. *Let $h(x)$ be a mother wavelet function, a and c are the scaling factor and the transfer factor, $x, a, c \in \mathbb{R}$, if $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$, and then the inner product type wavelet kernel function can be expressed as*

$$k(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} h\left(\frac{x_i - c_i}{a}\right) h\left(\frac{z_i - c_i'}{a}\right), \tag{12}$$

and the transfer invariant wavelet kernel function is

$$k(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} h\left(\frac{x_i - z_i}{a}\right). \tag{13}$$

Theorem 2. *Considering a common wavelet function (14)*

$$h(x) = \cos(1.75x)\left(-\frac{x^2}{2}\right) \tag{14}$$

if $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$*, then the wavelet kernel function is*

$$k(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n} h\left(\frac{x_i - z_i}{a}\right) = \prod_{i=1}^{n} \left\{\cos\left[1.75\left(x_i - z_i\right)\right] \exp\left[-\frac{\left\|x_i - z_i\right\|}{\left(2a^2\right)}\right]\right\}. \quad (15)$$

By changing the values of the scale parameter a, different scale wavelet kernel functions can be constructed.

In this paper, the multiscale wavelet kernel is introduced to the LSSVC, and the multiscale wavelet kernel LSSVC is defined as

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} \alpha_i y_i \left\langle \phi_j(\mathbf{x}), \phi_j(\mathbf{x}_i)\right\rangle + b\right); \quad (16)$$

namely,

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} \alpha_i y_i K_j(\mathbf{x}, \mathbf{x}_i) + b\right). \quad (17)$$

On the other hand, it is also an effective approach for the improvement of target recognition accuracy if we synthesize the features having multiresolution character and the multiscale kernel functions. In this paper, the 4-level local multiscale feature and the 4-scale wavelet kernels are synthesized; the scales of corresponding kernel functions are increased by 2 times. At the same time, the weights of kernels are determined by equal coefficients; namely, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1/4$; the schematic diagram is shown in Figure 2.

3. Robust ATR Complexity Analysis and Adaptability Analysis

3.1. Robust ATR Method and Complexity Analysis. The robust ATR procedure includes two stages which are the multiscale kernel classifier training and the recognition test; the schematic diagram is shown in Figure 3.

The multiscale kernel classifier training is mainly based on the MSTAR dataset; the steps are as follows.

Step 1. Finish the CFAR detection, respectively, for all the training target chips and obtain the target segmentation results.

Step 2. From the centers of the chips (namely, the centers of targets), execute the 4-level local multiscale decomposition for the target images.

Step 3. Encode the coefficients from the multiscale analysis, and extract the feature vectors of every target.

Step 4. Train the multiscale wavelet kernel classifier using the multilevel feature vectors.

The recognition test is based on the large-scene multiple targets image samples acquired in real-time; the steps are as follows.

Step 1. Do the CFAR detection for the large-scene image and segment the targets and regions of interest (ROI).

Step 2. Process the targets with the mathematical morphological, eliminate the false alert, and calculate the gravity center of each target, which is taken as the starting point of local multiscale decomposition.

Step 3. Conduct the 4-level local multiscale decomposition from the gravity center of each target.

Step 4. Extract the features of each target by sampling and construct the feature vector.

Step 5. Input the feature vectors to the multiscale wavelet kernel classifier, and output the recognition result.

In pattern recognition application using SVM, the training and testing are two different processes, so the algorithm complexity should not be understood as a whole. Here we mainly discuss the complexity in the training process, namely, the complexity solving of the quadratic programming problem. For a typical SVM training algorithm, its computational complexity is in the range of $\mathbf{O}(\mathbf{N_{sv}}^3 + \mathbf{LN_{sv}}^2 + \mathbf{dLN_{sv}}) \sim \mathbf{O}(\mathbf{dL}^2)$, where $\mathbf{N_{sv}}$ is the number of support vectors, $\mathbf{L}$ is the number of samples in training set, and $\mathbf{d}$ is the dimension of each sample, so in the worst case, the algorithm complexity is $\mathbf{O}(\mathbf{dL}^2)$. In our algorithm of this paper, the number of training samples $\mathbf{L} = 1622$; the sample dimension $\mathbf{d} = 161$. Considering the multiple kernel classifier, we introduced $\mathbf{m} = 4$ kernels to the algorithm, $\mathbf{m} \ll \mathbf{d}$ and $\mathbf{m} \ll \mathbf{L}$, and we still can think that the complexity of the multiple kernel training is $\mathbf{O}(\mathbf{dL}^2)$. So, we can conclude that the computational complexity in the multiple kernel classifier training process is $\mathbf{O}(\mathbf{dL}^2)$.

In the large-scene multiple targets ATR process, the complexity is mainly dependent on the kernel function computation between the testing sample vector and the support vectors. In this process, the complexity of each target sample is $\mathbf{O}(\mathbf{dN_{sv}})$. In the large-scene SAR image, suppose the number of the detected targets is $\mathbf{t}$, then the complexity of the whole testing may be $\mathbf{O}(\mathbf{tdN_{sv}})$. In practical applications, the value $\mathbf{t}$ usually is small; for example, in the experiments of adaptability analysis in our manuscript, there are 6 targets in the images. So we still can conclude that the complexity of the testing process is $\mathbf{O}(\mathbf{dN_{sv}})$.

3.2. Adaptability Analysis Method. The speckle may cause inconvenience to SAR image ATR, but this noise is inevitable and cannot be absolutely eliminated. So, the adaptability against speckle of the target recognition algorithm directly decides the usability and robustness. On the adaptability analysis against speckle, the main method is adding speckle into the large-scene multiple target image and then analyzing the recognition precision. When the speckle adding degree (SAD) is plus 1, the speckle is added into the whole image with mean 0 and variance 0.04. On this basis, some parameters such as the mean, the variance, the dynamic range, and the peak signal noise ratio of the image can be calculated. Then

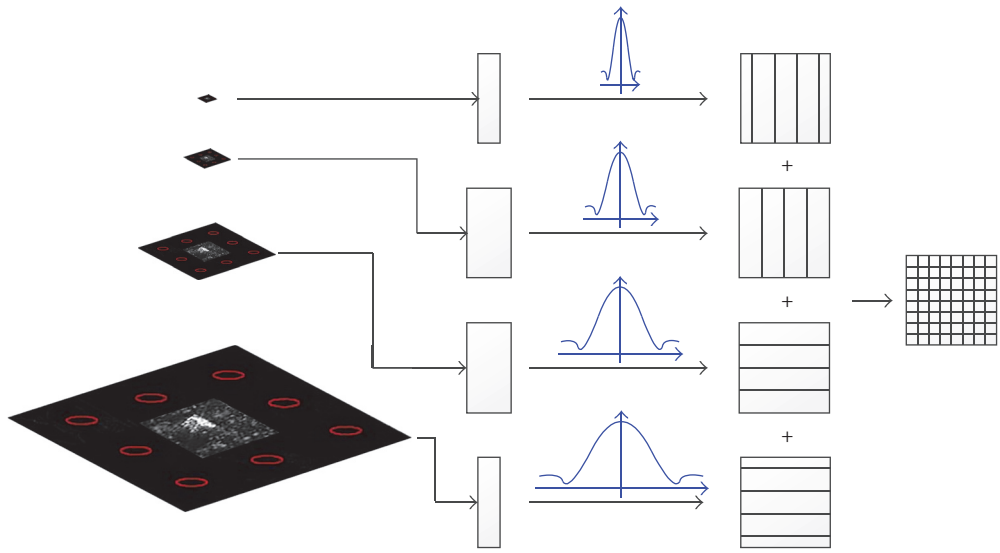


Figure 2: Synthesis schematic diagram of 4-level local multiscale feature and 4-scale wavelet kernel.

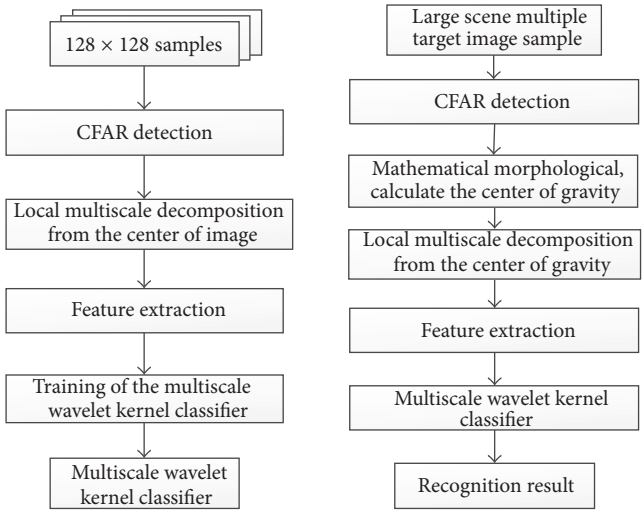


Figure 3: Flow diagram of multiscale wavelet kernel classifier training and large-scene multiple target ATR.

for the image with added speckle, through executing the target detection, feature extraction, and classification again, we can study the target detection and recognition precision at different speckle degree.

4. Experiments

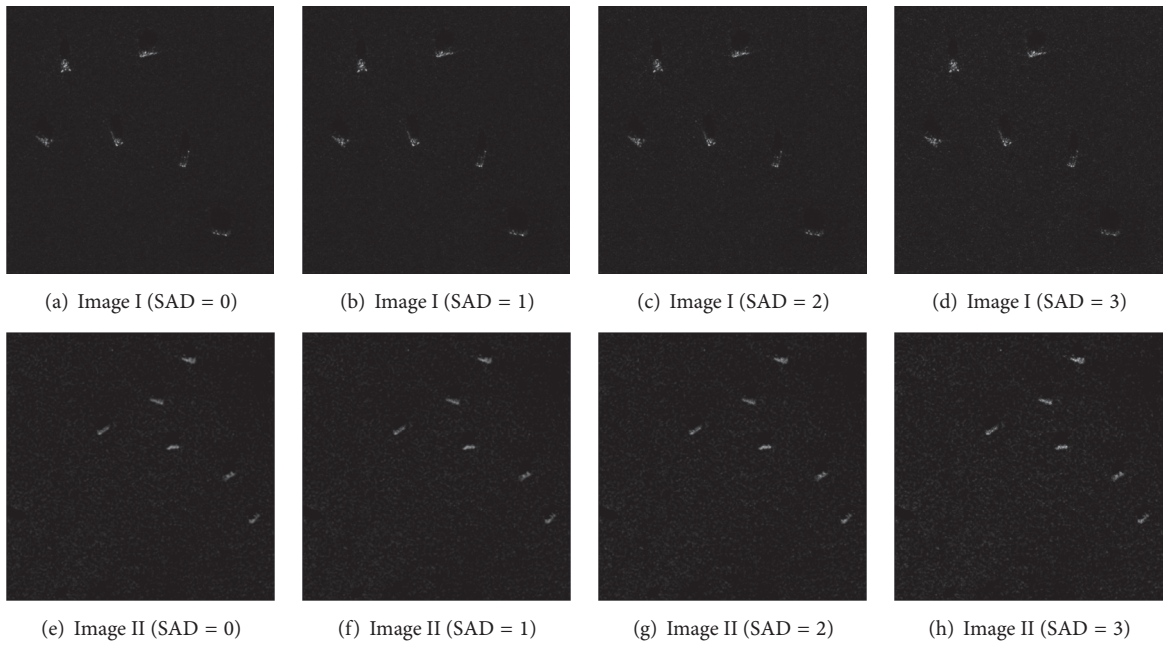
4.1. Single Target Recognition and Adaptability Analysis for Speckle. The single target recognition and speckle adaptability analysis are based on the MSTAR dataset; the numbers of the 3 classes samples are shown in Table 1. After feature extraction with local multiscale decomposition for all the sample chips, we design and train the classifier. Then the multiple-class classification problem is transformed into the two-class problem by "One VS One" method.

Table 1: MSTAR dataset.

Target type	Training set		Testing set	
	File name	Number	File name	Number
T72	SN_132	232	SN_132	196
	SN_812	231	SN_812	195
	SN_S7	228	SN_S7	191
BMP2	SN_9563	233	SN_9563	195
	SN_9566	232	SN_9566	196
	SN_C21	233	SN_C21	196
BTR70	SN_C71	233	SN_C71	196
SUM		1622		1365

Utilizing the training set, we also can obtain the optimal multiscale wavelet kernel classifier, where the scale factors of wavelet kernel are $a_1 = 800$, $a_2 = 400$, $a_3 = 200$, and $a_4 = 100$. The penalty coefficient $C = \{1, 10, 100, 1000\}$. After the features of the testing set being extracted, the feature vectors are sent to the classifier and the recognition precision is outputted. To analysis the algorithm performance in a nondistortion circumstance, some experiments and performance comparison between the proposed method and other typical methods are executed. The feature extraction methods, the feature dimension, the classifiers, and the classification precision are shown in Table 2.

From the experimental result, we can see that the fusion method with the multiscale feature and the multiscale kernel classifier gives a very high classification precision of 98.75% when there is no speckle added. In addition, the algorithm realizes the fast access and storage to nearly 3000 SAR images in a short time, which indicates good real-time performance. Comparing with the traditional method, the presented algorithm is far more advanced in fast detection of target and less dimension of feature vectors.



(a) Image I (SAD = 0) (b) Image I (SAD = 1) (c) Image I (SAD = 2) (d) Image I (SAD = 3)

(e) Image II (SAD = 0) (f) Image II (SAD = 1) (g) Image II (SAD = 2) (h) Image II (SAD = 3)

Figure 4: Result images with different speckle adding degrees.

Table 2: ATR result of different feature extraction methods and classifier.

Feature extraction methods	Dimension of features	Classifier	Recognition rate
Hu's moment	7	LSSVC	0.7313
Wavelet moment	7	LSSVC	0.8469
	17	LSSVC	0.9184
Wavelet moment + entropy	7	LSSVC	0.8776
	17	LSSVC	0.9660
PCA	196	LSSVC	0.9643
Local multiscale feature	161	LSSVC	0.9802
Local multiscale feature	161	Multiscale kernel classifier	0.9897

To analyze the adaptability against noise, the speckle with mean 0 and variance 0.04 is added to the MSTAR testing set. When SAD is plus 1, the speckle is added into the image one time. Through the comparison experiments under different SADs, the final recognition results are shown in Table 3.

When the speckle is added into the testing samples, the recognition precision comes to 93.41% with SAD = 1. As the enhancement of speckle, the recognition precisions reduce to 87.62% and 76.48% when SAD = 2 and SAD = 3, respectively, which are still the preferably correct ratios. The reasons lie in that there is only one target in each sample, and we already know the targets lie in the center of sample chips. So, even though the target structure changes, we can still do the local multiscale decomposition from the center of sample and extract the exactly proper features.

Table 3: Single target ATR result under different speckle adding degree.

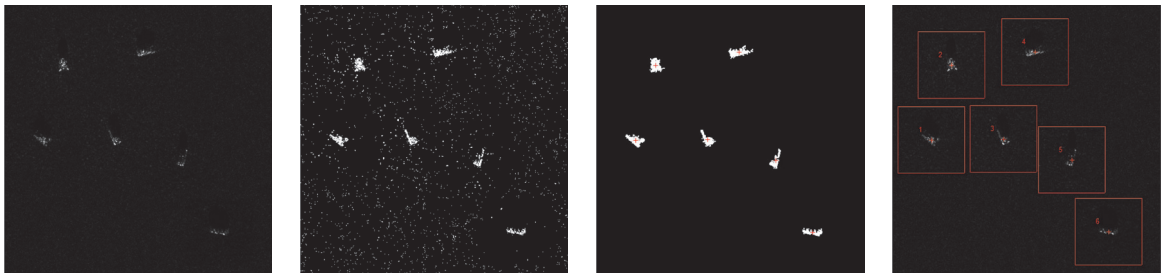
Speckle adding degree (SAD)	Number of testing sets	Correct recognition number	Recognition precision
SAD = 0 (no adding speckle)	1365	1348	98.75%
SAD = 1	1365	1275	93.41%
SAD = 2	1365	1196	87.62%
SAD = 3	1365	1044	76.48%

4.2. Large-Scene Multiple Target ATR and Adaptability Analysis for Speckle. For large-scene multiple target ATR, firstly, we must construct the large-scene multiple target images. Here, we randomly select 6 targets being correctly classified in the MSTAR testing set, where 2 targets are selected from each class, respectively, and the targets are embedded in the large-scene clutter images. Using this method, two large-scene multiple target images are formed, where Image I has a size of 512×512, with the image parameters of mean $\mu = 10.82$, variance $\sigma = 75.30$, and the dynamic range $D = 64.07$. Image II has a size of 768×768 and $\mu = 9.08$, $\sigma = 118.51$, and $D = 62.25$, respectively.

In the simulation tests, 3 degrees of speckle is added into the two large-scene multiple target images; the result images with different SADs are shown in Figure 4. Then using the same target segmentation, mathematical morphological processing, and center of gravity calculation, we can achieve the target detection and marking result. Figures 5 and 6 show the two-image target segmentation, detection, and marking result under the speckle adding degree 1 (SAD = 1). Then, begin from the center of gravity, execute the multiscale

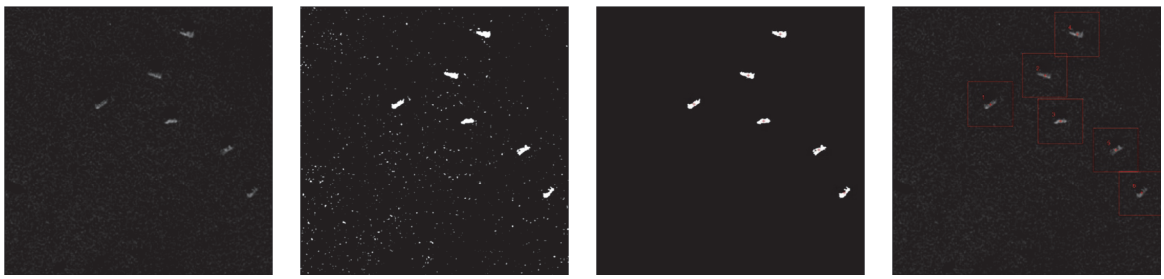
Table 4: Large-scene image parameters and multiple targets ATR result under different SADs.

Large-scene images	Speckle adding degree (SAD)	Image parameters μ, σ, D, PSNR	Number of detected targets	Number of correctly recognized targets
Image I	SAD = 1	10.96, 107.42, 65.43, 37.78	6	6
	SAD = 2	11.50, 146.34, 66.45, 31.72	6	5
	SAD = 3	12.59, 232.78, 67.88, 27.66	6	4
Image II	SAD = 1	8.88, 122.70, 63.08, 38.31	6	6
	SAD = 2	9.31, 162.29, 64.54, 32.24	6	5
	SAD = 3	10.21, 244.97, 66.15, 28.17	6	5



(a) Original image (SAD = 1) (b) CFAR detection and segmentation result (c) Morphological processing and gravity center location (d) Target detection and marking result

Figure 5: Multiple targets detection and marking result of the large-scene sample Image I.



(a) Original image (SAD = 1) (b) CFAR detection and segmentation result (c) Morphological processing and gravity center location (d) Target detection and marking result

Figure 6: Multiple targets detection and marking result of the large-scene sample Image II.

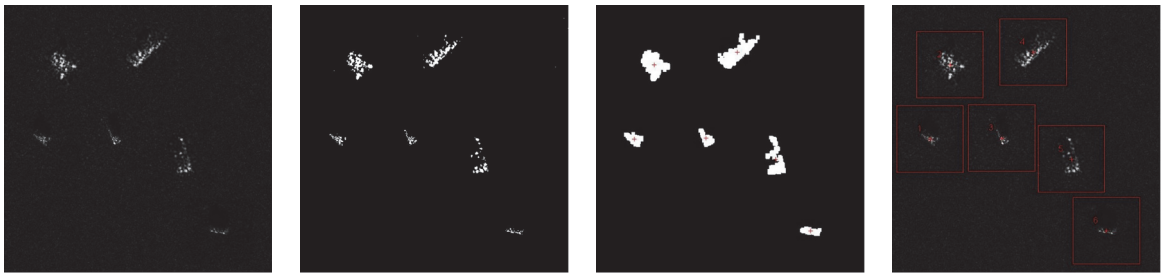
decomposition and feature extraction, and gain the feature vectors. Finally, send the feature vectors into the multiscale wavelet kernel classifier, and output the recognition results.

The large-scene multitarget image parameters and the ATR results under different SADs are recorded in Table 4. The experimental results indicate that when the SAD = 1, although the center of gravity of target is shifted from the geometrical center of the chips after target segmentation, the 6 targets still can be detected and correctly classified, which testifies the effectiveness and robustness of the presented algorithm. With the enhancement of speckle, the number of correctly recognized targets is decreased. The targets marked "6" and marked "1" in Image I are error recognized with the SAD = 2 and SAD = 3. In Image II, only the target marked "6" is error recognized with the two SADs; the rest 5 targets are correctly recognized. The main reason is the multiplicative characteristic of speckle, which can cause drastic variability to the target structure with target segmentation and mathematical morphological processing, so the center of gravity of target shifts dramatically. As a result, the local multiscale features are influenced. Overall, the feature extraction and classification method has good adaptability on speckle.

4.3. ATR with Scale and Rotation Transformation and Adaptability Analysis for Speckle. In this test, the rotation and scale transformations of the targets are introduced based on Image I and Image II in Section 4.2. For Image I, the targets marked "4," "2," and "5" are resized with scale parameters 2, 1.5, and 1.5, respectively, and with a rotation of 30°; then the targets are embedded into the original image and the new image named Image III is constructed. For Image II, the targets marked "1," "4," and "5" are resized with scale parameters 2, 1.5, and

Table 5: Large-scene image parameters and multiple targets ATR result with rotation and scale transformation under different SADs.

Large-scene images	Speckle adding degree (SAD)	Image parameters μ, σ, D, PSNR	Target marks	Correctly recognized target marks
Image III	SAD = 1	11.68, 180.76, 65.45, 36.44	2, 4, 5	2, 4, 5
	SAD = 2	12.26, 238.52, 66.85, 30.37	2, 4, 5	2, 4
	SAD = 3	13.43, 364.02, 68.33, 26.27	2, 4, 5	—
Image IV	SAD = 1	9.38, 157.41, 63.79, 37.45	1, 4, 5	1, 4, 5
	SAD = 2	9.84, 206.92, 64.89, 31.34	1, 4, 5	1, 4
	SAD = 3	10.80, 310.20, 66.55, 27.29	1, 4, 5	4



(a) Original image (SAD = 1) (b) CFAR detection and segmentation result (c) Morphological processing and gravity center location (d) Target detection and marking result

Figure 7: Multiple targets detection and marking result of the large-scene sample Image III with rotation and scale transformation.

(a) Original image (SAD = 1) (b) CFAR detection and segmentation result (c) Morphological processing and gravity center location (d) Target detection and marking result

Figure 8: Multiple targets detection and marking result of the large-scene sample Image IV with rotation and scale transformation.

1.5, respectively, and with a rotation of 30°; using the same method, Image IV is constructed.

In the tests, 3 degrees of speckle is added into Image III and Image IV. Then utilizing the same target segmentation, mathematical morphological processing with modulation of parameters, and center of gravity calculation, we can achieve the target detection and marking result. Figures 7 and 8 show the two-image target segmentation, detection, and marking result under the SAD = 1. Then, begin from the center of gravity, execute the multiscale decomposition and feature extraction, and gain the feature vectors. Finally, send the feature vectors into the multiscale wavelet kernel classifier, and output the recognition results.

Only considering the targets with rotation and scale transformation, the large-scene multitarget image parameters and the ATR results under different SADs are recorded in Table 5. Under the condition of SAD = 1, the 3 targets, respectively, in Image III and Image IV are still correctly recognized, which testifies the effectiveness of the feature extraction method and the robustness on rotation and scale transformation. But with the enhancement of the speckle, the recognition rate significantly decreased. For example, when SAD = 2, only 2 targets are correctly recognized, respectively, in Image III and Image IV; when SAD = 3, all 3 targets in Image III are error recognized, and only 1 target in Image IV can be correctly recognized. It is shown that the structure features of the targets are strongly changed after the rotation and scale; at the same time, the speckle has a far greater impact on large scale targets. As a result, the local multiscale features are extremely influenced, which leads to the error recognition of the targets.

To further testify the performance of the algorithm under rotation, scale transformation, and speckle, more large-scene image samples are constructed and tested. Based on Image I

Table 6: ATR results with more large-scene image samples under rotation and scale transformation.

Large-scene image datasets	Number of image samples	Number of targets	Number of correctly recognized targets	Recognition rate
Dataset I	12	72	63	87.5%
Dataset II	12	72	65	90.3%

and Image II, the targets in the two images are resized with scale parameter 1.5 firstly; then the targets are rotated at 30°, and finally, the speckle with SAD = 1 is added into the images. Ultimately, two datasets having 12 large-scene images, respectively, can be constructed from Image I and Image II. In each dataset, there are 72 targets in total with 3 classes.

The experiments are carried out with the two datasets, and the results are shown in Table 6. From the data in the table, we can find that the correct recognition rates of the three classes targets are 87.5% for Dataset I and 90.3% for Dataset II. The results once again indicate that the local multiscale feature has good adaptability on rotation, scale transformation, and speckle. Meanwhile, the fusion of local multiscale feature and the multiscale kernel classifier can bring better robustness and recognition rate for ATR systems.

5. Conclusions

Compared with the other image target recognition such as the face recognition, the gesture recognition, the fingerprint recognition, and the gait recognition, great obstacles are brought to the usability and recognition efficiency of SAR ATR as the strong speckle and low image resolution. Based on the fusion of multiscale feature and multiscale kernel machine, a robust full-process method from target detection, gravity center locating, local multiscale decomposition, and feature extraction to target classification with multiscale kernel LSSVC is studied, which can solve the multitarget ATR in large-scene SAR images with strong speckle effectively. Through adaptability analysis, the robust algorithm is testified having good adaptability on speckle. Meanwhile, the algorithm is well suitable for the requirement of practical applications. As you can imagine, the method can be applied to the vehicle target detection from the other imaging sensors such as visible light image and the infrared image; it also can be used for ship recognition in SAR images with complex sea clutters. On the other hand, to achieve better classification precision, a large number of labeled samples are needed to train an effective classifier for the support vector machines including the classifier presented in this paper. But the labeled sample acquisition is costly, laborious, and time-consuming; how to improve the classifier accuracy using unlabeled data has received considerable attention in medical applications and more recently in crowdsourcing and the event detection and event recounting problems to real video datasets [28]. So, in the next step, we look forward to making the significant improvement of performance and efficient implementation from this idea in SAR ATR.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was jointly supported by the National Natural Science Foundation for Young Scientists of China (Grant nos. 61202332, 61403397, and 61503389), China Postdoctoral Science Foundation (Grant no. 2012M521905), and Natural Science Basic Research Plan in Shaanxi Province of China (Grant no. 2015JM6313).

References

[1] R. Patnaik and D. Casasent, "MINACE filter classification algorithms for ATR using MSTAR data. Defense and security," *Proceedings of the SPIE-International Society for Optics and Photonics*, vol. 5807, pp. 100–111, 2005.

[2] S. Fukuda and H. Hirosawa, "Support vector machine classification of land cover: application to polarimetric SAR data," in *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS '01)*, pp. 187–189, July 2001.

[3] X. Liu, Y. L. Huang, J. F. Pei, and J. Y. Yang, "2DPCA-based two-dimensional marginal sample discriminant embedding for SAR ATR," in *Proceedings of the 33rd IEEE International Geoscience and Remote Sensing Symposium (IGARSS '13)*, pp. 2023–2026, Melbourne, Australia, July 2013.

[4] X. Liu, Y. L. Huang, J. F. Pei, and J. Y. Yang, "Sample discriminant analysis for SAR ATR," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 12, pp. 2120–2124, 2014.

[5] K. Yin, L. Jin, C. C. Zhang, and Y. F. Guo, "A method for automatic target recognition using shadow contour of SAR image," *IETE Technical Review*, vol. 30, no. 4, pp. 313–323, 2013.

[6] C. Liu, J. J. Yin, J. Yang, and W. Gao, "Classification of multi-frequency polarimetric SAR images based on multi-linear subspace learning of tensor objects," *Remote Sensing*, vol. 7, no. 7, pp. 9253–9268, 2015.

[7] Y. Huang, J. Peia, J. Yanga, B. Wang, and X. Liu, "Neighborhood geometric center scaling embedding for SAR ATR," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 1, pp. 180–192, 2014.

[8] J.-I. Park, S.-H. Park, and K.-T. Kim, "New discrimination features for SAR automatic target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 476–480, 2013.

[9] V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Sparsity-motivated automatic target recognition," *Applied Optics*, vol. 50, no. 10, pp. 1425–1433, 2011.

[10] M. G. Gong, J. J. Zhao, J. Liu, Q. G. Miao, and L. C. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 1, pp. 125–138, 2016.

[11] J. R. David, *Applications of artificial networks to synthetic aperture radar for feature extraction in noisy environments [Dissertation]*, California Polytechnic State University, 2013.

[12] D. A. E. Morgan, "Deep convolutional neural networks for ATR from SAR imagery," in *Algorithms for Synthetic Aperture Radar Imagery XXII*, vol. 9475 of *Proceedings of SPIE*, Baltimore, Md, USA, 2015.

[13] Q. Zhao and J. C. Principe, "Support vector machines for SAR automatic target recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 643–654, 2001.

[14] Y. Sun, Z. P. Liu, S. Todorovic, and J. N. Li, "Adaptive boosting for SAR automatic target recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 1, pp. 112–125, 2007.

[15] X. J. Chang, Y. Yang, G. D. Long, C. Q. Zhang, and A. G. Hauptmann, "Dynamic concept composition for zero-example event detection," in *Proceedings of the AAAI*, 2016.

[16] X. Chang, Y. L. Yu, Y. Yang, and A. G. Hauptmann, "Searching persuasively: joint event detection and evidence recounting with limited supervision," in *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*, pp. 581–590, ACM, 2015.

[17] G. Marino, *Analysis of performance of automatic target recognition systems [Ph.D. thesis]*, Cranfield University, 2012.

[18] B. Xu, Q. Zhang, and Y. P. Pan, "Neural network based dynamic surface control of hypersonic flight dynamics using small-gain theorem," *Neurocomputing*, vol. 173, pp. 690–699, 2016.

[19] B. Xu, Y. H. Fan, and S. M. Zhang, "Minimal-learning-parameter technique based adaptive neural control of hypersonic flight dynamics without back-stepping," *Neurocomputing*, vol. 164, pp. 201–209, 2015.

[20] X. Luo and J. Li, "Fuzzy dynamic characteristic model based attitude control of hypersonic vehicle in gliding phase," *Science China. Information Sciences*, vol. 54, no. 3, pp. 448–459, 2011.

[21] X. Luo, Y. X. Lv, M. Zhou, W. P. Wang, and W. B. Zhao, "A laguerre neural network-based ADP learning scheme with its application to tracking control in the Internet of Things," *Personal and Ubiquitous Computing*, vol. 20, no. 3, pp. 361–372, 2016.

[22] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[23] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *The Journal of Machine Learning Research*, vol. 9, no. 6, pp. 1179–1225, 2008.

[24] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[25] N. Kingsbury, D. B. H. Tay, and M. Palaniswami, "Multi-scale kernel methods for classification," in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pp. 43–48, Washington, DC, USA, September 2005.

[26] H. Q. Wang, F. C. Sun, Y. N. Cai, N. Chen, and L. G. Ding, "Multiple kernel learning methods," *Acta Automatica Sinica*, vol. 36, no. 8, pp. 1037–1050, 2010.

[27] L. Zhang, W. D. Zhou, and L. Jiao, "Wavelet support vector machine," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 34–39, 2004.

[28] X. J. Chang, Y. L. Yu, Y. Yang, and E. P. Xing, "They are not equally reliable: semantic event search using differentiated concept classifiers," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '16)*, 2016.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 5130603, 10 pages
http://dx.doi.org/10.1155/2016/5130603

Research Article

Text Summarization Using FrameNet-Based Semantic Graph Model

Xu Han,[1,2] Tao Lv,[1,2] Zhirui Hu,[3] Xinyan Wang,[4] and Cong Wang[1,2]

[1]*School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3]*Department of Statistics, Harvard University, Cambridge, MA, USA*
[4]*Air Force General Hospital, Beijing, China*

Correspondence should be addressed to Xinyan Wang; wangxinyan@china.com

Received 8 August 2016; Accepted 30 October 2016

Academic Editor: Xiong Luo

Text summarization is to generate a condensed version of the original document. The major issues for text summarization are eliminating redundant information, identifying important difference among documents, and recovering the informative content. This paper proposes a Semantic Graph Model which exploits the semantic information of sentence using FSGM. FSGM treats sentences as vertexes while the semantic relationship as the edges. It uses FrameNet and word embedding to calculate the similarity of sentences. This method assigns weight to both sentence nodes and edges. After all, it proposes an improved method to rank these sentences, considering both internal and external information. The experimental results show that the applicability of the model to summarize text is feasible and effective.

1. Introduction

With the era of big data, text resources are becoming more and more abundant. Natural Language Processing (NLP) techniques have developed rapidly. Text summarization is a research tool that has widely applied in NLP. It gives a summary which can help us understand the whole article immediately with least words. Text summarization is to generate a condensed version of the original documents by reducing documents in size while retaining the main characteristics [1]. But artificial text summarization needs much background knowledge and often requires long processing time, while qualities of summaries are not always good enough. For this reason, people began to focus on automatic text summarization. Automatic text summarization was first studied almost 60 years ago by Luhn [2] and has gained much attention in recent years. This method is faster than artificial ones and performs better than the average level. With the rapid growth of text resources online, various domains of text summarization are applied. For instance, a Question Answering (QA) System produces a question-based summary to offer information. Another example is the search result snippets in web search engine which can assist users to explore more [3]; short summaries for News Articles can help readers to obtain useful information about an event or a topic [4]; speech summarization automatically selects indicative sentences from original spoken document to form a concise summary [5].

Automatic text summarization is mainly facing the following two problems: one is to reduce redundancy and the other is to provide more information with less words. Sentences in-summary are those which can stand for parts of the whole article. So the repeated information should be reduced, while the main information should be maintained. The major issues for text summarization are as follows. To begin with, information included in text is often redundant, so it is crucial to develop a method to eliminate redundancy. It is very common that different words are used to describe the same object in a text. For that reason, naive similarity measures between words cannot faithfully describe the content

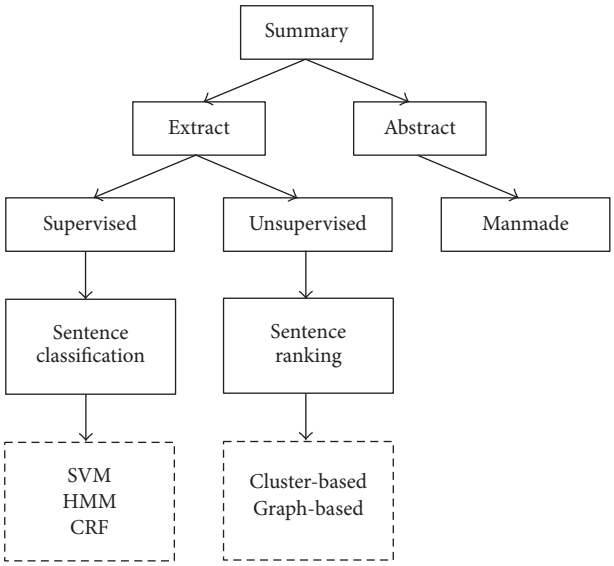


Figure 1: Summarization Approach Categories.

similarity. Another issue is identifying important difference among documents and covering the informative content as much as possible [6].

In this paper, to handle these issues, it proposes a Semantic Graph Model using FrameNet (FSGM) for text summarization that exploits the semantic information of sentence. FSGM treats sentences as vertexes while the semantic relationship between sentences as the edges. Firstly, it uses FrameNet to classify all sentences and then calculates the relevance values between sentences using semantic analysis and takes the values as weights of edges. Word embedding is used to combine similar sentences more accurately. Then based on the FSGM, sentences' values are scored by a variant of the improved PageRank graph ranking algorithm [7], considering both internal and external information. The experiments have been conducted on the DUC2004 data sets, and the results show that our model improves graph-based sentence extraction algorithms under various settings. So it can be used to enhance the performance. The experimental results show our model can be used for sentence extraction effectively.

The rest of the paper is organized as follows. Sections 2 and 3 briefly review the related work of current approaches about text summarization and the prepare work for experiments. The proposed Semantic Graph Model using FrameNet and word embedding is presented in Section 4 in detail. Section 5 is about the experiment, results, and relevant discussion. Finally, conclusion for this paper is in Section 6.

2. Related Work

Figure 1 shows the main summarization approaches [8]. There are two categories of summarization: abstract-based and extract-based. The abstractive summarization methods mostly generate the summary by compressing and reformulating the sentence. By this way, summary can keep its original length but has more effective information. However, the method requires complex linguistic processing, while the extract-based summarization measures various statistical significance for the words to locate the most central sentences in documents [9].

Extract-based approaches can be further categorized by supervised and unsupervised approaches. Supervised method proposes summarizing text as a classification problem in which all the sentences can be divided into either in-summary or not-in-summary. These methods are based on an essential assumption that each sentence should be independent. Naturally, it ignores the dependence between sentences. And then, there are additional approaches to build the relationship between sentences, such as using Hidden Markov Model (HMM) or Conditional Random Field (CRF).

Unsupervised method is to generate summaries by ranking sentences. Cluster-based approaches can classify sentences into groups and then choose the important ones in each group to form the final summary. Radev et al. [10] present MEAD with cluster centroids leverage. MEAD makes score for sentences by various features, assigning a representative centroid-based approach. Sentence-level semantic analysis (SLSS) and Symmetric Non-Negative Matrix Factorization (SNMF) are used in summarization by Wang et al. [11]. Mei and Chen [12] adapt fuzzy medoid-based clustering approach to produce subsets of similar sentences. Moreover, Latent Semantic Analysis (LSA), which helps adding and analyzing hidden topic features, is employed by Gong and Liu's work [13]. Wang et al. [14] generate summaries via discriminative sentence selection. He et al. [15] consider sentences in documents as a kind of signals from the perspective of compressive sensing. However, cluster-based approaches only considered the relationship between sentences, which may cause the semantic gap.

Another approach of unsupervised extract-based summarization [11, 16] uses graph-based model. TextRank [17] and LexRank [18] are first two graph-based models applied in text summarization, which use the PageRank-like algorithms to mark sentences. Then, other researchers have integrated the statistical and linguistic features to drive the sentence selection process, for example, the sentence position [19], term frequency [20], topic signature [21], lexical chains [22], and syntactic patterns [7, 23]. Ko and Seo [24] composed two sentences nearby into a bigram. Those bigrams were supposed to be context information. First, they extracted the bigrams by using the sentence extraction model. Then they used another extraction module to extract sentences from them. The ClusterCMRW and ClusterHITS models calculated the sentences scores by considering the cluster-level information in the graph-based ranking algorithm. Canhasi and Kononenko [25] improve matrix decomposition by employing the archetypal analysis for generic multidocument summarization. While coming to the document set, there must be more consideration about the document-level influence. But it did not consider the relationship between words and sentences. The DsR model [26] achieved it by using document-sensitive graph-based ranking model. But this method did not get a satisfied result. Yin et al. improved the summarization quality by adding extra information which

came from the query-extraction scenario. Goyal et al. [27] take Bernoulli model of randomness to index weights of sentences taking the context into consideration. The method proposed in [28] decomposed sentences by semantic role analysis, but while building the model, it did not use graph-based algorithms.

However, most of these graph-based methods only consider the relation of keyword cooccurrence, without considering the sentence-level dependency syntax. Those papers which use semantic information do not utilize the semantic information in the sentence-level. Thus, how to take advantage of the relationship between sentences needs further research. In this paper, it proposes sentence-level Semantic Graph Model. FSGM can build the relationships between sentences in a better way. Several experiments show ideal results in our model.

3. Preliminaries

In this section, it introduces related works for the experiments: firstly, the graph sorting algorithm, then the FrameNet, and finally word embedding.

3.1. Graph Sorting Algorithm. Graph sorting algorithm can calculate the nodes importance based on the graph structure. PageRank is one of its classical algorithms. The basic idea concerns the webpages' weight impacted both by usual links and reverse links, which means the more reverse links are, the heavier webpages' weight is. It builds graphs by using links structure, while the links are edges and webpages are nodes. The formula is as follows:

$$\mathrm{PR}\left(p_i\right) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{\mathrm{PR}\left(p_j\right)}{L\left(p_j\right)}, \tag{1}$$

where $p_1, p_2, \ldots, p_n$ are all nodes, $M(p_i)$ is a set which connect to p_i, $L(p_i)$ is p_i out-degree, N is the total number of nodes, and d is damping coefficient, usually set by 0.85.

In automatic text summarization, the graph is built by Natural Language Processing, text units are nodes, and the relationship of text units is edges. These relevancies often have its meaning, so the graph is basically unweighted undirected graph.

In undirected graph, every node has same in-degree and out-degree. In sparse condition, the convergence speed in undirected graph is faster than directed graph. The stronger the graph can connect, the faster it can converge. If the graph connectivity is strong enough, the convergence curves undirected and directed are nearly overlapped [29].

In weighted graph, the weight between nodes can be used; the formula is as follows:

$$\mathrm{PR}\left(p_i\right) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{w_{ij}\mathrm{PR}\left(p_j\right)}{L\left(p_j\right)}. \tag{2}$$

In graph sorting algorithm, the relevance between nodes is very important. In LexRank, if Cosine similarity between sentences is larger than the threshold, there is an edge. TextRank has another weight parameter comparing with PageRank. It gives edges weights and builds weighted graph. These two methods are simple and ignore some effective factors, such as semantic information and linguistic knowledge. Above all [30, 31], these methods are all considered more similar to judge the relevance between nodes. Corresponding solutions of these limitations are proposed in this paper. It used FrameNet combining the word embedding, to calculate the relevance between sentences; it will give more details in next section.

3.2. FrameNet. Traditional semantic sentence representations employ WordNet [17] or a corpus-based measure [9] in a classic bag of words, without considering sentence-level semantic information and word order. The proposed semantic sentence representation uses semantic frames instead of words to take both sentence-level semantic information and word order into account.

FrameNet, a lexical database based on a theory called frame semantics, contains a wealth of semantic lexical resources. Frame semantics studying the meaning of words and syntactic structure based on real corpus is proposed by Charles J. Fillmore. This theory uses empirical methods to find the close relationship between language and the human experience and tries to describe this relationship with a possible way. The basic idea is straightforward: the meanings of most words can best be understood on the basis of a semantic frame, a description of a type of event, relation, or entity, and the participants in it.

In the context of FrameNet, "frame" is a linguistic term that refers to a set of concepts to understand when people use natural language to describe an event or a semantic scene. Words that evoke a frame are called "lexical units" (LUs); for example, bike, carry, and drive are the LUs of the BRINGING frame. Each frame contains a series of semantic roles containers which called Frame Elements (FEs). FEs are related to words described in the context of events or objects in real corpus. There are two classes of FEs, core and noncore elements, which are determined by their importance to the corresponding frame. Different frames have different types and numbers of Frame Elements, and these differences can reflect semantic information in natural language. Figure 2 is an illustration of the BRINGING frame.

Frame-to-frame relation describes an overview semantic relationship, which is an asymmetric directional relation. Eight frame relations are defined, which are inheritance, perspective_on, subframe, precedes, causative_of, inchoative_of, using, and see_also.

Each frame is directly related to two frames, based on the orientation relationship; one is called super_frame and the other is called sub_frame. The BRINGING frame-to-frame relation is as follows:

Inherits from:

Is Inherited by: *Smuggling*

Perspective on:

Is Perspectivized in:

Uses: *Cause_motion*, *Motion*

BRINGING

Definition:

This frame concerns the movement of a Theme and an Agent and/or Carrier. The Agent, a person or other sentient entity, controls the shared Path by moving the Theme during the motion. In other words, the Agent has overall motion in directing the motion of the Theme. The Carrier may be a separate entity, or it may be the Agent's body. The Constant_location may be a subregion of the Agent's body or (a subregion of) a vehicle that the Agent uses.

Karl CARRIED the books across campus to the library on his head.

The FEs include Path, Goal, and Source. Area is an area that contains the motion when the path is understood as irregular. This frame emphasizes the path of movement as opposed to the FEs Source or Goal as in Filling or Placing.

FE core set(s):

{Goal, Path, Source}, {Agent, Carrier}

Lexical units:

airlift.v, bear.v, bike.v, bring.v, bus.v, carry.v, cart.v, convey.v, drive.v, ferry.v, fetch.v, fly.v, get.v, haul.v, hump.v, jet.v, lug.v, mobile.a, motor.v, paddle.v, portable.a, row.v, schlep.v, shuttle.v, take.v, tote.v, transport.n, transport.v, truck.v, trundle.v, wheel.v

Figure 2: Definition of the frame.

Is Used by: *Convoy*, *Sending*

Subframe of:

Has Subframe(s):

Precedes:

Is Preceded by:

Is Inchoative of:

Is Causative of:

See also: *Cause_motion*, *Motion*, *Passing*, *Sending*

To sum up, the strengths of FrameNet are as follows.

(a) It contains a wealth of semantic information (semantic roles) and specifies the scene of semantic roles appeared under predicates.

(b) The definitions of semantic roles are intuitive and straightforward.

(c) It defines relationships between the predicates.

The weaknesses are listed below.

(a) Frames are determined by predicates, so new predicates need to define new frame.

(b) The expression of the semantic predicates is determined by other related words; the selection of frames is a thorny problem, especially in the case of polysemy.

(c) FrameNet may parse some phrases directly into semantic roles, but these phrases may be separable.

3.3. Word Embedding. Word embedding is a core technique, which brings deep learning into NLP research. In the past, it often uses a simple method, called one-hot representation, to transfer words into vectors. Every word is shown as a long vector; the dimension of the vector is the length of the thesaurus. It is represented by sparse matrix. For example, word "mic" is saved as $[0, 0, 0, 0, 1, \ldots, 0, 0]$ and word "microphone" is saved as $[0, 0, 0, 1, 0, \ldots, 0, 0]$. It is simple but has a fatal weakness; it cannot distinguish synonym, such as "microphone" and "mic." This phenomenon is called semantic gap. Meanwhile, the curse of dimensionality will appear when the dimension is large enough [32].

Hinton [33] proposed "distributed representation" in 1986. It uses low dimension real numbers to represent word vectors; the length of the dimension used to be 50 or 100. For example, word "mic" can be saved as $[0.58, 0.45, -0.15, 0.18, \ldots, 0.52]$. This method avoids the semantic gap by calculating the distances between words. Word embedding can give words similarity at the semantic level, which is an effective addition for FrameNet. At first, it used FrameNet to compute sentence similarity at a sentence-level. Although it achieved good results, the similarity between sentences calculation relies only on simple comparison within the framework of words. Word embedding transfers all the words in the sentence. In the calculation of sentence similarity, it can effectively reduce the redundancy of sentences by combining the subject or object [34]. In this paper, it uses the lexicon generated by Wikipedia words.

4. FrameNet-Based Semantic Graph Model

Sarkar [35] uses selected key phrases as key concepts identified from a document to create summary of document, while Sankarasubramaniam et al. [36] leverage Wikipedia as the knowledge base to understand the salient concepts of documents. However, the above techniques may fail to semantically analyze newly coined words; Heu et al. [37] employ the tag clusters used by Flickr, a Folksonomy system. Inspired by above researchers, FSGM adapts FrameNet to detect key concepts from documents.

In FSGM, it utilizes the semantic information while considering automatic text summarization, to avoid semantic gap as well as ignorance of grammatical structure by other methods. FrameNet is the basis of FSGM to represent semantic information. The Frame Elements in FrameNet can express rich semantic information of the framework. And FSGM requires the relationship between frames to calculate similarities between sentences. In order to make full use of semantic information, word embedding is applied in computation of sentence similarity. In construction of the semantic graph, sentences should be aggregated when sentence similarity value is greater than the threshold α, and the aggregated sentence should assign new aggregated weights. Thus, the number of nodes in the semantic graph n is not greater than the number of sentences in the text. Graph

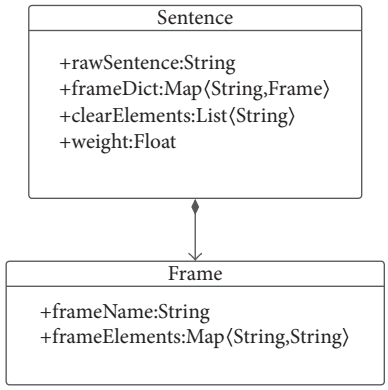


Figure 3: UML form of the structure.

sorting algorithm is then applied to the semantic graph. After the calculation, the aggregation and nodes weights are combined to select the most representative sentence. Detailed steps are as follows.

(a) Use FrameNet to parse the text, to identify the frame, and annotate the sentences with LUs and FEs, and generate the sentence node.

(b) Calculate the similarity between sentence nodes by combining the FrameNet with the word embedding.

(c) Aggregate sentences when the similarity value is greater than threshold α.

(d) Calculate the weight of sentence nodes by graph ranking algorithm.

(e) The final weights of the sentence nodes are calculated by using the method of combination of aggregate weights and relationship weights.

(f) Select the most representative sentence to generate summary.

4.1. Structure of Sentence Node. In order to facilitate further steps after preprocessing, sentences should be stored in specific structure. The structure of sentence node in FSGM must include original text, recognized frame, Frame Elements, lexical elements after preprocessing, and sentence weight. It should be noted that a sentence may contain multiple frames, and it designs the most basic data structure for storing the semantic representation of the sentence. Figure 3 shows the UML form of the structure.

In Figure 3, frame structure includes frame name and collection of Frame Elements. Sentence structure includes original sentence, collection of frames, collection of lexical elements after preprocessing, and sentence weight.

After computing sentence similarity, FSGM will aggregate appropriate sentences, for the purpose of reducing redundancy and promoting diversity while generating summary.

4.2. Sentence Semantic Parsing. When FSGM parses sentences, it first annotates LUs in sentences to identify frames inside, then locates other FEs in sentences according to the FEs defined in the frames, and stores parsed semantic

Figure 4: Semantic parsing.

information in predefined sentence structure. FSGM uses an open-source semantic parser SEMAFOR. It first analyzes sentence with a rule-based system to recognize target LUs and then identifies frames by using a statistical model. FEs are captured by another statistical model at the last step. Figure 4 is an example of semantic parsing.

For the sentence "Kate drove home in a stupor." two frames are identified from the sentence: BRINGING and FOREIGN_OR_DOMESTRIC_COUNTRY. Moreover, Frame Elements Agent and Theme are annotated in the frame BRINGING.

4.3. Sentence Semantic Similarity. Sentence semantic similarity is the kernel of the FSGM, which describes differences between sentences in semantic level by combining the FrameNet and word embedding. Lin et al. [38] propose a similarity measure for text by computing the similarity between two documents with respect to a feature. FSGM regards frame as the feature between sentences and takes following three cases into account.

If frames identified in both sentences are the same and the Frame Elements defined in the frames are also the same, the similarity between two sentences is 1, indicating in the sight of FrameNet that the semantics is the same. If the frame is different, it defined similarity as

$$\mathrm{Sim}\left(S_i, S_j\right) = \frac{\sum_{k=0}^{n} \theta_k \times \mathrm{Distance}\left(\mathrm{SF}_{ik}, \mathrm{SF}_{jk}\right)}{\sum_{k=0}^{n} \theta_k}, \quad (3)$$

where SF is word vector of the lexicon elements and θ is the coefficient of lexicon elements, which has 3 types. The first is that elements are under the same frame, the second is that elements are under the 8 frames mentioned above, and the third is all the other conditions. In this paper, θ is set by 1.2, 1.1, and 1.0. n is the smaller number in lexicon elements, set by $\min(\mathrm{num}(\mathrm{SF}_i), \mathrm{num}(\mathrm{SF}_j))$. Lexicon elements either are under frame or are not.

Word vectors are as follows:

$$\mathrm{SF} = \frac{\sum_{i=0}^{n} E_i}{n}, \quad (4)$$

where n is the number of elements in lexicon setting and E_i is the number of i's word vector.

$\mathrm{Distance}(\mathrm{SF}_i, \mathrm{SF}_j)$ is the Cosine distance:

$$\mathrm{Distance}\left(\mathrm{SF}_i, \mathrm{SF}_j\right) = \frac{\sum_{k=0}^{n} v_{ik} \times v_{jk}}{\sqrt{\sum_{k=0}^{n} v_{ik}^2} \sqrt{\sum_{k=0}^{n} v_{jk}^2}}, \quad (5)$$

where v_{ik} is SF_i's value in k dimension.

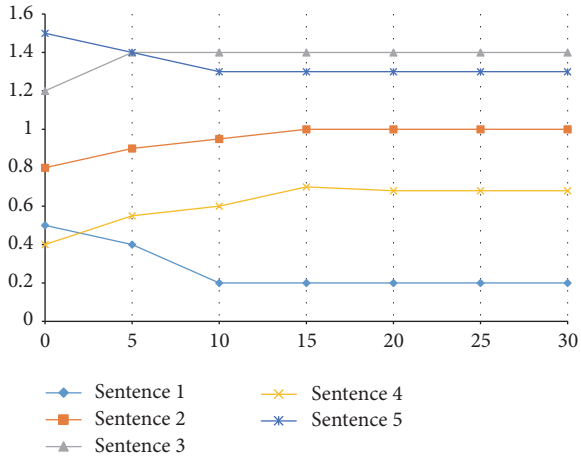


Figure 5: Example of convergence curves.

Word embedding generates a multidimension word vector for each word; [34] proved that adding up all the word vectors in one sentence and calculating the Cosine distance is an effective method to get the sentence similarity.

4.4. Construct Semantic Graph. After transforming sentences into semantic sentence representation and calculating of the semantic similarity between pairwise sentences, the sentence nodes need to be merged for variability and diversity. The threshold of the similarity is 1, which means FSGM only merges the same sentences.

FSGM builds a weighted undirected semantic graph $G = \{V, E\}$, where V is the vertex set and E is the edge set of the graph. Each sentence in documents is represented in frames and is built to be a graph vertex. If the semantic similarity between them is larger than some threshold, these representative vertexes in the semantic graph will be linked by edges and the edge weight is assigned by the value of the pairwise semantic similarity.

Based on the weighted undirected graph, every sentence is evaluated by applying the PageRank-like algorithm. Equation (6) shows how to calculate the weight of each vertex in graph,

$$\text{weight}\left(V_i\right) = (1-d) + d \times \sum_{V_j \in \ln(V_i)} \frac{\text{Sim}\left(S_i, S_j\right) \times \text{weight}\left(V_j\right)}{\sum_{V_k \in \text{Out}(V_j)} \text{Sim}\left(S_j, S_k\right)}, \quad (6)$$

where weight(V_i) represents the weight of V_i which also represents the sentence S_i, d represents the damping factor, usually set to 0.85, $\ln(V_i)$ is the set of vertexes which is connected to V_i, since the graph is undirected, $\text{Out}(V_i) = \ln(V_i)$, and $\text{Sim}(S_i, S_j)$ is the semantic similarity between S_i and S_j. The convergence threshold η is set to 10^{-5}. In actual calculation, an initial value is given for V_i and then updated by (6). Experiments show that (6) usually converges in 20–30 iterations in a sentence semantic graph.

Figure 5 shows the weight change of five sentences in a document with a given initial value C. The abscissa represents the number of iterations; the ordinate represents the sentence weight of current state. In the undirected weighted graph, vertex weight converges very quickly. After a few iterations, the difference between the vertex weight values is far less than η.

With sentence-level semantic graph, it can obtain the weight of each sentence in the document. After calculating the similarity, it gathers the sentences if the values are larger than the threshold α and generate a new node. Given the new node its weight is based on the number of combining sentences. Then, if the similarity values are larger than the threshold β, we link these two sentences. After that, it generates a weighted graph. The convergence results in graph ranking model are irrelevant with the initial weight vector. After iteration, it finally gets a FSGM.

4.5. Sentence Semantic Weight. Word embedding can represent word senses in word embedding vectors. As the dimension of word embedding vectors is consistent between words, words in a sentences can sum their word embedding vectors up to represent the sentence. Also sentences in a document can sum their vectors to represent the document. So the semantic representation of document can be word embedding vectors.

The vectors of the document contain all semantic information. The sentence semantic weights should be the amount of information that the vectors of sentence contained. So the weight is calculated by the Cosine distance between the document vectors and the sentence vectors.

4.6. Sentence Selection. It should consider both the semantic weight and the weight in the text structure when generating summary. FSGM combined the semantic weight and the relation weight; the formula is as follows:

$$W = \mu W_s + (1-\mu) W_g, \quad (7)$$

where μ is the coefficient of semantic weight, $(1-\mu)$ is coefficient of relation weight, W is the final weight, W_s is the semantic weight for sentence node, and W_g is the relation weight for sentence node.

5. Experiments

In this section, it introduces the data set, the evaluation metric, and the result of text summarization using proposed FSGM.

5.1. Setup. Firstly, to compare our methods, it uses several document summarization baseline systems. The definitions of these systems are as follows.

Lead. For each topic, it would return the leading sentences of every document.

Random. For each topic, it would select sentences randomly.

GraphRank. Rank sentences by graph-based algorithms using traditional bag-of-word.

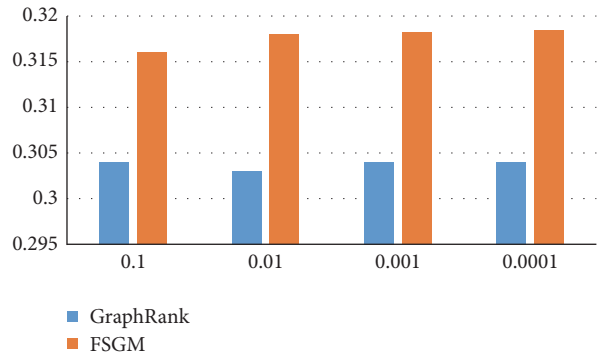


Figure 6: Sensitivity of ROUGE-1 with respect to threshold η.

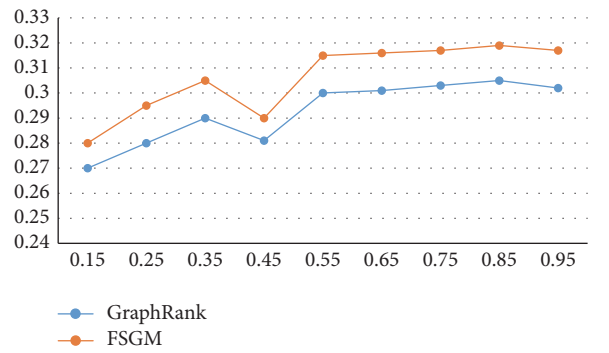


Figure 7: Sensitivity of ROUGE-1 with respect to damping factor d.

5.2. Data Set. In the experiments, it chooses DUC 2004 dataset for text summarization. It is the collections of newspaper documents from TREC. It has the original summaries. Besides, the datasets are public. Therefore, lots of researchers select DUC 2004 to study text summarization. For each document, NIST human assessors wrote two summaries: one is 200 words and the other is 400 words. And for each task, there are two summarizer models. There are 5 tasks in DUC 2004; each task has 15 clusters. It has chosen 15 clusters randomly from the 75 clusters.

5.3. Evaluation Metrics. ROUGE is a performance evaluation method widely applied by DUC. It chooses this method to measure our FSGM. It measures the summary quality by counting the overlaps between a set of reference summaries and candidate summary. After that, there are several kinds of automatic evaluation methods, such as ROUGE-N, ROUGE-W, ROUGE-L, and ROUGE-SU. Equation (8) shows the compute step for ROUGE-N which is a n-gram recall:

$$\text{ROUGE-}N = \frac{\sum_{S\in\{\text{RefSum}\}} \sum_{n\text{-gram}\in S} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{S\in\{\text{RefSum}\}} \sum_{n\text{-gram}\in S} \text{Count}(n\text{-gram})}, \quad (8)$$

where n is the length of the n-gram and Ref is the reference summaries. In a candidate summary and the reference summary, $\text{Count}_{\text{match}}(n\text{-gram})$ is the max of n-gram, and in the reference summaries, the number of n-gram is $\text{Count}(n\text{-gram})$. The longest common subsequence (LCS) has been used in ROUGE-L statistics; while ROUGE-W is based on weighted LCS, ROUGE-SU is on the basis of skip-bigram plus unigram. All these four evaluation methods in ROUGE can form three scores, which are recall, precision, and F-measure.

It uses the ROUGE toolkit 1.5.5 for evaluation. It calculates these scores by our FSGM method and compared with three other systems, (Lead, Random, and GraphRank). It uses the scores of ROUGE-1, ROUGE-2, and ROUGE-L.

5.4. Result and Discussion. For graph-based text summarization method, the value of threshold η and damping factor d will affect the performance of the graph-based algorithm. It analyzes the sensitivity of both parameters for two graph-based systems involved in our experiment including GraphRank and FSGM. The threshold η is varied in the set $\{0.1, 0.01, 0.001, 0.0001\}$ and damping factor d is varied in the range between 0.05 and 0.95 with an interval of 0.1. From Figure 6, it can be inferred that the choice of threshold η in the set $\{0.1, 0.01, 0.001, 0.0001\}$ is quite insensitive to both systems. And the experiments show that when η is set to 0.0001, both systems gave the best results. Figure 7 shows that the results are not sensitive to the damping factor d in the range between 0.65 and 0.95 of both systems.

$d = 0.85$ is chosen for further experiments as it gave the best results.

In order to investigate the effects of performance by different parameters, it compares the default setting of several parameters with designed parameters.

Table 1 illustrates the performance comparison of FSGM with and without word embedding. The table clearly shows that word embedding can better utilize the semantic information to calculate similarity between sentences. It should mention that although word embedding can finitely improve the performance of generating summary, the option of adapting word embedding or not should be considered thoroughly.

Different settings of coefficients of lexical elements in the semantic similarity are compared in Table 2. The coefficient is a vector of the three elements. Each element is the weight of the related collections of lexical elements. It can be inferred from the table that the default setting achieved the best performance. The default setting assigns reasonable weights to three types of lexical elements collections. The θ_3 assigns all types of lexical elements with the weight of one, but it mixes different types of collections without consideration.

Table 1: Performance comparison of FSGM with word embedding and without word embedding.

Method	FSGM (with word embedding, default settings)	FSGM (without word embedding, default setting)
ROUGE-1	0.325	0.315
ROUGE-2	0.052	0.049
ROUGE-L	0.298	0.293

Table 2: Performance comparison of FSGM with different coefficients of lexical elements.

Parameter	θ_{default} (1.2, 1.1, 1.0)	θ_1 (2, 1.5, 1.0)	θ_2 (1.5, 1.3, 1.0)	θ_3 (1.0, 1.0, 1.0)	θ_4 (0.5, 0.8, 1.0)
ROUGE-1	0.325	0.315	0.322	0.310	0.298
ROUGE-2	0.052	0.049	0.049	0.047	0.042
ROUGE-L	0.298	0.305	0.303	0.293	0.289

Table 3: Performance comparison of FSGM with different thresholds of combining sentences.

Parameter	α_{default} (1)	α_1 (0.95)	α_2 (0.90)	α_3 (0.85)	α_4 (0.5)
ROUGE-1	0.325	0.320	0.320	0.317	0.198
ROUGE-2	0.052	0.049	0.052	0.046	0.019
ROUGE-L	0.298	0.300	0.295	0.293	0.135

Table 4: Performance comparison of FSGM with different thresholds of estimating sentence relationship.

Parameter	β_{default} (0.5)	β_1 (0.9)	β_2 (0.7)	β_3 (0.3)	β_4 (0.1)
ROUGE-1	0.325	0.167	0.240	0.214	0.156
ROUGE-2	0.052	0.019	0.032	0.024	0.013
ROUGE-L	0.298	0.126	0.145	0.156	0.113

Table 5: Performance comparison of FSGM with different coefficients of estimating semantic weight.

Parameter	μ_{default} (0.3)	μ_1 (0.9)	μ_2 (0.7)	μ_3 (0.5)	μ_4 (0.1)
ROUGE-1	0.325	0.298	0.320	0.307	0.306
ROUGE-2	0.052	0.043	0.051	0.049	0.045
ROUGE-L	0.298	0.270	0.295	0.293	0.273

Table 3 is the performance comparison of FSGM with different thresholds of combining sentences. When the threshold is set by 1, the performance is the best. When it comes lower, all parameters are worse. And when it comes to 0.5, which means that it combines two sentences when they are only half same, ROUGE parameters make no sense. The more similar the two sentences are when combined, the better their performance is.

Table 4 is the performance comparison of FSGM with different thresholds of estimating sentence relationship. The best performance appears on 0.5. This line almost obeys the normal distribution, while the sentences should not be too similar nor too dissimilar. When β is small, there are little edges; when β is too big, nearly all lines link between nodes.

Table 5 is the performance comparison of FSGM with different coefficients of estimating semantic weight. This parameter has less influence than others. The best setting is $\mu = 0.3$.

Moreover, this paper analyzes and compares the FSGM performance on the DUC2004 with several baseline systems. Table 6 presents the results that are achieved by Lead, Random, GraphRank, and FSGM. The table clearly shows the proposed FSGM outperformed other systems. It can be inferred from the results that text summarization gets better performance using sentence-level semantic information.

It should notice that Lead's performance outperformed Random in every ROUGE measure, and its results of measure ROUGE-2 and ROUGE-L even approach graph-based system GraphRank. Lead method only chooses the first sentence of the document as the abstract, while the result is similar to GraphRank. It means that the location of sentences must be considered when generating abstract. Although Lead method is simple, its performance is instable, because it greatly depends on the context.

Unlike Lead, GraphRank analyzes the context based on relations of words. It builds a graph model based on sentence similarity and analyzes their occurrence relations. So the performance is much better than baseline. But it is not enough to consider the sentence-level semantic similarity. It only focuses on the relationship between sentences, ignoring sentence itself as basic morpheme. On such basis, FSGM considers the semantic similarity between sentences, so it gets the best performance. Meanwhile, as it considers the relationship between sentences, the performance is very stable.

While there are multiple choices of graph ranking algorithm, it choses PageRank-like algorithm as the default ranking algorithm of FSGM. It choses another famous graph ranking algorithm HITS for performance comparison, which is also widely applied in measuring the importance of graph vertexes. Both HITS and PageRank-like algorithm applied in the FSGM achieve the best results. It can be inferred from Table 6 that taking sentence-level semantic information into consideration can improve the performance of general graph ranking algorithm. The reason why PageRank-like algorithm does better than HITS in the experiment may be because that the former is topic independent while the latter is topic related. The FSGM based on HITS may be more suitable in query-based tasks.

6. Conclusion

In this paper, it reviews the common methods of text summarization and proposes a Semantic Graph Model using FrameNet called FSGM. Besides the basic functions, it particularly takes sentence meaning and words order into consideration, and therefore it can discover the semantic relations between sentences. This method mainly optimizes the sentences nodes by combining similar sentences using word embedding. Also, giving the sentences its weight and optimizing the PageRank can make the model more rigorous. The results show that FSGM is more effective from the understanding of sentence-level semantic.

Above all, if it can take more semantic information into account, it may probably get a better result. In the future work, it prepares to build a multiple-layer model to further show the

Table 6: Performance comparison on DUC2004 using ROUGE evaluation methods.

Parameter	Lead	Random	GraphRank	FSGM (PageRank-like)	FSGM (HITS-like)
ROUGE-1	0.292	0.290	0.304	0.325	0.310
ROUGE-2	0.043	0.041	0.041	0.052	0.045
ROUGE-*L*	0.271	0.264	0.265	0.298	0.280

accuracy rate of application in text summarization. And in this paper, it only applies FSGM to a test corpus. Nowadays, text from social media is the main resource. And there will be more serious problems about the credibility. It will research on the social media content in the future.

Disclosure

This paper is based on the authors' paper "Text Summarization Using Sentence-Level Semantic Graph Model" from 2016 4th IEEE International Conference on Cloud Computing and Intelligence Systems.

Competing Interests

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

Acknowledgments

This work is supported by Basic Research of the Ministry of Science and Technology, China (2013FY114000).

References

[1] K. S. Jones, "Automatic summarising: factors and directions," in *Advances in Automatic Text Summarization*, I. Mani and M. Maybury, Eds., pp. 1–12, MIT Press, Cambridge, Mass, USA, 1999.

[2] K. M. Svore, L. Vanderwende, and C. J. C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pp. 448–457, June 2007.

[3] T. Hirao, Y. Sasaki, and H. Isozaki, "An extrinsic evaluation for question-biased text summarization on QA tasks," in *Proceedings of the NAACL Workshop on Automatic Summarization*, 2001.

[4] X. Wan and J. Zhang, "CTSUM: extracting more certain summaries for news articles," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, pp. 787–796, Queensland, Australia, July 2014.

[5] B. Chen, S.-H. Lin, Y.-M. Chang, and J.-W. Liu, "Extractive speech summarization using evaluation metric-related training criteria," *Information Processing & Management*, vol. 49, no. 1, pp. 1–12, 2013.

[6] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Computational Linguistics*, vol. 28, no. 4, pp. 399–408, 2002.

[7] E. Baralis, L. Cagliero, N. Mahoto, and A. Fiori, "Graph Sum: discovering correlations among multiple terms for graph-based summarization," *Information Sciences*, vol. 249, no. 16, pp. 96–109, 2013.

[8] X. Li, S. Zhu, H. Xie et al., "Document summarization via self-present sentence relevance model," in *Database Systems for Advanced Applications*, pp. 309–323, Springer, Berlin, Germany, 2013.

[9] P. Goyal, L. Behera, and T. M. McGinnity, "A context-based word indexing model for document summarization," *IEEE Transactions on Knowledge & Data Engineering*, vol. 25, no. 8, pp. 1693–1705, 2013.

[10] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.

[11] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 307–314, ACM, July 2008.

[12] J.-P. Mei and L. Chen, "SumCR: a new subtopic-based extractive approach for text summarization," *Knowledge & Information Systems*, vol. 31, no. 3, pp. 527–545, 2012.

[13] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th ACM Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, pp. 19–25, New Orleans, La, USA, September 2001.

[14] D. Wang, S. Zhu, T. Li et al., "Comparative document summarization via discriminative sentence selection," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 3, pp. 1963–1966, 2009.

[15] R. He, J. Tang, P. Gong, Q. Hu, and B. Wang, "Multi-document summarization via group sparse learning," *Information Sciences*, vol. 349-350, pp. 12–24, 2016.

[16] T. Li, "A general model for clustering binary data," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 188–197, August 2005.

[17] S. Park, J.-H. Lee, D.-H. Kim, and C.-M. Ahn, "Multi-document summarization based on cluster using non-negative matrix factorization," in *SOFSEM 2007: Theory and Practice of Computer Science: 33rd Conference on Current Trends in Theory and Practice of Computer Science, Harrachov, Czech Republic, January 20–26, 2007. Proceedings*, vol. 4362 of *Lecture Notes in Computer Science*, pp. 761–770, Springer, Berlin, Germany, 2007.

[18] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, pp. 269–274, San Francisco, Calif, USA, August 2001.

[19] R. Katragadda, P. Pingali, and V. Varma, "Sentence position revisited: a robust light-weight update summarization 'Baseline' algorithm," in *Proceedings of the 3rd International Workshop Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3 '09)*, pp. 46–52, Boulder, Colo, USA, June 2009.

[20] C.-Y. Lin and E. Hovy, "Identifying topics by position," in *Proceedings of the 5th conference on Applied Natural Language Processing*, pp. 283–290, Washington, DC, USA, April 1997.

[21] C.-Y. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proceedings of the 18th Conference on Computational Linguistics (COLING '00)*, pp. 495–501, Saarbrücken, Germany, August 2000.

[22] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *Proceedings of the ACL Workshop Intelligent Scalable Text Summarization*, pp. 10–17, Madrid, Spain, July 1997.

[23] M. H. Haggag, "Semantic text summarization based on syntactic patterns," *International Journal of Information Retrieval Research*, vol. 3, no. 4, pp. 18–34, 2013.

[24] Y. Ko and J. Seo, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1366–1371, 2008.

[25] E. Canhasi and I. Kononenko, "Multi-document summarization via Archetypal Analysis of the content-graph joint model," *Knowledge & Information Systems*, vol. 41, no. 3, pp. 821–842, 2014.

[26] F. Wei, W. Li, Q. Lu, and Y. He, "A document-sensitive graph model for multi-document summarization," *Knowledge and Information Systems*, vol. 22, no. 2, pp. 245–259, 2010.

[27] P. Goyal, L. Behera, and T. M. McGinnity, "A context-based word indexing model for document summarization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1693–1705, 2013.

[28] S. Harabagiu and F. Lacatusu, "Topic themes for multi-document summarization," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, pp. 202–209, August 2005.

[29] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in *Proceedings of the ACL on Interactive Poster and Demonstration Sessions*, p. 20, Association for Computational Linguistics, 2004.

[30] R. Ferreira, F. Freitas, L. De Souza Cabral et al., "A four dimension graph model for automatic text summarization," in *Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies (IAT '13)*, vol. 1, pp. 389–396, November 2013.

[31] R. Ferreira, R. D. Lins, F. Freitas, S. J. Simske, and M. Riss, "A new sentence similarity assessment measure based on a three-layer sentence representation," in *Proceedings of the ACM Symposium on Document Engineering (DocEng '14)*, pp. 25–34, Fort Collins, Colo, USA, September 2014.

[32] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1137–1155, 2003.

[33] G. E. Hinton, "Learning distributed representations of concepts," in *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, vol. 1, pp. 1–12, Amherst, Mass, USA, 1986.

[34] M. J. Kusner, Y. Sun, N. I. Kolkin et al., "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on Machine Learning (ICML '15)*, pp. 957–966, Lille, France, 2015.

[35] K. Sarkar, "Automatic single document text summarization using key concepts in documents," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 602–620, 2013.

[36] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh, "Text summarization using Wikipedia," *Information Processing & Management*, vol. 50, no. 3, pp. 443–461, 2014.

[37] J.-U. Heu, I. Qasim, and D.-H. Lee, "FoDoSu: multi-document summarization exploiting semantic analysis based on social Folksonomy," *Information Processing & Management*, vol. 51, no. 1, pp. 212–225, 2015.

[38] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 7, pp. 1575–1590, 2014.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 6842891, 12 pages
http://dx.doi.org/10.1155/2016/6842891

Research Article

Cloud Model Approach for Lateral Control of Intelligent Vehicle Systems

Hongbo Gao,[1,2,3] Xinyu Zhang,[3] Yuchao Liu,[4] and Deyi Li[1,2,4]

[1]*State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China*
[2]*State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100083, China*
[3]*Information Technology Center, Tsinghua University, Beijing 100083, China*
[4]*The Institute of Electronic System Engineering, Beijing 100039, China*

Correspondence should be addressed to Xinyu Zhang; xyzhang@tsinghua.edu.cn

Received 6 June 2016; Revised 9 September 2016; Accepted 28 September 2016

Academic Editor: Xiong Luo

Studies on intelligent vehicles, among which the controlling method of intelligent vehicles is a key technique, have drawn the attention of industry and the academe. This study focuses on designing an intelligent lateral control algorithm for vehicles at various speeds, formulating a strategy, introducing the Gauss cloud model and the cloud reasoning algorithm, and proposing a cloud control algorithm for calculating intelligent vehicle lateral offsets. A real vehicle test is applied to explain the implementation of the algorithm. Empirical results show that if the Gauss cloud model and the cloud reasoning algorithm are applied to calculate the lateral control offset and the vehicles drive at different speeds within a direction control area of ±7°, a stable control effect is achieved.

1. Introduction

In academic and industrial circles, studies on intelligent vehicles have drawn considerable attention. Such studies play an important role in the research on vehicles and intelligent transportation. Control methods are the key in the study of intelligent vehicles. Vehicle model parameters are extremely complex. The system model equation is nonlinear, and its system parameters constantly change over time. Research on vehicle control theory includes lateral tracking control and longitudinal tracking control. Lateral tracking control includes the support vector machine (SVM) method, the sublevel control method [1], the traditional PID (Proportional-Integral-Derivative) method [2], and the intelligent method. The latter includes the fuzzy control [3, 4] and neural network control methods [5, 6]. Longitudinal tracking control includes the coordination of the brake and the accelerator as well as the antijamming capability of the control accuracy. One important study on intelligent vehicle control is the Urban Challenge, which was organized by DARPA (Defense Advanced Research Projects Agency) in 2007. The champion, "BOSS," used a road navigation and regional navigation control strategy [7]. The third placer, the "Odin" team, used a control model [8] based on driver behavior. The "Talos" team applied the output path of the navigator and the speed commands of the low-level controller output, using the method based on a RRT (Rapidly-exploring-Random-Tree) [9], and generated a dynamic trajectory feasible tree through countless random samples, thereby expanding the typical RRT [10]. Current self-adaptive control methods of intelligent vehicles modify parameters of PID on the basis of changes in intelligent vehicle states and object properties, thereby improving control. They mainly include adaptive control reference models [11], adaptive control fuzzy models [12, 13], adaptive control neural networks models [14, 15], and adaptive control evolutionary models [16].

The current study aims to improve the accuracy, robustness, and adaptability to various road conditions of the vehicle control algorithm. First, the convergence of vehicles toward trajectory tracking errors is investigated from the perspective of nonlinear system stability, which is the premise of vehicle tracking trajectory. Subsequently, the robustness and control algorithm that can adapt to the environment is also considered, thereby ensuring control performance when

the running conditions of a vehicle are drastically changed. Finally, the function of vehicle motion control is expanded, which enables vehicles to complete automatic overtaking task, adaptive cruise task, automatic parking task, flowing into traffic task, and so on.

In most of the studies cited above, some researches only focused on lateral tracking control and some researches only focused on longitudinal tracking control, without considering driving speed and driving direction as input values. When intelligent driving tasks increase in complexity, the control systems cited earlier are unable to adapt to complex tasks. In addition, the control system should be able to guarantee stability. The main contributions of our study are as follows. (1) A new uncertainty control system according to the Gauss cloud model (GCM) and cloud reasoning is illustrated. (2) The new model considers both speed and direction, whereas velocity and direction are mutually constrained. (3) The speed control rules for intelligent driving vehicles are constructed, with reference to human driving experience.

This paper is organized as follows. Section 1 presents the lateral control of intelligent vehicle. Section 2 presents the GCM, the GCM algorithm, and cloud reasoning, including a preconditioned Gauss cloud generator (PGCG), a postconditioned Gauss cloud generator (PCGCG), and a rule generator. Section 3 describes the lateral control algorithm for intelligent vehicle systems and cloud controller rules. Section 4 provides the results of the experiment and analysis performed using the cloud control algorithm. Finally, the results of experiment are illustrated in Section 5.

2. Model and Problem Formulation

2.1. Gauss Cloud Model. The Gauss distribution (GD) is one of the most important distributions in probability theory, in which the general characteristics of random variables are represented as means of the mean and variance of two numbers. As a fuzzy membership function, the bell-shaped membership function is mostly used in sets, which is typically expressed through the analytical expressions of $m(x) = \exp\{-(x-a)^2/2b^2\}$. This study presents a cloud model based on the GD, called the Gauss cloud model (GCM), which is defined as follows [17, 18].

Definition 1. U is expressed in a precise numerical quantitative domain. $C(Ex, En, He)$ is a qualitative concept on U. If the value of x ($x \in U$) is a random realization of the qualitative concepts of C, then the "expectation" of the GD $x \sim N(Ex, En'^2)$ is denoted as Ex, and its "variance" is denoted as En'^2. Meanwhile, the "expectation" of GD $En' \sim N(En, He^2)$ is denoted as En, and its "variance" is denoted as He^2. En' is the full form of GD $En' \sim N(En, He^2)$ and is a random realization [19]. The certainty degree of x in C is satisfied via $m(x) = \exp\{-(x-Ex)^2/2(En')^2\}$. The distribution of x in the domain of U is called a Gauss cloud (GC) [20]. The GC algorithm is presented as follows [17, 20].

The GC Algorithm

Input. Three figures (Ex, En, He) and the number of cloud drops n.

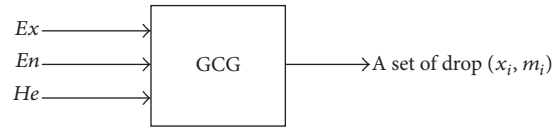


Figure 1: The GCG.

Output. A sample set that represents concept extension and its certainty (x_i, m_i), $i = 1, \ldots, n$.

(1) Generate Gauss random $En' \sim N(En, He^2)$.

(2) Generate Gauss random $x \sim N(Ex, En'^2)$.

(3) Calculate the certainty: $m(x) = \exp\{-(x-Ex)^2/2(En')^2\}$.

(4) Repeat (1)–(4) until the number of cloud drops is n.

The algorithm causes distribution drops, called cloud distribution (CD). The algorithm of GCM can be obtained through a cloud generator (CG), which forms a forward Gauss cloud generator (GCG), as shown in Figure 1. The Gauss random number generation method is the foundation of the whole algorithm. It generates uniform random numbers in $[0, 1]$ and uses them to calculate the Gauss random number. Random number sequences are determined through the uniform random function of a seed. The method of using uniform random numbers to generate a Gauss random number is described in detail in [21]. GC distribution (GCD) is different from the GD because the GCD algorithm uses the Gauss random number twice, in which one random number is the basis of another random number. Among these,

(1) when $He = 0$, the algorithm generates a precise value of En and the value of x is transformed into a GD,

(2) when $He = 0$ and $En = 0$, the value of x of the algorithm generation is an exact value of Ex, and $m \equiv 1$.

From (1) and (2), certainty can be concluded as a special case of uncertainty, and the GD is a special case of the GCD.

For a qualitative concept of a steering angle of positive and negative 40°, given that $Ex = 80°$, $En = 1$, and $He = 0.1$, 1000 cloud drops are generated. The distribution of drops and its certainty degree of $C(x, m)$ are shown in Figure 2.

2.2. Cloud Reasoning

2.2.1. Preconditioned Gauss Cloud Generators and Postconditioned Gauss Cloud Generators. Knowledge forms a concept and its relationship with communicating and abstracting. The relationship among concepts forms certain rules, from which rules library and rules generator can be established through knowledge reasoning based on GC. Rules include preconditioned and postconditioned rules. Preconditioned rules include one or several rules, whereas postconditioned rules express the results and specific control actions generated by the preconditioned rules. In the control field, "perception-action" can establish the rule library based on the relationship among concepts, thereby realizing control of uncertainty.

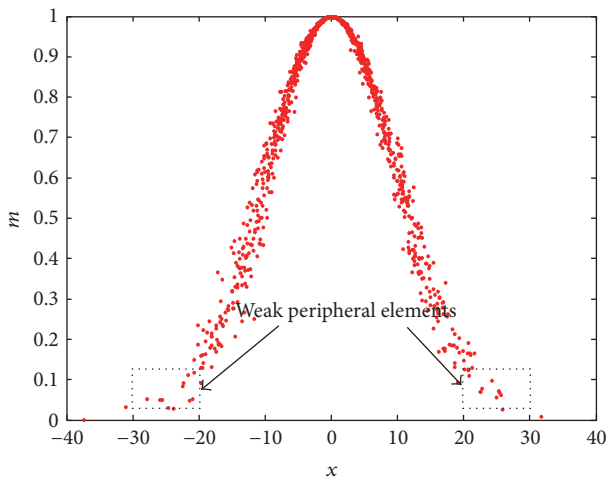


Figure 2: The distribution of 1000 drops.

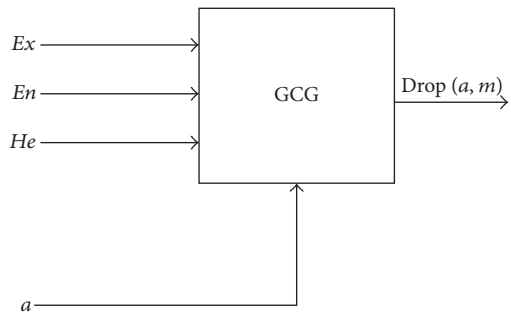


Figure 3: The PGCG.

A preconditioned Gauss cloud generator (PGCG) and a postconditioned Gauss cloud generator (PCGCG) are composed of the GCG, which is defined as follows.

Definition 2. Assume the following rule:

$$\text{If } A, \text{ then } B, \tag{1}$$

where A corresponds to concepts C_1 in universal sets U_1 and B corresponds to concepts C_2 in universal sets U_2. a is a specific value in universal sets U_1, where the GCG generates a specific value of a based on the concept C_1 of the certainty degree of m distribution, and $m \in [0, 1]$, which is called a PGCG [22], as shown in Figure 3.

The PGCG algorithm is presented as follows [23].

The PGCG Algorithm

Input. Three figures (Ex, En, He) and a specific value a.

Output. The distribution of drops (a, m).

(1) Generate Gauss random $En' \sim N(En, He^2)$.

(2) Calculate the certainty: $m(x) = \exp\{-(x - Ex)^2 / 2(En')^2\}$.

(3) Generate the distribution of drops (a, m).

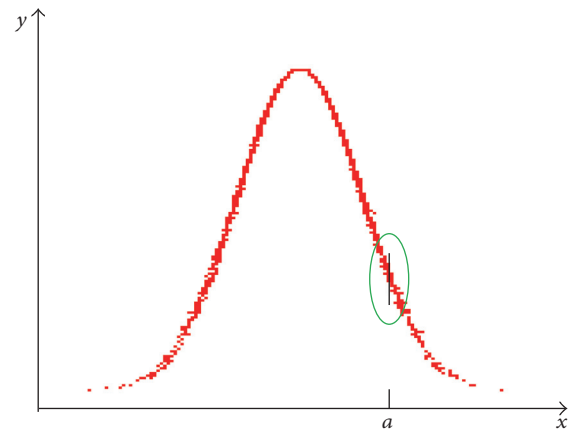


Figure 4: Cloud drop distribution of the PGCG.

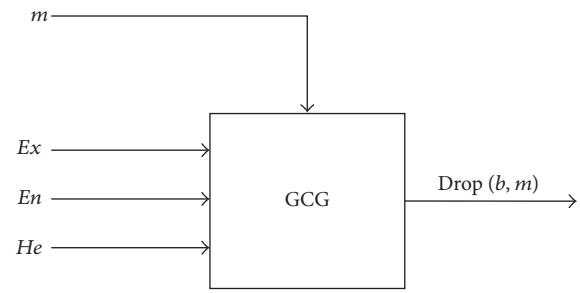


Figure 5: The PCGCG.

As shown in Figure 4, the distribution of drops (a, m) of the specific value of a and the certainty degree of m is on the line of $x = a$.

Definition 3. Assume the following rule:

$$\text{If } A, \text{ then } B, \tag{2}$$

where A corresponds to concepts C_1 in universal sets U_1 and B corresponds to concepts C_2 in universal sets U_2. The certainty degree of m belongs to $[0, 1]$. The GCG generates the certainty degree of m drop distribution, which is satisfied by applying concepts C_2 in universal sets U_2, called PCGCG [24], as shown in Figure 5.

The PCGCG algorithm is presented as follows [25, 26].

The PCGCG Algorithm

Input. Three figures (Ex, En, He) and certainty degree m.

Output. The drop distribution (b, m).

(1) Generate Gauss random $En' \sim N(En, He^2)$.

(2) Calculate the certainty: $b = Ex \pm En' \sqrt{-2 \ln m}$.

(3) Generate the distribution of drops (b, m).

As shown in Figure 6, the drop distribution (b, m) of the cloud drop specific value of b and the certainty degree of m is on the line of $y = m$.

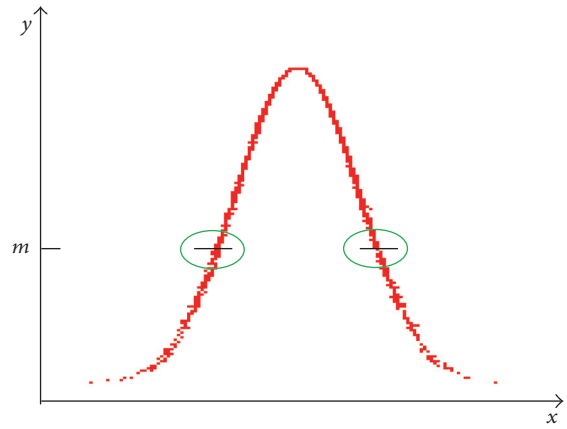


Figure 6: Cloud drop distribution of the PCGCG.

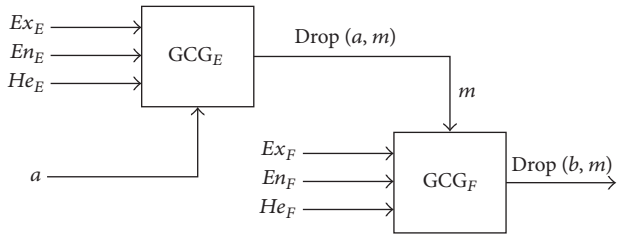


Figure 7: The SCSRGCG.

2.2.2. Rule Generator

Definition 4. Assume the following rule:

$$\text{If } E, \text{ then } F, \tag{3}$$

where E is the PGCG that generates the drop distribution (a, m) with a specific value of a and a certainty degree of m. F is the PCGCG that generates the drop distribution (b, m) of the cloud with a specific value of b and a certainty degree of m, which is called the single-condition single-rule GCG (SCSRGCG) [27, 28]. The composition diagrams of PGCG and PCGCG are shown in Figure 7.

The SCSRGCG algorithm is presented as follows.

The SCSRGCG Algorithm

Input. Three figures (Ex_E, En_E, He_E), three figures (Ex_F, En_F, He_F), and a specific value a.

Output. The drop distribution (b, m).

(1) Generate Gauss random $En'_E \sim N(En_E, He_E^2)$.

(2) Calculate the certainty: $m = \exp\{-(x - Ex_E)^2 / 2(En'_E)^2\}$.

(3) Generate Gauss random $En'_F \sim N(En_F, He_F^2)$.

(4) If $a < Ex$, then calculate the certainty: $b = Ex_F - En'_F\sqrt{-2\ln m}$.

(5) If $a > Ex$, then calculate the certainty: $b = Ex_F + En'_F\sqrt{-2\ln m}$.

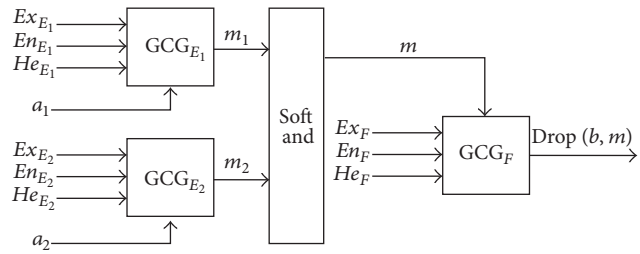


Figure 8: The DCSRGCG.

(6) Generate the distribution of drops (b, m).

The SCSRGCG implies an uncertainty transfer in the conceptual reasoning process. In the universal sets U_1 of the PGCG, the distribution of the certainty degree of m belongs to the specific value of a, whereas the certainty degree of m is the input of the PCGCG that generates the drop distribution (b, m) of the cloud specific value of b and the certainty degree of m. The processing of the certainty value of a to the certainty value of b is uncertain.

Definition 5. Assume the following rule:

$$\text{If } E_1, E_2, \text{ then } F, \tag{4}$$

where E_1 is the PGCG that generates the drop distribution (a_1, m_1) of the specific value of a_1 and the certainty degree of m_1, E_2 is the PGCG that generates the drop distribution (a_2, m_2) of the specific value of a_2 and the certainty degree of m_2, and F is the PCGCG that generates the drop distribution (b, m) of the cloud drop specific value of b and the certainty degree of m. The certainty degree m is obtained from the "soft and" of m_1 and m_2, which is called a double-condition single-rule GCG (DCSRGCG). The composition diagrams of two PGCGs and one PCGCG are shown in Figure 8.

The "soft and" is expressed via 2D GCM $(1, En_x, He_x, 1, En_y, He_y)$, which expresses the uncertainty of "and" of m_1 and m_2; the result of "and" is expressed by the certainty degree of m [29]. The degree of "soft and" can be realized by adjusting the values of En_x, He_x, En_y, and He_y when $En_x = En_y = 0$ and $He_x = He_y = 0$. Then, "soft and" becomes an "and" operation. The "soft and" output is presented in Figure 9, which shows the distribution of the drops and their certainty degree (x, y, m), with $x \in [0, 1]$, $y \in [0, 1]$.

The DCSRGCG can establish numerous conditions of single-rule GCG (MCSRGCG) based on its composition principle. The SCSRGCG and the MCSRGCG are stored in the rule library and applied in qualitative knowledge reasoning and intelligent control field.

3. Lateral Control Approach of an Intelligent Vehicle System

The control of an intelligent vehicle mainly comprises the control for speed and angle under conditions of car-following driving, lane-changing driving, and intersection driving, with car-following driving being the most common. Using

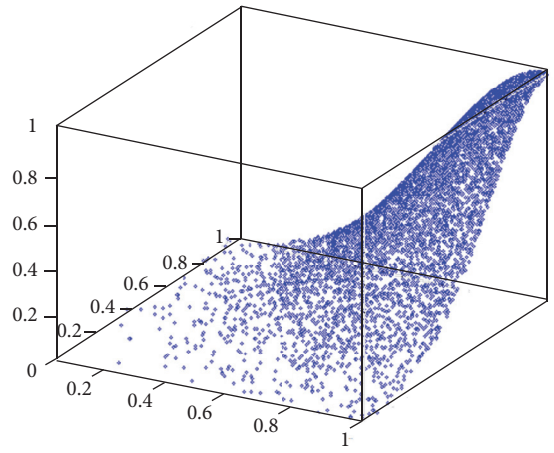


Figure 9: Quantitative transformation of the qualitative concept "soft and."

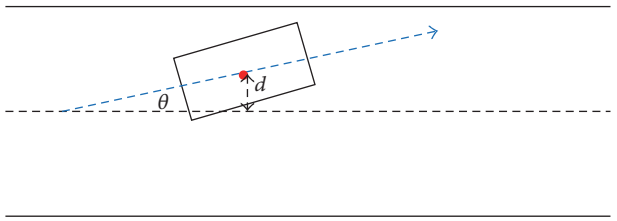


Figure 10: Distance d between the geometric center of a vehicle and the central axis of a lane, as well as the included angle θ between heading direction and the central axis.

this state as an example, vehicle speed and angles can be intelligently controlled once cloud reasoning and cloud control are introduced.

Under the condition of car-following driving, an intelligent vehicle should constantly adjust its speed according to obstacles, such as vehicles and pedestrians, while driving efficiently and avoiding collisions. The angle control of an intelligent vehicle aims to keep the car in the middle of the road, with an equal distance between the left/right lane line and the center of the vehicle while driving. Furthermore, the heading direction should remain in accordance with the lane line.

The speed and angle controls of intelligent vehicles are both typical double-conditional and single-rule controllers. In Figure 10, the solid line represents the real lane line, whereas the thick dotted line represents the central axis of the lane calculated according to two lane markings. The black dots represent the geometric center of an intelligent vehicle, which is represented by a rectangle. Vehicles achieve angle control by calculating the distance d between the geometric center of an intelligent vehicle and the central axis of the lane, as well as the included angle θ between heading direction and the central axis.

The input of the cloud controller is the distance d (in meters) between the geometric center of an intelligent vehicle and the direction of the central axis of the lane and the included angle θ (in degrees) between heading directions and lane line. The output is the steering wheel angle δ (in degrees). On the basis of a brief summary of an actual driving experience, several qualitative conclusions are drawn:

(1) If the intelligent vehicle does not veer off the middle of the lane and the heading direction of that vehicle remains in accordance with the axis of the lane, then the steering wheel should be returned to the zero position to keep the car moving straight forward. That is, if d and θ are near 0, δ should be 0, thereby enabling the vehicle to proceed normally.

(2) If the vehicle offsets to the right, then turn the wheel to the left to try and return the vehicle to the center of the lane. For a higher offset value, a greater adjustment angle of the steering wheel is necessary. That is, if d is more than 0, then δ is less than 0. A positive correlation exists between δ and d's absolute value.

(3) If the vehicle offsets to the left, then turn the wheel to the right and try to return the vehicle to the center of the lane. For a greater offset value, a greater adjustment angle of the steering wheel is necessary. That is, if d is less than 0, then δ is more than 0. A positive correlation exists between δ and d's absolute value.

(4) If the included angle between the heading direction and the central axis of the lane is greater than 0, which indicates that the vehicle is drifting toward the right front of the axis, then turn the wheel to the left and try to return the vehicle to the center of the lane. For a higher offset value, a larger adjustment angle of the steering wheel is necessary. If θ is more than 0, then δ is less than 0. A positive correlation exists between δ and θ's absolute value.

(5) If the included angle between the heading direction and the central axis of the lane is less than 0, which indicates that the vehicle is drifting toward the left front of the axis, then turn the wheel to the right and try to return the vehicle to the center of the lane. For a higher offset value, a larger adjustment angle of the steering wheel is necessary. If θ is less than 0, then δ is more than 0. A positive correlation exists between δ and θ's absolute value.

In the next section, we will describe the linguistic value sets of the input and the output, define the range of different linguistic values, and establish the cloud controller and its control rules based on the aforementioned five qualitative rules.

3.1. Cloud Controller Rules. The variables d, θ, and δ can be described using five qualitative concepts, namely, "positive more," "positive less," "near-zero," "negative less," and "negative more." The input and output variables define the five qualitative concepts and construct a corresponding cloud regulation generator.

The detailed car-following state and the speed control rules for intelligent vehicles are shown in Table 1. The rule set RS(d) of distance d between the geometric center of an

Table 1: (a) Rule sets RS(d) of distance d between the geometric center of an intelligent vehicle and the central axis of the lane. (b) Rule sets RS(θ) of included angle θ between heading direction and the central axis of the lane.

(a)

	Axis-line-distance d		Steering wheel angle δ
	Positive more		Negative more
	Positive less		Negative less
If	Zero	Then	Zero
	Negative less		Positive less
	Negative more		Positive more

(b)

	Axis-line-angle θ		Steering wheel angle δ
	Positive more		Negative more
	Positive less		Negative less
If	Zero	Then	Zero
	Negative less		Positive less
	Negative more		Positive more

Table 2: Parameter setting of the qualitative concepts of the composition of the speed control rules of an intelligent vehicle.

Parameter	Positive greater	Positive less	Zero	Negative less	Negative greater
RS(d)	(1.5, 0.2, 0.004)	(0.5, 0.15, 0.003)	(0, 0.08, 0.001)	(−0.5, 0.15, 0.003)	(−1.5, 0.2, 0.004)
RS(θ)	(10, 1.2, 0.02)	(5, 1, 0.02)	(0, 1, 0.01)	(−5, 1, 0.02)	(−10, 1.2, 0.02)
δ	(20, 3, 0.005)	(10, 2, 0.02)	(0, 2, 0.008)	(−10, 2, 0.02)	(20, 3, 0.05)

intelligent vehicle and the central axis of the lane is shown in Table 1(a). The rule set RS(θ) of the included angle θ between heading direction and the central axis of the lane is shown in Table 1(b). The parameter settings of the qualitative concepts in the rules are shown in Table 2.

3.2. Lateral Control Algorithm. The GCM and cloud reasoning can express human inference and decision and both exhibit strong robustness in solving the control problems of complicated systems. This study applies the GCM and the steering behavior imposed by cloud reasoning on drivers to build models. The model is shown in Figure 11.

The flowchart of the steering control algorithm shown in Figure 11 consists of two modules: steering wheel adjustment angle decision and steering wheel adjustment speed decision. The former comprises double-condition single-input and single cloud controllers, which are used to independently estimate the preview drift angle and the preview cornering distance. These controllers are called controller a and controller d, where u_a refers to the output of controller a and u_d refers to the output controller d. $u_a, u_d \in [-1, 1]$ denotes adapting the expectation direction of the adjustment angle of the steering wheel. The left represents the negative values, whereas the right represents the positive values. A value stands for an expectation degree. When the value is closer to 1, the expectation degree is higher, and vice versa. The adjustment angle decision will constantly determine the adjustment of the steering wheel in terms of controller output. The module of the steering wheel adjustment speed decision, which consists of double-condition single-input and single-cloud controllers, adopts the waterfall structure connection. It inputs the variables for the controller, which outputs the longitudinal velocity of the vehicle, and obtains the information of the adaptation speed of the wheel after adjusting speed. Negative values indicate adjusting the wheel to the left, whereas positive values indicate adjusting the wheel to the right. When the absolute value is closer to 1, the steering speed is faster. Conversely, the steering speed is lower when the absolute value is farther from 1.

4. Experiment Result and Analysis

4.1. Experiment Setup

4.1.1. Hardware Architecture of an Intelligent Vehicle System. The on-board sensor configuration of an intelligent vehicle comprises a radar sensor, a vision sensor, and a positioning sensor. The radar sensor consists of two separate Universal Transverse Mercator (UTM) single laser radars on the left and right of the body, a forward SICK single laser radar, a forward four-layer laser radar, and a backward millimeter wave radar. The vision sensor comprises three front-facing cameras, two rear-facing cameras, and two lateral cameras set in both rear-view mirrors. The positioning sensor consists of the Global Positioning System (GPS) and an inertial measurement unit (IMU), as is shown in Figure 12. This study is based on a "MengShi" intelligent vehicle, as is shown in Figure 13. All types of sensors are mainly applied to sense the surroundings of the vehicle for real-time acquisition of its location, posture, speed, and time.

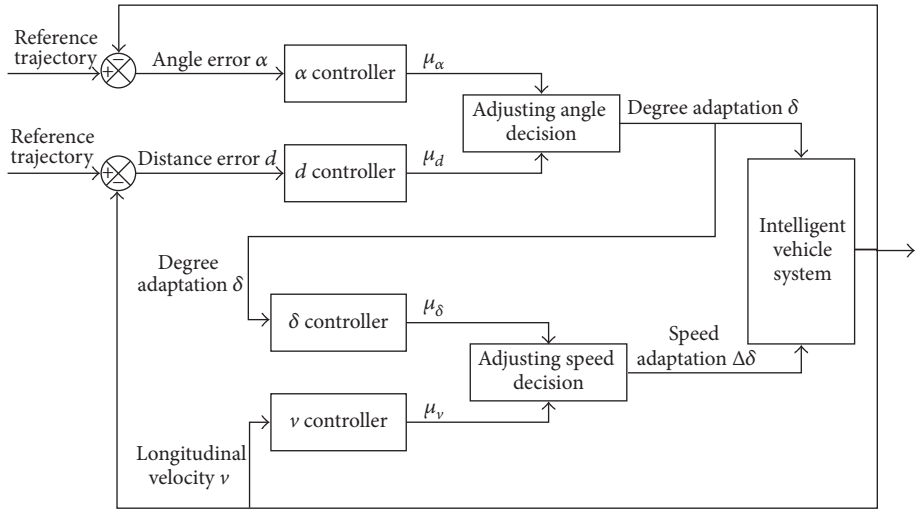


Figure 11: Lateral control algorithm flowchart.

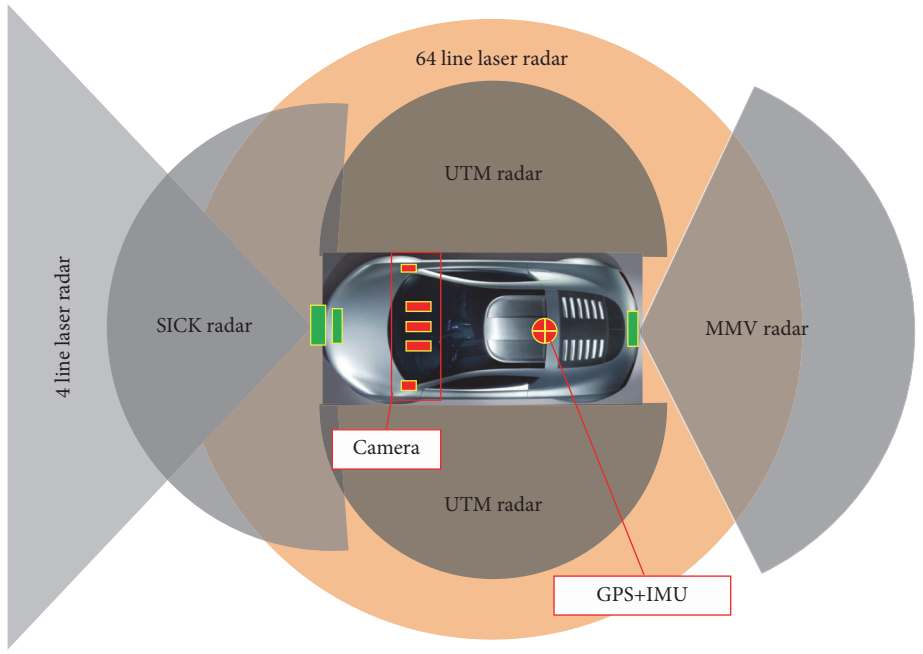


Figure 12: Experiment sensor configuration.

Figure 13: "MengShi" intelligent vehicle.

4.1.2. Software Architecture of an Intelligent Vehicle System. The design and development of intelligent vehicles are aimed at studying the key techniques of multi-interaction and collaborative driving based on visual and auditory information. The software architecture of intelligent vehicle systems is shown in Figure 14. This architecture comprises a human computer interaction (HCI) layer, a sensor and sensing layer, a planning and decision layer, and a control layer.

HCI Layer. This layer receives the touch commands and emergency braking instructions of the driver and relays them

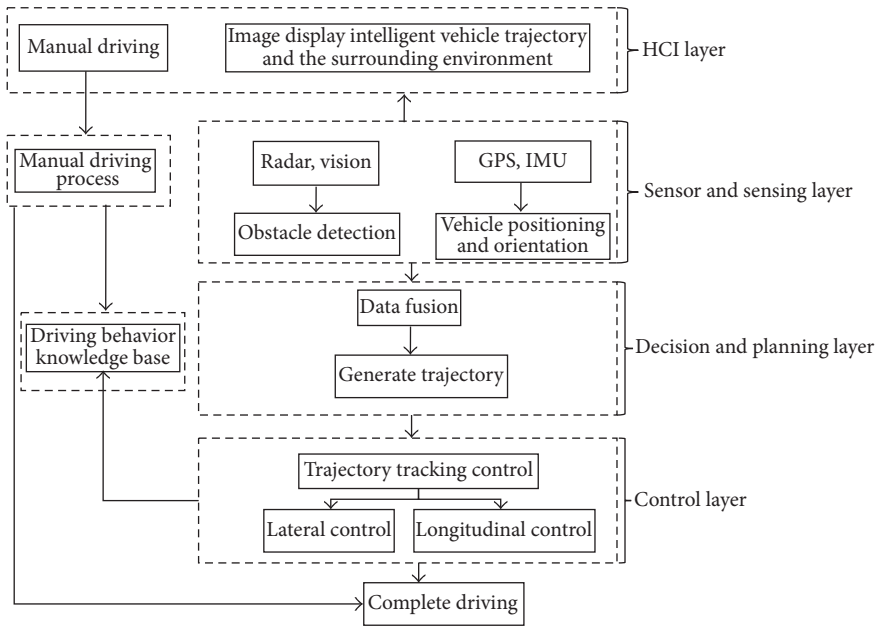


Figure 14: Software architecture of an intelligent vehicle system.

to the control layer. It simultaneously provides the driver with feedback information from the surroundings and other vehicles through sounds and images.

Sensor and Sensing Layer. This layer consists of a radar sensor, a vision sensor, a GPS sensor, and an IMU sensor. It focuses on completing the collection of sensor data. To realize the "plug and play" feature of the sensor, the standard data format of various sensors should be normative, which requires transforming the specific data format of the sensor to the standard format understood by an intelligent vehicle. The sensor data collected in this layer is delivered to the sensory module. The sensing layer focuses on sensor data analysis, road edge identification, obstacle detection, traffic sign detection, and body state estimation, which can facilitate the planning and decision of an intelligent vehicle.

Decision and Planning Layer. This layer focuses on path planning and navigation, which determine the driving pattern of an intelligent vehicle by analyzing environment data and vehicle data from the sensory module. This layer also determines the position of the vehicle in a detailed electronic map and generates the traveling track according to the coordinates of the target point. Human intervention and obstacles also influence the track.

Control Layer. This layer controls vehicles to enable them to proceed based on track data and current vehicle state. It also receives human instructions and performs acceleration/deceleration and steering operations. This layer directly outputs the control order to the accelerator, as well as the braking and steering controller, of the vehicle.

4.1.3. Experimental Environment. The Beijing-Tianjin Expressway, which spans the Taihu Toll Station and the Dongli Toll Station, covers 121 km of shuttle distance. Rain is moderate rain in Tianjin, with a small amount of water on the ground. The weather is rainy in the Tianjin section of the Beijing-Tianjin Expressway. When the sun occasionally shines, the weather remains sunny until reaching Beijing, where it is cloudy. The temperature outside the vehicle is 32°C, and that on the road is 40°C. Visibility is over 200 m. The experiment path is designated by the blue line in Figure 15.

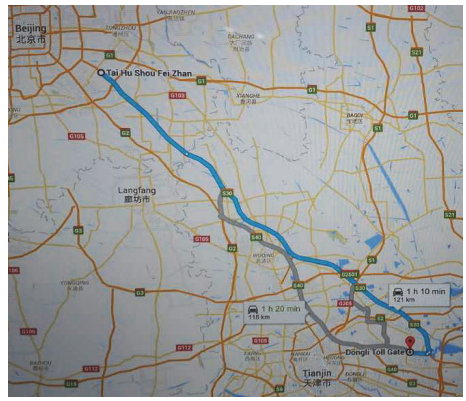


Figure 15: Experiments paths.

4.2. Experiment Result and Analysis. When the intelligent vehicle proceeds, θ marks the included angle between the intelligent vehicle and the lane line. Negative and positive values refer to drifting to the left and right, respectively. d marks the distance between the geometric center of the vehicle and the lane line. The instant velocity is obtained using GPS. The control angle of the steering wheel target is calculated through a decision algorithm and recorded.

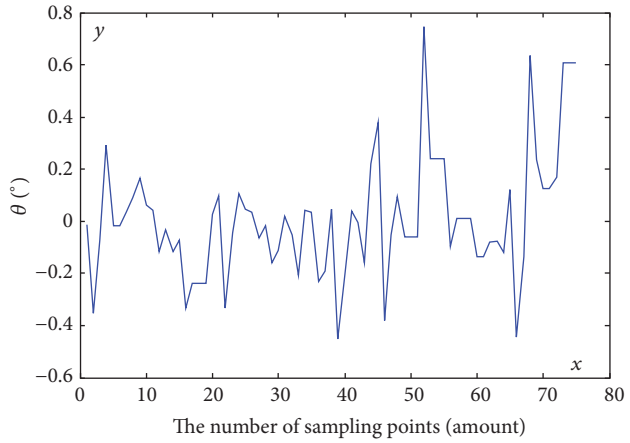


Figure 16: Variation curve of the included angle θ between the vehicle body and the lane line.

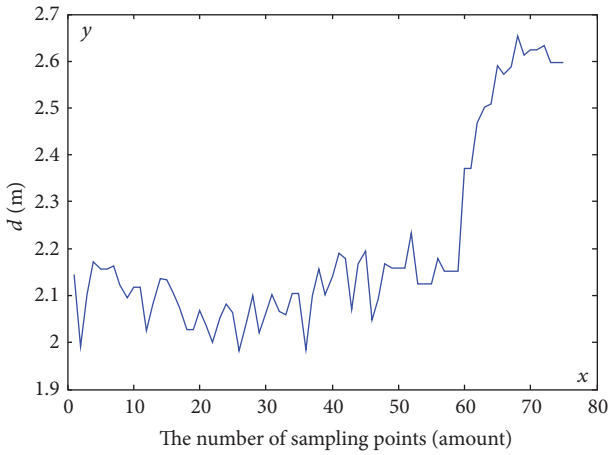


Figure 17: Variation curve of the distance d between the vehicle body and the lane line.

4.2.1. Analysis of Maintenance Situations on Lanes under Different Speeds

(1) Vehicle Speed Lower Than 80 km/h. When speed is lower than 80 km/h, a section of real-time data (75 in total) is randomly selected for analysis. Figure 16 shows the variation curve of the included angle θ between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the included angle θ (unit: degree). The value of the included angle θ between the lane line and the heading direction of the vehicle body should range from -0.5° to 0.8°, with a fluctuation range within 1.3°. Figure 17 shows the variation curve of distance d between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the distance d (unit: m). The distance d should range from 0.2 m to 0.8 m, with a fluctuation range within 0.6 m. The data further demonstrate that the second half of the driving drifts toward the left side of the lane line by a wide margin, but the overall situation remains good.

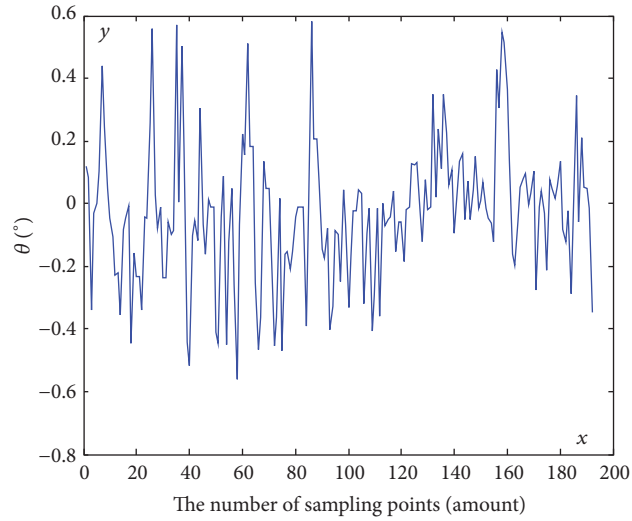


Figure 18: Variation curve of the included angle θ between the vehicle body and the lane line.

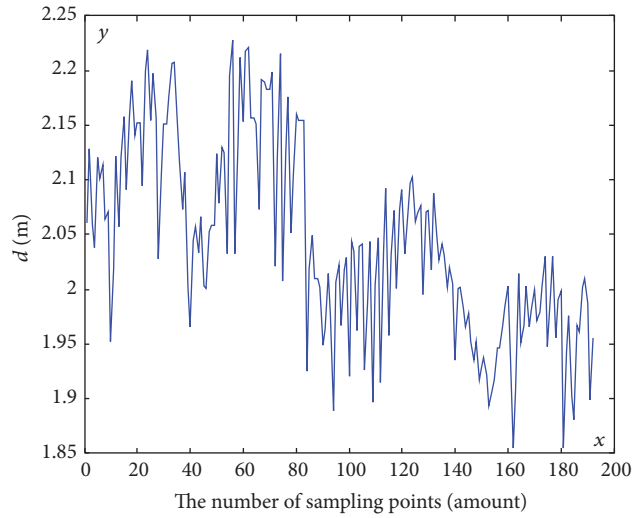


Figure 19: Variation curve of distance d between the vehicle body and the lane line.

(2) Vehicle Speed between 80 km/h and 90 km/h. When speed is between 80 km/h and 90 km/h, a section of real-time data (192 in total) is randomly selected analysis. Figure 18 shows the variation curve of the included angle θ between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount), The y-axis represents the included angle θ (unit: degree). The value of the included angle θ between the lane line and the heading direction of the vehicle body should range from -0.6° to 0.6°, with a fluctuation range within 1.2°. Figure 19 shows the variation curve of distance d between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the distance d (unit: meter). The distance d should range from -0.1 m to 0.4 m, with a fluctuation range within 0.5 m. The data show that the situation of the lane remains good.

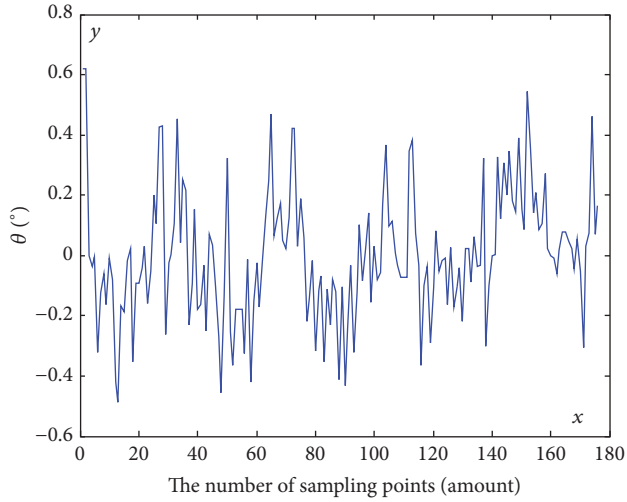


Figure 20: Variation curve of the included angle θ between the vehicle body and the lane line.

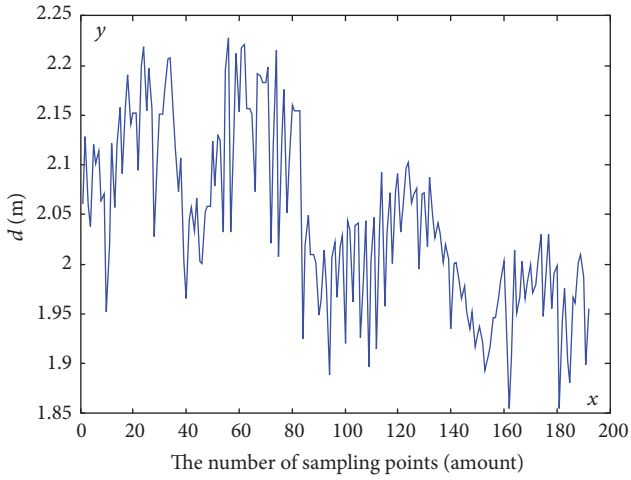


Figure 21: Variation curve of distance d between the vehicle body and the lane line.

(3) Vehicle Speed between 90 km/h and 100 km/h. When speed is between 90 km/h and 100 km/h, a section of real-time data (176 in total) is randomly selected for analysis. Figure 20 shows the variation curve of the included angle θ between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the included angle θ (unit: degree). The value of the included angle θ between the lane line and the heading direction of the vehicle body should range from -0.5° to 0.6°, a range of 1.1°. Figure 21 shows the variation curve of the distance d between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the distance d (unit: meter). The distance d should range from -0.25 m to 0.05 m, with a fluctuation range of 0.3 m. The data show that the situation of the lane remains good.

(4) Vehicle Speed Greater Than 100 km/h. When speed is greater than 100 km/h, a section of real-time data (96 in total) is randomly selected for analysis. Figure 22 shows the variation curve of the included angle θ between the vehicle body and the lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the included angle θ (unit: degree). The value of the included angle θ between the lane line and the heading direction of the vehicle body should range from -0.4° to 0.7°, a fluctuation range within 1.3°. Figure 23 shows the variation curve of the distance d between the vehicle body and lane line. The x-axis represents the number of sampling points (unit: amount). The y-axis represents the distance d (unit: meter). The distance d should range from -0.6 m to -0.2 m, with a fluctuation range within 0.4 m. The data show that the lane situation of the lane remains good.

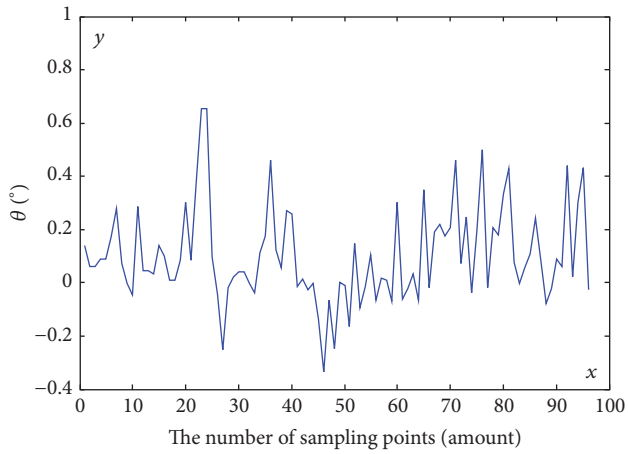


Figure 22: Variation curve of the included angle θ between the vehicle body and the lane line.

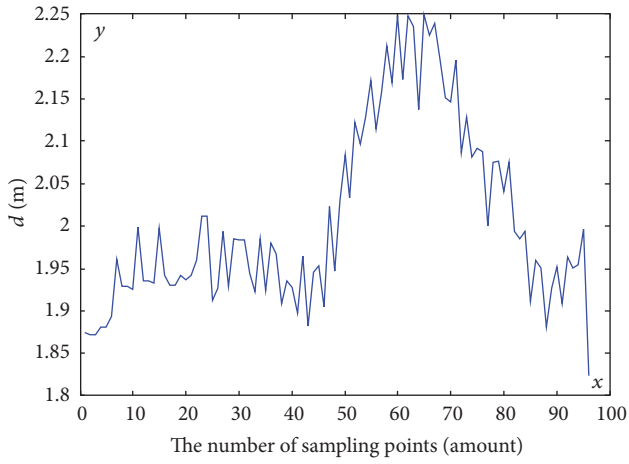


Figure 23: Variation curve of distance d between the vehicle body and the lane line.

4.2.2. Analysis of the Control Angle of the Steering Wheel. The curve graph of the control angle of the steering wheel is shown in Figure 24. The x-axis represents the traveled mileage of the vehicle (unit: km). The data recording interval is 50 ms. The y-axis represents the control angle of the steering wheel (unit: degree).

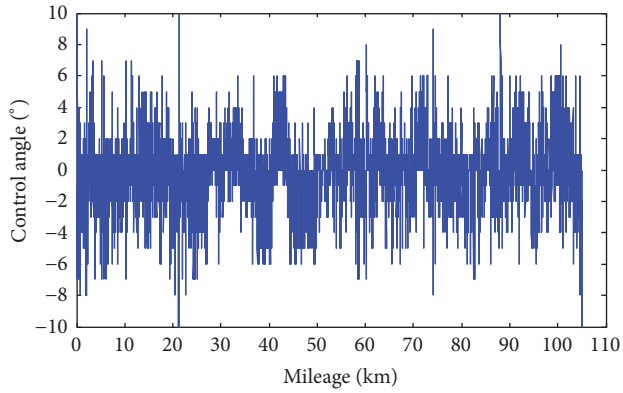


Figure 24: Curve graph of steering wheel control.

Steering to the right creates a negative value, whereas steering left creates a positive value. Approximately 81% of the angles of the steering wheel range from −3° to 3°, 3% range from −6° to 6°, and the maximum angle ranges from −7° to 7°. According to relevant laws, the floating range of a manually operated steering wheel ranges from −7.5° to 7.5°, which is a stable operation.

5. Conclusion

This study proposes a novel type of lateral control migration algorithm for intelligent vehicles. On the basis of the GCM and cloud reasoning, it also presents the qualitative concept cloud parameterization of the speed control rule for a vehicle on an expressway, designs a lateral control algorithm for an intelligent vehicle, and provides the speed control rules for different car-following conditions. The lateral controller of the vehicle, which is based on the GCM and the cloud reasoning algorithm, can be adapted to various speeds. Therefore, 81% of the angles of the steering wheel range from −3° to 3°, 3% range from −6° to 6°, and the maximum angle, which can achieve stable control, is within the range of −7° to 7°.

Competing Interests

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

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant nos. 61035004, 61273213, 61300006, 61305055, 90920305, and 61203366, the National Key Research and Development Program of China under Grant no. 2016YFB0100903, and the National High Technology Research and Development Program ("863" Program) of China under Grant no. 2015AA015401.

References

[1] D. J. Zhuang, Y. U. Fan, and Y. Lin, "The vehicle directional control based on fractional order PD~ μ controller," *Journal of Shanghai Jiaotong University*, vol. 41, no. 2, pp. 278–283, 2007.

[2] Y. Li, K. H. Ang, and G. C. Y. Chong, "PID control system analysis and design," *IEEE Control Systems Magazine*, vol. 26, no. 1, pp. 32–41, 2006.

[3] T. Hessburg and M. Tomizuka, "Fuzzy logic control for lateral vehicle guidance," *IEEE Control Systems*, vol. 14, no. 4, pp. 55–63, 1994.

[4] R. Choomuang and N. Afzulpurkar, "Hybrid Kalman filter/Fuzzy logic based position control of autonomous mobile robot," *International Journal of Advanced Robotic Systems*, vol. 2, no. 3, pp. 207–213, 2005.

[5] G. M. Scott, J. W. Shavlik, and W. H. Ray, "Refining PID controllers using neural networks," *Advances in Neural Information Processing Systems*, vol. 4, no. 5, pp. 746–757, 2008.

[6] R. J. Wai, "Tracking control based on neural network strategy for robot manipulator," *Neurocomputing*, vol. 69, no. 7–9, pp. 425–445, 2003.

[7] P.-J. He, K.-F. Ssu, and Y.-Y. Lin, "Sharing trajectories of autonomous driving vehicles to achieve time-efficient path navigation," in *Proceedings of the IEEE Vehicular Networking Conference (VNC '13)*, pp. 119–126, IEEE, Boston, Mass, USA, December 2013.

[8] K.-W. Min and J.-D. Choi, "Design and implementation of autonomous vehicle valet parking system," in *Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC '13)*, pp. 2082–2087, October 2013.

[9] S. M. Lavalle and J. J. Kuffner, "Rapidly-exploring random trees: progress and prospects," in *Algorithmic & Computational Robotics: New Directions*, pp. 293–308, CRC Press, 2010.

[10] E. Frazzoli, Z.-H. Mao, J.-H. Oh, and E. Feron, "Resolution of conflicts involving many aircraft via semidefinite programming," *Journal of Guidance, Control, and Dynamics*, vol. 24, no. 1, pp. 79–86, 2001.

[11] R. H. Byrne and C. T. Abdallah, "Design of a model reference adaptive controller for vehicle road following," *Mathematical and Computer Modelling*, vol. 22, no. 4–7, pp. 343–354, 1995.

[12] F. A. A. Cheein, C. De La Cruz, T. F. Bastos, and R. Carelli, "Slam-based cross-a-door solution approach for a robotic wheelchair," *International Journal of Advanced Robotic Systems*, vol. 7, no. 2, pp. 155–164, 2010.

[13] A. Ghanbari and S. M. R. S. Noorani, "Optimal trajectory planning for design of a crawling gait in a robot using genetic algorithm," *International Journal of Advanced Robotic Systems*, vol. 8, no. 1, pp. 29–36, 2011.

[14] P. C. Park and J. Miller, "Convergence properties of associtive memory storage for learning control system," *Automation and Remote Contorl*, vol. 1989, no. 2, pp. 254–286, 1989.

[15] S. Puntunan and M. Parnichkun, "Online self-tuning precompensation for a PID heading control of a flying robot," *International Journal of Advanced Robotic Systems*, vol. 3, no. 4, pp. 323–330, 2006.

[16] M. Doi and Y. Mori, "Generalized minimum variance control for time-varying system," *Transactions of the Society of Instrument and Control Engineers*, vol. 45, no. 6, pp. 298–304, 2011.

[17] D. R. Li, S. L. Wang, and D. Y. Li, *Cloud Model*, Spatial Data Mining, 2015.

[18] D. Y. Li, Y. C. Liu, and Y. Du, "Artificial intelligence with uncertainty," *Journal of Software*, vol. 15, no. 11, pp. 1538–1594, 2004.

[19] B. H. Cao, D. Y. Li, K. Qin et al., "An uncertain control framework of cloud model," in *Proceedings of the International Conference on Rough Set and Knowledge Technology (RSKT '10)*, pp. 618–625, Beijing, China, October 2010.

[20] D. Li, C. Liu, and W. Gan, "A new cognitive model: cloud model," *International Journal of Intelligent Systems*, vol. 24, no. 3, pp. 357–375, 2009.

[21] H. X. Gao, *Statistical Calculation*, Peking University Press, Beijing, China, 1995.

[22] D. Y. Li and Y. Du, *Artificial Intelligence with Uncertainay*, National Defence Industry Press, Beijing, China, 2005.

[23] H. Gao, J. Jiang, L. Zhang, L. Yuchao, and D. Li, "Cloud model: detect unsupervised communities in social tagging networks," in *Proceedings of the International Conference on Information Science and Cloud Computing Companion (ISCC-C '13)*, pp. 317–323, December 2013.

[24] H. B. Gao, X. Y. Zhang, T.-L. Zhang, Y.-C. Liu, and D.-Y. Li, "Research of intelligent vehicle variable granularity evaluation based on cloud model," *Acta Electronica Sinica*, vol. 44, no. 2, pp. 365–373, 2016.

[25] H. Liu and F. Sun, "Semi-supervised ensemble tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1645–1648, April 2009.

[26] H. P. Liu, F. C. Sun, and M. Y. Yu, "Vehicle tracking using stochastic fusion-based particle filter," in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS '07)*, pp. 2735–2740, San Diego, Calif, USA, 2007.

[27] K. C. Di, *The Theory and Methods of Spatial Data Mining and Knowledge Discovery*, Wuhan Technical University of Surveying and Mapping, 1999.

[28] Y. Du, *Research and Applications of Association Rules in Data Mining*, LA University of Science and Technology, 2000.

[29] D. Y. Li, Y. Du, G. D. Yin et al., "Commonsense knowledge modeling," in *Proceedings of the 16th World Computer Congress*, pp. 34–45, Beijing, China, August 2000.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 6876241, 9 pages
http://dx.doi.org/10.1155/2016/6876241

Research Article

A Smart High-Throughput Experiment Platform for Materials Corrosion Study

Peng Shi, Bin Li, Jindong Huo, and Lei Wen

National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing, China

Correspondence should be addressed to Peng Shi; shipengustb@sina.com

Received 6 August 2016; Accepted 4 October 2016

Academic Editor: Wenbing Zhao

Materials corrosion study is based on plenty of contrast experiments. Traditional corrosion experiments are time-consuming and require manual corrosion grade evaluating during the experiment. To improve the efficiency of experiment, a high-throughput experiment platform is designed and accomplished. The platform mainly consists of high-throughput corrosion reaction facility, data acquisition system, and data processing system. The corrosion reaction facility supports high-throughput materials corrosion reactions under various conditions. The data acquisition system is mainly responsible for capturing the images of samples' surface, collecting electrochemical signals, and storing them into the computer in real time. The data processing system treats the acquired data and evaluates the degree of materials corrosion in real time by program automatically. The platform not only reduces the occupation of the equipment but also improves the efficiency of sample preparation and experiment occurrence. The experimental data shows that the platform can accomplish high-throughput corrosion contrast experiment easily and reduce the time cost obviously.

1. Introduction

Corrosion is one of the most popular materials' failure causes. It will significantly reduce materials' strength, plastic, and toughness properties, shorten the service life, and cause catastrophic accidents [1]. To investigate the factors and mechanisms of materials corrosion, a lot of experiments must be conducted. A small change of the experimental parameters often means a series of new similar experiments, which will undoubtedly consume a large amount of resources. To improve the efficiency of this kind of experiments, high-throughput method is adopted.

The idea of high-throughput originates from combinatorial chemistry [2], aiming to shorten the product development period and reduce costs by large-scale chemical synthesis technology. Its essence is to accomplish large amount of repetitive experiments in parallel and obtain experimental results through the novel design of experiment scheme and equipment. High-throughput characterization technology has been widely used in gene sequencing, drug screening, and so on. In recent years, new materials development begins to adopt high-throughput technology to improve development efficiency [3]. However, there are few reports on high-throughput materials experiment method. In this paper, we try to improve the efficiency of materials corrosion experiment by high-throughput technology.

A high-throughput corrosion experiment platform has been designed and accomplished. The platform mainly comprises high-throughput corrosion reaction facility, data acquisition system, and data processing system. The corrosion reaction facility supports high-throughput materials corrosion reactions under various conditions. The data acquisition system is mainly responsible for capturing the images of samples' surface, collecting electrochemical signals, and storing them into the computer in real time. The data processing system treats the acquired data and evaluates the degree of materials corrosion in real time by program automatically.

The platform has accomplished some high-throughput materials corrosion experiments, which not only reduces the occupation of the equipment but also improves the efficiency of sample preparation and experiment conduction. Different

from traditional artificial corrosion evaluation, the platform can treat the collected data and automatically generate corrosion evaluation results. The contribution of our method is that it provides the possibility of large-scale materials corrosion experiment and version-based data analysis.

The rest of the paper is arranged as follows. Section 2 introduces the related works about high-throughput experiment and corrosion image processing. Section 3 describes the structure and implementation of the platform. The key algorithm and programing method are presented in Section 4. The experimental procedure and result analysis are shown in Section 5. Section 6 analyzes the performance of our method. At last, Section 7 summarizes our contributions and points out the future work.

2. Related Works

2.1. High-Throughput Technology. High-throughput is a technology to execute parallel task with multicharacterization methods. It has been applied in several fields, including high-throughput sequencing, high-throughput screening, and combination technology. High-throughput sequencing [4] brings great promotion on the development of biological gene sequencing. It can measure millions of DNA molecules' sequence at the same time. This makes it possible to analyze the transcriptome and the genome of a species in detail. High-throughput screening [5] is a method for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry. It allows a researcher to quickly conduct millions of chemical, genetic, or pharmacological tests with robotics, data processing and control software, liquid handling devices, and sensitive detectors [6]. Combination technology, with multichannel parallel synthesis and high-throughput rapid characterization, can synthetize samples of different compositions fleetly by finite steps, investigate its structure and properties efficiently, and develop new materials with required properties finally [7].

2.2. Corrosion Image Processing and Evaluation. Image processing is a technology to analyze images and obtain some desired features by computer. It generally includes image compression, enhancement, restoration, matching, description, and recognition. Image description, matching, and recognition are usually adopted to accomplish some smart applications.

During the study of materials corrosion, the appearance of surface is important to evaluate the corrosion degree. The idea of image processing can be used in the grade evaluation of materials corrosion. In 1981, Itzhak et al. scanned the surface of AISI 304 material by digital scanner, which has been soaked for 20 minutes in 50°C 10% $FeCl_3$ solution [8]. The corrosive pitting is counted and the corrosion rate is calculated by computer program according to the scanned images. Codaroa et al. analyzed the appearance of pitting of Al-Ti alloy and gave a quantitative description method for pitting. The method can be utilized to represent the evaluation procedure of corrosion [9]. Wang et al. collected the morphology of sea water corrosion of carbon steel

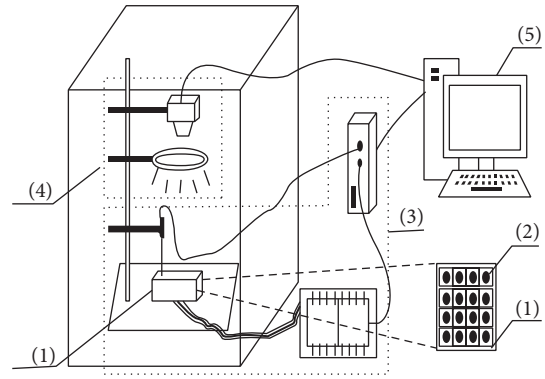


Figure 1: The structure of high-throughput corrosion experiment platform ((1) corrosion reaction facility, (2) specimen, (3) electrochemical signal acquisition system, (4) image acquisition system, and (5) computer).

[10]. They adopted grey correlation method to establish the relationship between the corrosion degree and the apparent color and edge. Xu et al. investigated the relationship between the apparent grey value of image and the depth of corrosive pitting in the specimen by fractal dimension method [11]. They found out that the relationship was nearly linear. In this paper, corrosion images are automatically processed in real time to evaluate the corrosion grade of specimens.

3. Platform Structure

With the idea of high-throughput, a smart materials' corrosion experiment platform is designed. It consists of high-throughput corrosion reaction facility, data acquisition system, and data processing system, shown in Figure 1. All the systems are connected to transfer signal and data.

3.1. High-Throughput Corrosion Reaction Facility. High-throughput corrosion reaction facility indicates the container where the corrosion reaction occurs. We design two kinds of reaction units for high-throughput corrosion reaction: single-solution reaction unit and multiple-solution reaction unit.

3.1.1. Single-Solution Reaction Unit. The appearance of single-solution reaction unit is shown in Figure 2. It uses corrosion resistant materials, such as epoxy resin, to package several specimens into one component. Single-solution reaction unit can only investigate the corrosion behaviors of different specimens in the same solution. The specimens in one component can be distinguished from materials, roughness, and other properties. Since the specimens are packaged, they can be treated together to improve the sample making efficiency, such as polishing and surface processing.

3.1.2. Multiple-Solution Reaction Unit. The appearance of multiple-solution reaction unit is shown in Figure 3. It is a reaction vessel consisting of a set of parallel tactic groove liquid pool arrays. Each liquid pool can contain different

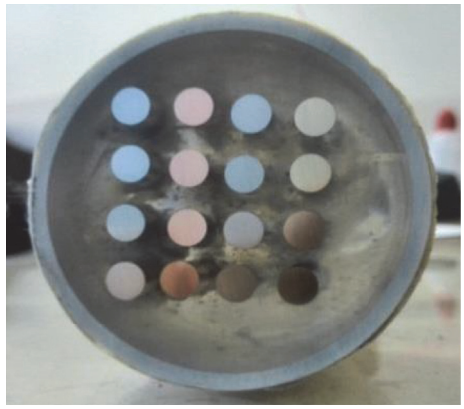


Figure 2: Single-solution reaction unit.

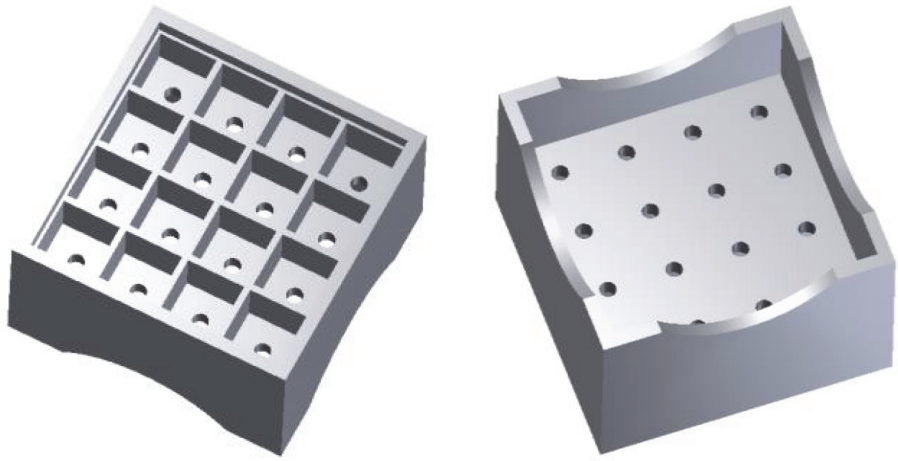


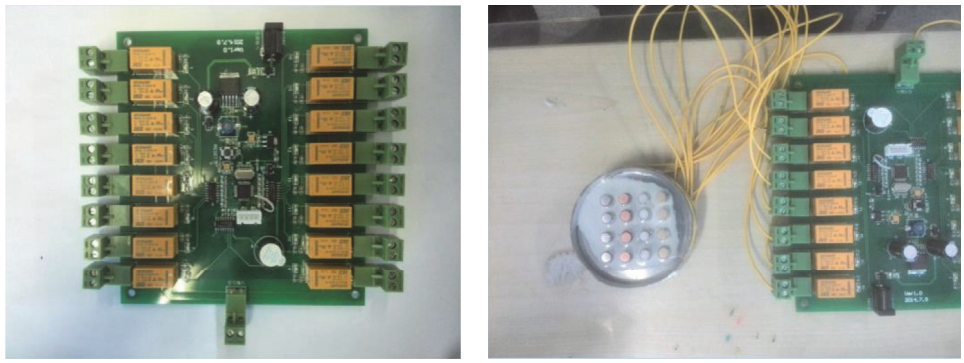
Figure 3: Multiple-solution reaction unit.

solutions during one experiment. Every specimen can be arranged into one liquid pool. After one specimen is put in, the bottom of the pool should be sealed up to avoid the leaking of solution. During the experiment procedure, each liquid pool can be plugged with an electrode. Thus, the electrochemical signals of every specimen can be collected independently. The advantage of multiple-solution reaction unit is that different specimens can be tested in different solutions at the same time. One experiment may produce independent image and electrochemical data of every specimen. It can obviously improve the corrosion reaction efficiency.

3.2. Data Acquisition System

3.2.1. Electrochemical Signal Acquisition System. Electrochemical signal is an important characterization parameter during materials corrosion reaction. It is usually detected by electrochemical workstation, representing the current and potential of the specimen and solution. The electrochemical signal acquisition system is composed of an electrochemical workstation, a reference electrode, and a multiplexer. The electrochemical workstation and reference electrode are working together to collect the signal of current and potential. In our high-throughput materials corrosion experiments, the two-electrode system was used to form a closed loop with a working electrode and a reference electrode. We use "Interface 5000" electrochemical workstation from Gamry® Company. The working electrode is a bolt type materials' electrode, and the reference electrode is commonly Ag/AgCl electrode used in the study of electrochemical corrosion. Ag/AgCl electrode can be utilized as a probe part of reference electrode in the solution.

Since high-throughput experiment needs to get the electrochemical signals of all the specimens, a multiplexer is designed to accomplish the function of parallel signal collection. The multiplexer is developed based on the circuit board. It consists of several input ports and one output port. Here we show a multiplexer with 16 input ports. All the specimens in the corrosion reaction unit can be connected with the input ports. The output port is connected with an electrochemical workstation to analyze the signal from the input ports. After circuit programing, we can realize the connecting between one input port and the output port at one moment. Thus, the multiplexers can take turns to query 16 input ports in accordance with the way specified by



(a) (b)

Figure 4: Multiplexer and its connection mode. (a) Multiplexer. (b) Connecting with specimens.

Table 1: Daheng camera parameters.

Name	Parameter
Model	MER-500-7UC
Interface	Mini USB 2.0
Resolution	2592 (H) × 1944 (V)
Frame rate	7 fps
Sensor	1/2.5″ CMOS
Pixel size	2.2 μm × 2.2 μm
Spectrum	Black-and-white/color

the program. The appearance of the multiplexer is shown in Figure 4(a). Figure 4(b) shows the connecting mode with the experimental sample. With the help of the multiplexer, one single channel electrochemical workstation can accomplish the function of multichannel electrochemical workstation. Although there may have time delay on signal collection, it hardly influence the result because of slow corrosion action.

3.2.2. Image Acquisition System. Since the surface appearance of specimen is one important characterization for corrosion study, specimen surface image must be collected continuously. The image acquisition system is composed of a camera, a ring light, and a light shield.

We choose "MER-500-7UC" camera from Daheng® Graphics Company. The specific parameters of the camera are shown in Table 1. The camera adopts USB 2.0 port to connect with computer and provides programing interface to control its action. Its size is only 29 mm ∗ 29 mm ∗ 29 mm and it is easy to be used in various environments.

For enough light on the specimens, we take LED ring light as a light source. It can supply an even light for the measured object. To avoid the influence from the environmental light, a light shielding cover was used. The stable light condition makes the direct comparison of images at different stages possible. Light shield selection was relatively simple, as long as it can cover the whole system and block the environmental light in.

3.3. Data Processing System. Data processing of our platform is automatically accomplished by programing, which is running on a computer. The data produced from the platform mainly includes the electrochemical signals and specimen surface images. The program can separate the electrochemical data of each specimen from the electrochemical workstation file, which contains the mixed data of all the specimens. The program can also process the surface image of specimens. The corrosion grade of each specimen is evaluated by image processing based program. The details of the method and algorithm will be introduced in Section 4.

4. Algorithm and Programing

4.1. Electrochemical Signal Extracting. In our high-throughput experiment, the electrochemical data of all specimens are collected together. The electrochemical workstation collects signals sequentially and stores them into one DAT format file. To analyze the electrochemical signal of every specimen, the signal data of every specimen should be extracted from the DAT file. Based on the time stamp in the file and multiplexer, each specimen's electrochemical data can be attained by a program.

To avoid the error of data collecting, we set the data collecting with a higher frequency than the realistic demand. In the electrochemical data split program, the data with big error is picked out. An average data value of the remaining data with small error is adopted as the electrochemical signal value of the specimen by the program. After the program treatment, the data in DAT file is separated into several electrochemical signals. Every group of independent signals is corresponding to one specimen.

4.2. Image-Based Smart Corrosion Evaluation. The surface morphology and corrosion grade evaluation are important for the study of corrosion mechanism. During traditional corrosion experiment, the specimen corrosion grade is evaluated manually. With our smart high-throughput experiment platform, the specimen corrosion grade can be evaluated by the program automatically.

In the past decades, image processing technology has been successfully applied in corrosion description and evaluation. Most of the related works focused on the analysis of grey image. Here we give a quick corrosion degree evaluation method for color corrosion image. For the RGB image, each pixel point has three color channels: red, green, and blue. Each color channel can take any value from 0 to 255 (with 8 bits' format), so the whole color space is a total of 256^3 (about 16 million) kinds of color combination. The calculation of comparing the distribution of these 16 million combinations is too large, so we design a quick method to evaluate the color image by characteristic value. The RGB value from 0 to 255 is divided into four ranges: 0 to 63 is the first range; 64 to 127 is the second range; 128 to 191 is the third range; and 193 to 255 is the fourth range. Each color channel can be divided into four districts, so 4^3 (64) kinds of color combinations can be reformed. The 64 kinds of combinations can be regarded as a space of 64 dimensions.

Here we define the concept of characteristic vector C of an image. C is a 64-dimension vector. The element of C is the total number of pixels with the color of certain dimension, denoted as c_i. Thus a color image can be represented by a 64-dimensional vector. For example, the characteristic vector of the left top brass specimen in Figure 8(a) can be denoted as

$$C = (7414, 230, 0, 0, 8, \ldots, 109, 0, 0, 3415, 53929). \quad (1)$$

The details of RGB district and characteristic vector are shown in Table 2. The elements in column "c" compose a vector, that is, the characteristic vector. It can be regarded as a "fingerprint" of an image. The similarity between two images can be determined by comparing their characteristic vectors [12]. In this way, we can improve the efficiency of image similarity determination while ensuring the relative uniqueness.

To quickly evaluate the corrosion grade of a specimen, we establish a standard corrosion image database. The database consists of a series of standard corrosion images with manual corrosion grade evaluation. The characteristic vectors of the images are also stored in the database. The main idea of quick corrosion grade evaluation is to specify one standard image's grade to an image, whose characteristic vector is the nearest to the standard image. The procedure of corrosion grade evaluation based on characteristic vector similarity mainly consists of the following five steps:

(1) Load the corrosion images to be evaluated.

(2) Calculate the characteristic vector of the image to be evaluated.

(3) Compare the calculated characteristic vector with that of the standard image one by one.

(4) Find out the standard image with the nearest characteristic vector.

(5) Output the corrosion grade of the standard image as that of the input image.

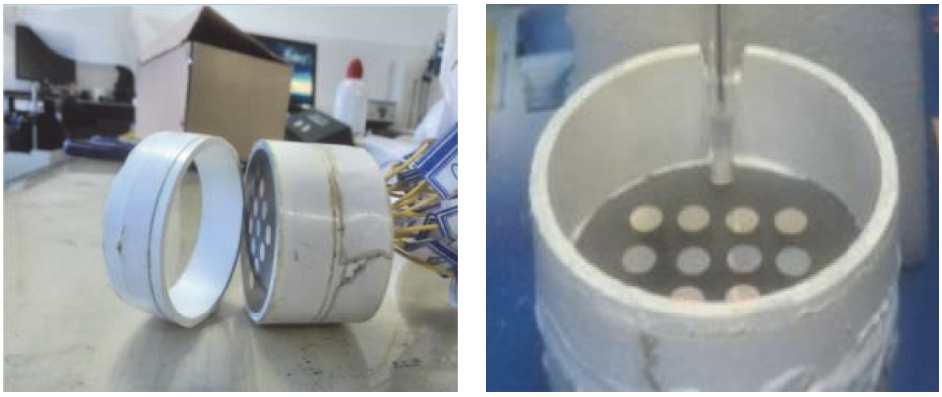
Based on the procedure above, we have developed a program to calculate corrosion grade of collected images. Consequently, the high-throughput experiment platform can output the corrosion grade of every specimen in real time.

Table 2: RGB and characteristic vector of image.

R	G	B	c
0	0	0	7414
0	0	1	230
0	0	2	0
0	0	3	0
0	1	0	8
0	1	1	372
0	1	2	88
0	1	3	0
0	2	0	0
0	2	1	0
0	2	2	10
0	2	3	1
0	3	0	0
1	0	0	891
1	0	1	13
1	0	2	0
1	0	3	0
1	1	0	592
1	1	1	3462
1	1	2	355
1	1	3	0
1	2	0	0
1	2	1	101
1	2	2	882
1	2	3	53110
1	3	0	11053
2	0	0	1146
2	0	1	0
2	0	2	0
2	0	3	0
2	1	0	2552
2	1	1	9040
2	1	2	47
2	1	3	0
2	2	0	0
2	2	1	8808
2	2	2	8
2	2	3	0
2	3	0	16
3	0	0	11
3	0	1	0
3	0	2	0
3	0	3	0
3	1	0	856
3	1	1	1376
3	1	2	0
3	1	3	0
3	2	0	0
3	2	1	3650
3	2	2	6260
3	2	3	109
3	3	0	0

5. Experiments and Data Analysis

With the smart high-throughput experiment platform, we have conducted several experiments of materials corrosion in NaCl solution.



(a) (b)

Figure 5: The solution container on single-solution corrosion unit. (a) Sticking method. (b) Solution container with an electron.

Table 3: The polishing sandpaper types and surface roughness.

Sandpaper type	200#	400#	800#	1500#
Surface roughness/μm	75.0	35.0	21.8	12.6

5.1. Experiment Preparation. Four kinds of materials, including aluminum, brass, copper, and steel, are selected as the objects to be investigated. Every kind of material is made into four specimens as parallel reference separately. They are made into bar sample, with 5 mm diameter and 20 mm length. The intersecting surface is polished as the corrosion surface, by 200#, 400#, 800#, and 1500# sandpaper, respectively. Thus the effect of surface roughness on materials corrosion can be studied. The polishing sandpaper types and surface roughness are shown in Table 3.

The corrosion experiment is conducted in 3.5% NaCl solution. Here we use the single-solution corrosion reaction unit as reaction container. After the specimens are polished, a pipe made of PTFE (Polytetrafluoroethylene) is stuck to the package surface by 704 silicone rubber. Then the NaCl solution can be contained in the pipe space. The appearance of stuck single-solution unit is shown in Figure 5(a). A reference electrode is plugged into the container to collect the electrochemical signals, shown in Figure 5(b).

5.2. Experimental Data

5.2.1. Electrochemical Signal. The experiment monitors the open circuit potential signal of the specimens by two-electrode system. Ag/AgCl electrode is adopted as the reference electrode of the corrosion system. The reference electrode is connected with the electrochemical workstation. The specimens can be regarded as working electrode. All the working electrodes are connected with the input ports of the multiplexer one by one with extension wire. The output port of the multiplexer is directly connected with the Gamry electrochemical workstation. Thus the specimens, electrochemical workstation, multiplexer, and reference electrode compose a connected loop.

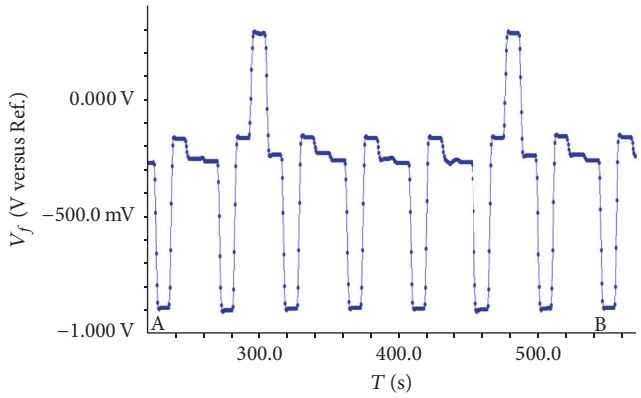
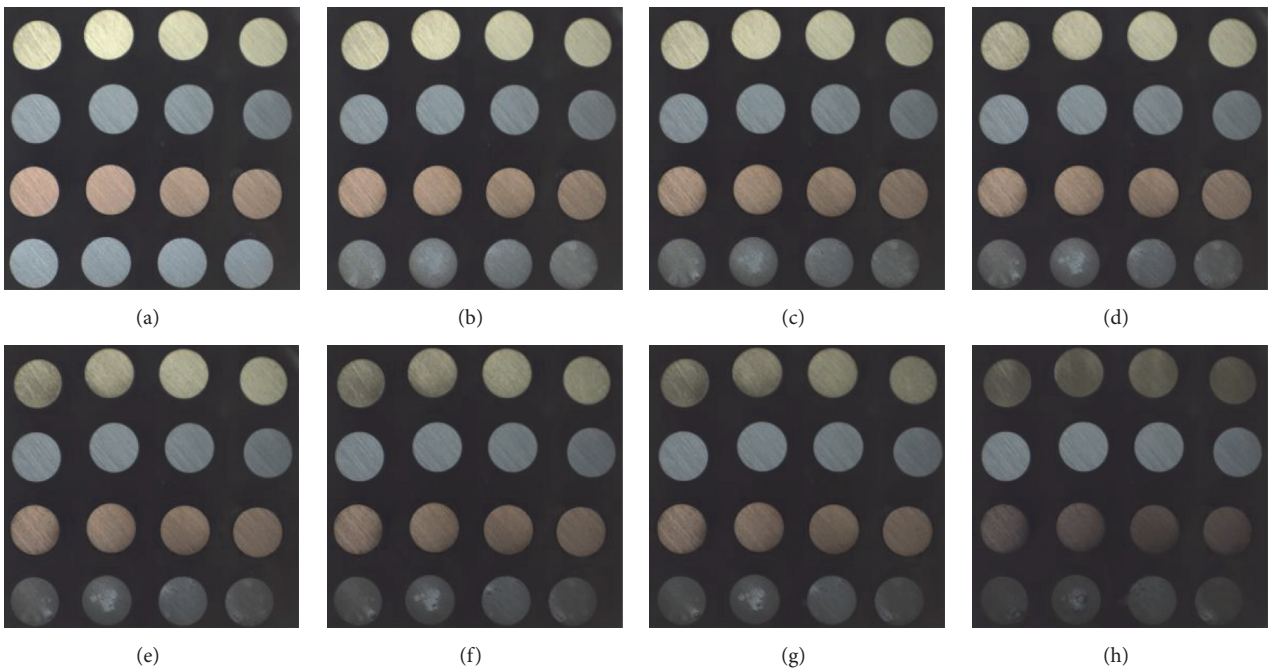


Figure 6: Open circuit potential of a full sampling period.

In order to obtain stable data output, we maintain 5-second period of switching on with each specimen by multiplexer control program. At the same time, the Gamry potential signal scanning period is set to 0.5 seconds. Consequently, the signal scanning may complete about 10 times in each specimen's switching on interval. After removing the data with big offset, the average of the remaining data is calculated. The average value is specified as the potential value of the specimen in the switching on period. The platform can accomplish a complete scan circle for all the specimens in 80 seconds. The data of one complete circle is shown in Figure 6. Every point in the diagram is original data from the electrochemical workstation.

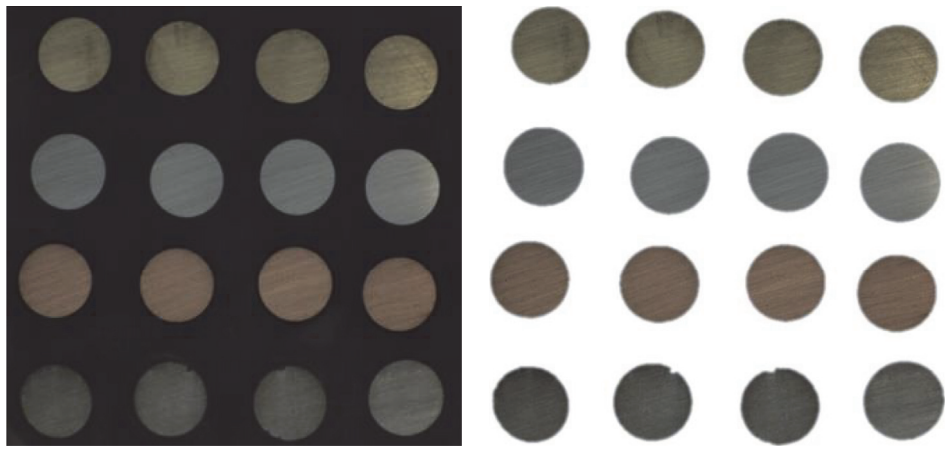
5.2.2. Surface Image. During the experiment, the image acquisition system records the specimens' surface image automatically. Since the corrosion action is not fast, the system captures the surface image once per hour by the program. The corrosion morphology images of the specimens within 0~7 hours are shown in Figure 7. From the images, researchers can investigate the corrosion morphology. The corrosion morphology comparison on different materials and different corrosion time can be executed easily.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7: High-throughput corrosion images of one-hour interval. (a) $t = 0$ h, (b) $t = 1$ h, (c) $t = 2$ h, (d) $t = 3$ h, (e) $t = 4$ h, (f) $t = 5$ h, (g) $t = 6$ h, and (h) $t = 7$ h.



(a) (b)

Figure 8: High-throughput image segmentation. (a) Image before segmentation. (b) Single specimen image.

5.3. Data Processing and Analysis. The data produced by high-throughput experiment is automatically processed in real time. Electrochemical signal processing is to separate the original data into 16 independent signals of every specimen. Image processing is mainly to evaluate the corrosion grade of every specimen.

5.3.1. Electrochemical Signal Processing. The electrochemical data of all the specimens collected by the electrochemical workstation are stored in a DAT format file. According to the program setting, the workstation collects data twice in one second, shown in Figure 6. The potential data of all the specimens is in one file. The data of each specimen is separated into several electrochemical signals by the program. The program considers two factors. One is the collecting time, by which the certain specimen can be located. The other is the data value. Because of the error of detecting, there may be some data with big offset. The program removes the offset data and calculates the average of the remaining normal data. The average value is specified as the potential value of the specimen in the switching on period. Thus every specimen contains a series of independent potential data.

The open circuit potential data after processing is shown in Figure 9. It presents the electrochemical signals of high-throughput experiment of carbon steel, brass, copper, and

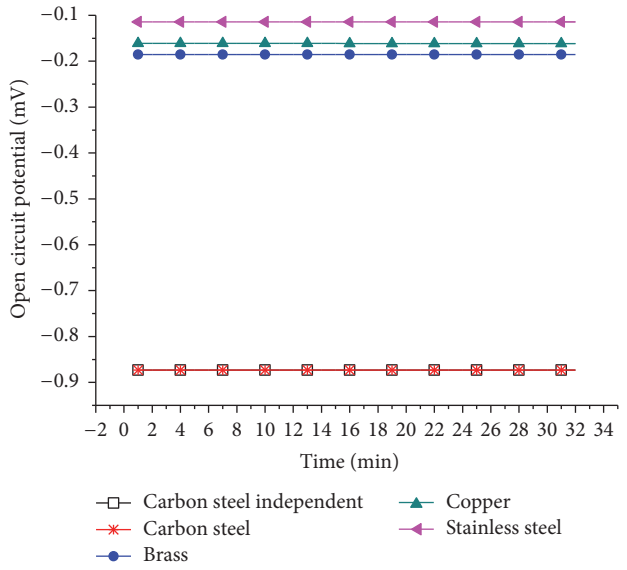


Figure 9: Electrochemical signals after processing.

stainless steel. To verify the correctness of high-throughput experiment, we conduct a carbon steel corrosion experiment with traditional independent experiment. The data of the independent experiment is also shown by black squares in Figure 9. From the comparison result, the independent experimental potential data is consistent with the data obtained by high-throughput experiment of the same materials in the same condition. Therefore, the phenomenon shows the correctness of high-throughput experiment in one respect.

5.3.2. Image Processing and Corrosion Grade. Since high-throughput experiment produces images with multiple specimens, the image of every single specimen should be extracted. Based on image edge detection method, the image of single specimen can be easily attained. The high-throughput image segmentation is shown in Figure 8. After the treatment by the program, the image of every specimen is extracted, shown in Figure 8(b).

Based on the standard GB/T 6461-2002, China standard of corrosion evaluation, the corrosion grade can be evaluated. The standard proposes two basic parameters: corrosion area ratio and corrosion grade, to evaluate the corrosion grade. The corresponding relationship between corrosion grade and corrosion area ratio is shown in Table 4. Corrosion area ratio is the percentage ratio between corrosion surface area and total surface area. The corrosion grade is from 1 to 10. The higher the grade is, the more serious the corrosion status is.

To verify the correctness of corrosion evaluation, the corrosion images are also submitted to an expert of materials corrosion. The evaluation result of 7 h experiment by our program and the expert is shown in Table 5. From the table, it can be seen that there are some differences of the evaluated grade. It is because the expert uses his/her eyes to evaluate the corrosion grade by his/her experiences. After the confirmation of the expert, the program produced result is better than manual result.

Table 4: The corresponding relationship of grade value and corrosion area ratio.

Corrosion area ratio/%	Corrosion grade value
No corrosion	10
$0 < A \leq 0.1$	9
$0.1 < A \leq 0.25$	8
$0.25 < A \leq 0.5$	7
$0.5 < A \leq 1.0$	6
$1.0 < A \leq 2.5$	5
$2.5 < A \leq 5.0$	4
$5.0 < A \leq 25$	3
$25 < A \leq 50$	2
$50 < A$	1

Table 5: Corrosion evaluation result of 7 h experiment.

Specimen position in the sample	Grade value by program	Grade value by expert
1,1	7	7
1,2	7	7
1,3	7	7
1,4	6	7
2,1	6	6
2,2	6	6
2,3	6	6
2,4	5	6
3,1	8	8
3,2	7	8
3,3	8	8
3,4	8	8
4,1	2	2
4,2	3	3
4,3	2	2
4,4	4	4

Table 6: Time cost comparison of one-by-one, parallel and high-throughput experiment mode.

Steps	One-by-one/minute	Parallel/minute	High-throughput/minute
Polish	960	960	240
Encapsulation	0	0	30
Experiment	6720	420	420
Handle	140	140	140
Total	7820	1520	830

6. Performance Analysis and Discussion

Compared with traditional corrosion experiments, the high-throughput experiment platform shows high efficiency, especially in time cost. Table 6 shows the time and equipment cost comparison of a typical high-throughput corrosion experiment with parallel experiment and one-by-one experiment.

Here, the high-throughput experiment with 16 specimens is taken as an example. From the data in Table 6, it is clear that the high-throughput experiment costs much less time than traditional experiment way. The parallel experiment and high-throughput experiment cost less time than one-by-one experiment. However, more testing equipment is occupied by parallel experiment than one-by-one experiment and high-throughput experiment. Although the high-throughput method takes some time for sample packaging, it reduces the sample polishing time. The parallel experiment costs more time for polishing because the high-throughput sample can be polished together. From the total time cost data in Table 6, it is obvious that high-throughput experiment has the highest efficiency. From the analysis above, it can be concluded that the high-throughput experiment method designed in this paper has a great advantage in the equipment occupation and the overall time consumption.

7. Conclusion and Future Work

In this paper, we propose a smart high-throughput experiment platform for the research of materials corrosion. Intelligent analysis of the corrosion of materials was successfully accomplished by the experimental platform. The corrosion grade can be calculated by the program automatically. The calculated results are matched closely with the manual analyzed results. Unfortunately, for the experiment with little surface changing, the image-based corrosion evaluation may not work well.

In the future, we will try to collect other types of data during materials corrosion, such as microstructure, solution changes, and other characterization data. By analyzing the new types of data, it is expected to accomplish more comprehensive understanding of the mechanisms of materials corrosion. At the same time, through adding the new type of data to the standard library, the corrosion evaluation results can be more accurate.

Competing Interests

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

Acknowledgments

This work is supported by "The Fundamental Research Funds for the Central Universities (FRF-TP-14-060A2)" of China—"The Research on Parallel Data Processing Method for Materials High-Throughput Experiment."

References

[1] N. Canter, "Materials corrosion preventives: protect materials and specific applications," *Tribology & Lubrication Technology*, vol. 718, pp. 5–6, 2015.

[2] P. Dydio, P.-A. R. Breuil, and J. N. H. Reek, "Dynamic combinatorial chemistry in chemical catalysis," *Israel Journal of Chemistry*, vol. 53, no. 1-2, pp. 61–74, 2013.

[3] P. de Lima-Neto, A. N. Correia, R. A. C. Santana et al., "Morphological, structural, microhardness and electrochemical characterisations of electrodeposited Cr and Ni-W coatings," *Electrochimica Acta*, vol. 55, no. 6, pp. 2078–2086, 2010.

[4] L. Chen, L.-Y. Wang, S.-J. Liu et al., "Profiling of microbial community during in situ remediation of volatile sulfide compounds in river sediment with nitrate by high-throughput sequencing," *International Biodeterioration & Biodegradation*, vol. 85, pp. 429–437, 2013.

[5] L. M. Junker and J. Clardy, "High-throughput screens for small-molecule inhibitors of *Pseudomonas aeruginosa* biofilm development," *Antimicrobial Agents and Chemotherapy*, vol. 51, no. 10, pp. 3582–3590, 2007.

[6] G. R. Eldridge, H. C. Vervoort, C. M. Lee et al., "High-throughput method for the production and analysis of large natural product libraries for drug discovery," *Analytical Chemistry*, vol. 74, no. 16, pp. 3963–3971, 2002.

[7] P. A. White, A. E. Hughes, S. A. Furman et al., "High-throughput channel arrays for inhibitor testing: proof of concept for AA2024-T3," *Corrosion Science*, vol. 51, no. 10, pp. 2279–2290, 2009.

[8] D. Itzhak, I. Dinstein, and T. Zilberberg, "Pitting corrosion evaluation by computer image processing," *Corrosion Science*, vol. 21, no. 1, pp. 17–22, 1981.

[9] E. N. Codaroa, R. Z. Nakazatoa, A. L. Horovistizb, L. M. F. Ribeirob, R. B. Ribeirob, and L. R. O. Heinb, "An image processing method for morphology characterization and pitting corrosion evaluation," *Materials Science and Engineering A*, vol. 334, no. 1-2, pp. 298–306, 2002.

[10] S. Wang, D. Kong, and S. Song, "Diagnosing corrosion modality system of metallic material in seawater based on fuzzy pattern recognition," *Acta Metallrugica Sinica*, vol. 37, no. 5, pp. 517–521, 2001.

[11] S. Xu, Y. Weng, and X. Li, "Characterization for corrosion pit distribution by using fractal dimension of image," *Journal of the Chinese Society of Corrosion and Protection*, vol. 27, no. 2, pp. 109–113, 2007.

[12] K. Belaid, O. Echi, and R. Gargouri, "Two classes of locally compact sober spaces," *International Journal of Mathematics and Mathematical Sciences*, vol. 2005, no. 15, pp. 2421–2427, 2005.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 5635673, 9 pages
http://dx.doi.org/10.1155/2016/5635673

Research Article

RVLBPNN: A Workload Forecasting Model for Smart Cloud Computing

Yao Lu,[1] John Panneerselvam,[2] Lu Liu,[1,2] and Yan Wu[1,3]

[1]*School of Computer Science and Telecommunication Engineering Jiangsu University, Jiangsu, China*
[2]*Department of Computing and Mathematics, University of Derby, Derby, UK*
[3]*Department of Computer Science, Boise State University, Boise, USA*

Correspondence should be addressed to Lu Liu; l.liu@derby.ac.uk

Received 28 July 2016; Accepted 19 September 2016

Academic Editor: Wenbing Zhao

Given the increasing deployments of Cloud datacentres and the excessive usage of server resources, their associated energy and environmental implications are also increasing at an alarming rate. Cloud service providers are under immense pressure to significantly reduce both such implications for promoting green computing. Maintaining the desired level of Quality of Service (QoS) without violating the Service Level Agreement (SLA), whilst attempting to reduce the usage of the datacentre resources is an obvious challenge for the Cloud service providers. Scaling the level of active server resources in accordance with the predicted incoming workloads is one possible way of reducing the undesirable energy consumption of the active resources without affecting the performance quality. To this end, this paper analyzes the dynamic characteristics of the Cloud workloads and defines a hierarchy for the latency sensitivity levels of the Cloud workloads. Further, a novel workload prediction model for energy efficient Cloud Computing is proposed, named RVLBPNN (Rand Variable Learning Rate Backpropagation Neural Network) based on BPNN (Backpropagation Neural Network) algorithm. Experiments evaluating the prediction accuracy of the proposed prediction model demonstrate that RVLBPNN achieves an improved prediction accuracy compared to the HMM and Naïve Bayes Classifier models by a considerable margin.

1. Introduction

Cloud Computing is emerging as a prominent computing paradigm for various business needs, as it is known to be a low-cost any-time computing solution. The on-demand service access features of the Cloud Computing help the Cloud clients to adopt or transform their business model to Cloud datacentres for computing and storage resources [1]. This increasing number of Cloud adoptions by various business domains over the recent years is also reflected in the increase in the number of Cloud service providers. An immediate impact of this is that Cloud datacentres are addressed to be one of the major sources of energy consumers [2] and environmental pollutants. To this end, Cloud datacentres are addressed to be causing energy, economic, and environmental impacts to an irresistible margin. It has been reported [3] that ICT (Information Communication Technology) energy consumption will contribute up to 50% of the total energy expenditures in the United States in the next decade, which was just 8% in the last decade. Energy efficient computing has been promoted and researched under various dimensions for the purpose of reducing the energy consumption levels of the datacentre whilst processing workloads and cooling the server resources. It is worthy of note that cooling system in a typical Cloud datacentre would incur considerable amount of energy cost of those spent towards the actual task execution [4]. Thus it is apparent that energy efficient Cloud Computing is one of demanding characteristics of Cloud Computing.

Resource management driven by forecasting the future workloads is one of the possible ways of achieving energy efficiency in Cloud Computing. In general, the intrinsic dynamic [5] nature of the Cloud workloads imposes complexities in scheduling, resource allocation, and executing workloads in the datacentres. Predicting the nature of the future workloads

can help reduce the energy consumption levels of the server resources by the way of effectively scheduling the incoming workloads with the most appropriate level of resource allocation. Alongside energy efficient computing, predictive analytics in Cloud Computing also benefits [5] effective resource utilization, optimum scalability of resources, avoiding process failures, capacity planning, network allocation, task scheduling, load balancing, performance optimization and maintaining the predetermined QoS (Quality of Service) and SLAs (Service Level Agreements), and so forth.

Owing to the extravagant dynamicity of Cloud workloads, understanding the characteristic behaviors of the Cloud workloads at the datacentre environment is often a complex process. Mostly, Cloud workloads are of shorter duration and arrive more frequently at the datacentres and are generally not computationally more intensive unlike scientific workloads. Furthermore, every submitted workloads are bound to a certain level of latency sensitiveness [6] which decides the time within which the workload has to be processed. Workloads with increased latency sensitivity levels usually demand quicker scheduling from the providers. This implies that an effective prediction model should possess the qualities of understanding the inherent characteristics and nature of the Cloud workloads and their corresponding behaviors at the datacentres.

Despite the existing and ongoing researches, Cloud Computing still demands extensive analyses of the Cloud entities for the purpose of modelling the relationship between the users and their workload submissions and the associated resource requirements. An effective prediction model should necessarily incorporate the knowledge of three important characteristic events in a datacentre environment in order to achieve reliable level of prediction accuracy. Firstly, the volume and the nature of the workloads submitted are driven by the users based on their requirements and resource demands. Increased amounts of jobs submissions obviously demand increased amounts of resource allocation and thus causes increased energy expenditures. Secondly, the actual execution of the workloads would not necessarily consume all the allocated resources. The immediate implication is that increased proportions of allocated resources remain idle during task execution and incur undesirable energy consumptions. Finally, the user behavioral pattern of job submission and associated resource consumption are subjected to change over time.

The intrinsic dynamism of both the Cloud workloads and the server resources should be effectively captured [7] by the prediction model over a prolonged observation period. Existing works in analyzing the intrinsic characteristics of the Cloud entities have not contributed suffice inferences [5, 8, 9] required for an effective prediction model. Imprecise knowledge of such aforementioned parameters of the Cloud entities would increase the prediction error margin, which would directly affect the Quality of Service (QoS) by violating the Service Level Agreement (SLA). With this in mind, this paper proposes a novel forecast model named RVLBPNN (Rand Variable Learning Rate Backpropagation Neural Network), based on an improved BPNN (BP Neural Network) for accurately predicting the user requests. Exploiting the latency sensitivity levels of the Cloud workloads, our proposed model predicts the user requests anticipated in the near future in a large-scale datacentre environment.

The rest of this paper is organized as follows: Section 2 introduces the previous works in Cloud workload prediction modelling. Section 3 presents a background study on Cloud workloads, exhibiting the dynamic nature of the Cloud workloads. The computational latencies affecting the Cloud workloads are defined in Section 4 and Section 5 proposes our prediction model based on the modified BP Neural Network. Our experiments are presented in Section 6 and Section 7 concludes this paper along with our future research directions.

2. Related Works

A number of researches are being conducted with the motivation of promoting green computing in the recent past. For instance, the approach of capacity management and VM (Virtual Machine) placement have been the strategies of [10, 11]. A workload placement scheme, called BADP, combines task's behavior to place data for improving locality at the cache line level. Further, [11] proposes a remaining utilization-aware (RUA) algorithm for VM placement. In general, workload placement and task allocation can be more effective when driven by a proactive prediction of the incoming workloads. Time series [12] approach incorporates the repeatable behaviors such as periodicity and timely effects of the various Cloud entities such as VMs and users and explores the temporal and spatial correlations in order to provide the prediction results. However, such technique usually explores the entities individually and often leads to inaccurate results resulting from the random behaviors of the individual entities.

A multiple time series approach [13] has been proposed to improve the prediction accuracy, by the way of analyzing the Cloud entities at the group level rather individually. Nonlinear time series approach works with the assumption that the observations are real valued and such techniques often require special emphasis on extracting the chaotic invariants for prediction analysis. Autoregression (AR) is a prediction technique [14] which usually predicts the next state transition by recursively acting on the prediction values. However, AR method has a conspicuous shortcoming that the prediction errors will be accumulated for long term prediction analysis because of the recursive effect. Another drawback of AR methods is that they only deliver accurate forecasts for datasets characterized with reasonable periodicity, which is shown in [15], where a number of different linear prediction models based on AR have been deeply analyzed. Poisson process [16] models the incoming workload arrival pattern for prediction analysis and has the capability of capturing complex nonexponential time varying workload features. Moving average approaches [14, 16] such as first-order and second-order moving average techniques used for prediction analysis cannot capture important features required to adapt to the load dynamics.

Recently, Bayes and Hidden Markov Modelling (HMM) [5] approaches were analyzed in our earlier works for evaluating their prediction efficiency in Cloud environments.

Byes technology predicts the future samples based on a predefined evidence window. The adjacent samples contained in the evidence window should be mutually correlated for delivering a reliable prediction output. Thus Bayes model will lose efficiency in a dynamic Cloud environment. However, Bayes model could still be deployed in situations where there are less fluctuations among the workload behavior. HMM is a probabilistic approach which is used to predict the future sate transition from the current state. In spite of the dynamic nature of the Cloud workloads, probabilistic approach may not scale well for predicting the future workloads with a reliable level of prediction accuracy. Despite the existing works, Cloud Computing still demands a smart prediction model with the qualities of relative high precision and the capacity of delivering a reliable level of prediction accuracy. With this in mind, this paper proposes a novel prediction model named RVLBPNN. Exploiting the workload characteristics, our proposed model achieves a reliable level of prediction accuracy. Our proposed model has been sampled and tested for accuracy based on a real life Cloud workload behaviors.

3. Background

3.1. Cloud Workloads. Cloud workloads arrive at Cloud datacentres in the form of jobs [15] submitted by the users. Every job includes certain self-defining attributes such as the submission time, user identity, and its corresponding resource requirements in terms of CPU and memory. A single job may contain one or more tasks, which are scheduled for processing at the Cloud servers. A single task may have one or more process requirements. Tasks belonging to a single job may also be scheduled to different machines but it is desirable to run multiple processes of a single task in a single machine. Tasks are also bound to have varied service requirements such as throughput, latency, and jitter, though they belong to the same job. The tasks belonging to the same job not necessarily exhibit higher correlating properties among them. Thus, tasks within the same job might exhibit greater variation in their resource requirements. Tasks might also interact among each other during their execution. Furthermore, two jobs with the same resource requirements may not be similar in their actual resource utilization levels because of the variation found among the tasks contained within the jobs. Based on the resource requirements, tasks are scheduled either within the same or across different servers. Usually, the provider records the resource utilization levels of every scheduled task and maintains the user profiles.

The attributes encompassed by the Cloud workloads, such as type, resource requirements, security requirements, hardware, and network constraints, can be exploited to derive the behaviors the Cloud workloads. Interestingly, Cloud workloads behave distinctively with different server architectures. Such distinctive workload behaviors with different server architectures strongly influence the CPU utilization, with the memory utilization generally remaining stable across most of the server architectures. Thus the resource utilization highly varies across the CPU cores compared to the memory or disc, as the disc utilization mostly shows similar utilization patterns across different server architectures. Thus the behaviors of workloads at the Cloud processing environment are strongly correlated with the CPU cores compared to RAM capacity of the machines at the server level. The capacity levels of CPU and memory in a physical server usually remain static. Resource utilization levels are more dynamic and vary abruptly under different workloads. Such dynamic parameters of the server architectures are usually calculated as the measure of the number of cycles per instruction for CPU and memory access per instruction for memory utilizations, respectively. Thus the task resource usage is usually expressed as a multidimensional representation [5] encompassing task duration in seconds, CPU usage in cores, and memory usage in gigabytes. It is commonly witnessed that most of the allocated CPU and memory resource are left unutilized during task execution.

3.2. Characterizing Workloads. User demand often changes over time which reflects the timely variations of the resource consumption levels of the workloads generated as they are driven by the users. User demands are generally influenced by the time-of-the-day effects, showing a repeating pattern [13] in accordance to the changing business behaviors of the day and by the popular weekend effects showing weekend declines and weekday increase trend in the arrival of the workloads. The relationships [17] between the workloads and user behaviors are primarily the integral component in the understanding of the Cloud-based workloads and their associated energy consumptions.

Different workloads will have different processing requirements such as CPU, memory, throughput, and execution time, and this variation results from the characteristic behaviors of different users. Nowadays, Cloud environments are more heterogeneous composing different servers with different processing capacities. In order to satisfy the diverse operational requirements of the Cloud user demands, normalization of this machine heterogeneity is now becoming an integral requirement of the Cloud providers, by which virtually homogenizing the heterogeneous server architectures and thereby eliminating the differentiation found in both the hardware and the software resources. In general, the different forms of workloads from the provider's perspectives include computation intensive with larger processing and smaller storage, memory intensive with larger storage and smaller processing, workloads requiring both larger processing and larger storage, and communication intensive with more bandwidth requirements. Workloads are usually measured in terms of the user demands, computational load on the servers, bandwidth consumption (communication jobs), and the amount of storage data (memory jobs). User demand prediction modelling requires an in depth quantitative and qualitative analysis of the statistical properties and behavioral characteristics of the workloads including job length, job submission frequency, and resource consumption of the jobs, which insists that the initial characterization of the workloads is more crucial in developing an efficient prediction model. Rather than the stand-alone analysis of the above stated workload metrics, modelling the relationships between them across a set of workloads is more significant in order to achieve more reliable prediction results. Statistical

properties [18] of the workloads are more significant for the prediction accuracy since they remain consistent in longer time frames. Some of the important characteristics of Cloud workloads affecting prediction accuracy include job length, job submission frequency, resource request levels, job resource utilization, and self-similarity.

3.3. Categorizing Workloads. In the Cloud Computing service concept, the workload pattern, the Cloud deployment types (public, private, hybrid, and community), and the Cloud service offering models (SaaS, PaaS, and IaaS) are closely interconnected with each other. From the perspectives of the Cloud service providers, the incoming Cloud workloads can be categorized into five major types [9] as static, periodic, unpredictable, continuously changing, and once-in-a-lifetime workloads. Static and periodic workloads usually follow a predictable pattern in their arrival frequency. Continuously changing workloads exhibit a pattern of definite variations characterized by regular increasing and declining trend in their arrival frequencies. Unpredictable workloads exhibit a random behavior in their arrival frequency and are the most challenging type of workloads for prediction analysis. Once-in-a-lifetime workloads are the rarely arriving workloads and their submissions are mostly notified by the clients.

4. Workload Latency

Latency plays an important role at various levels of processing the workloads in a Cloud processing infrastructure. This paper mainly focuses the influence of the workload latency sensitivity upon prediction accuracy. The most dominating types of latencies are the network latency and the dispatching latency, both of which actually result from the geographical distribution of the users and the Cloud datacentres. Both of these latencies depend on the Round Trip Time (RTT) [19], which defines the time interval between the user requests and the arrival of the corresponding response. Another type of latency existing in the process architecture is the computational latency which is the intracloud latency [20] found among the processing VMs located within a single datacentre. This latency depends on both the software and the hardware components [6, 21] such as CPU architecture, runtime environment, and memory, guests and host operating system, instruction set, and hypervisor used. CPU architecture, Operating System, and the scheduling mechanisms are the most dominating factors of this type of in-house computing latency, and efficient handling of such resources helps reducing the impacts of the computational latencies.

Jobs submitted at the Cloud datacentre undergo various levels of latencies depending on the nature of their process requirements and the end-user QoS expectations. Since a single job might contain a number of tasks, the latency sensitivity of every single tasks has to be treated uniquely. A single definition of the computing latency cannot fit all types of jobs or tasks, since every job is uniquely viewed at the datacentre. For instance, processing a massive scientific workload may span across several days or months, in which latencies of a few seconds are usually acceptable. Common example of the latency sensitive workloads is the World Wide Web, among which different applications have different latency levels. The acceptable level of latencies is usually the measure of the end-user tolerances. Workloads resulting from users surfing the internet are generally latency-insensitive. Jobs including online gaming and stock exchange data are the commonly witnessed latency sensitive applications. The level of sensitivity is determined by the allowed time-scale for the providers to provide an undisrupted execution of the workloads for delivering the desired levels of QoS, ranging from a few microseconds to a few tens of microsecond end-to-end latencies.

The taxonomy of the latency levels of the Cloud workloads studied in this paper are attributed from level 0 representing the least latency sensitive tasks to level 3 representing the most latency sensitive tasks. Least latency sensitive tasks (level 0) are nonproduction tasks [22] such as development and nonbusiness critical analyses, which do not have a significant impact on the QoS even if these jobs are queued at the back end servers. Level 1 tasks are the next level of latency sensitive tasks and are generally the machine interactive workloads. Level 2 tasks are the real time machine interactive workloads and the latency tolerance levels of these tasks stay at tens of milliseconds. Level 3 tasks are the most latency sensitive tasks with latency tolerance levels at the range of submilliseconds, and are generally the revenue generating user requests such as stock and financial analysis. Workloads characterizing an increased level of latency sensitivity are usually treated with higher scheduling priorities at the datacentres. Latency analysis has a prime importance in greening the datacentre, since every job or task submitted to the Cloud has its own level of latency tolerances, directly affecting not only the various workload behaviors at the datacentres but also the end-user QoS satisfaction.

Based on our earlier analysis conducted on a Cloud dataset [23], we perform a latency aware quantification of the jobs submitted to the datacentre comprising a total of 46093201 tasks in our recent work [24]. Figure 1 illustrates a day-wise submission of tasks across the observed 28 days. Figure 2 quantifies the total number of task submissions in terms of their latency sensitivity levels. It can be observed that most of the task submissions are least latency sensitive accounting for 79.52% of the total task submissions, followed by level 1, level 2, and level 3 with 12.46%, 7.54%, and 0.47%, respectively.

5. Proposed Prediction Model

This section describes our novel prediction model aimed at predicting the anticipated workloads in a large-scale datacentre environment.

5.1. BP Neural Network Method. BP Neural Network is a multilayer hierarchical network composed of upper neurons and fully associated lower neurons. Upon training the input samples into this multilayer network structure, the transformed input values are propagated from the input layer through the middle layer, and the values are outputted by the neurons in the output layer. The error margins

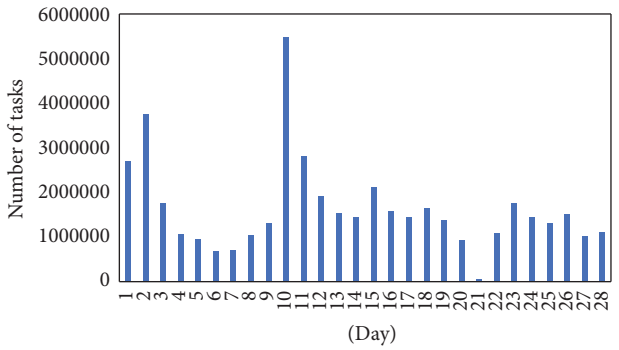


Figure 1: Total number of task submissions.

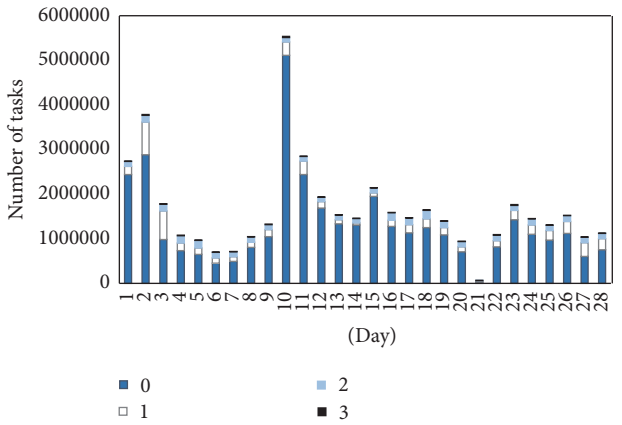


Figure 2: Latency-wise task submission.

between the actual and the expected output are normalized by the way of the output neurons adjusting the connection weights of the neurons in both the middle layer and the input layer. This back propagation mechanism of connection weight adjustment enhances the correctness of the network responses of the neurons to the input values. As the BP algorithm implements a middle hidden layer with associated learning rules, the network neurons can effectively identify the hidden nonlinear pattern among the input samples.

5.2. BP Neural Network Architecture. A typical neuron model can be derived according to characteristics of the neurons [25–28], which is shown in Figure 3. In this figure, $X_1, X_2, \ldots, X_n$ are n input data to the neurons; $a_{i1}, a_{i2}, \ldots, a_{in}$ are the weight factor of $X_1, X_2, \ldots, X_n$, respectively; $g()$ is a nonlinear function; O_i is the output result; and λ_i is the threshold.

Based on the above neuron structure, we make $O_i = g(P_i)$, where, $P_i = \sum_{j=1}^{n} a_{ij}X_j - \lambda_i$. In the formula, X represents the input vector, a_i represents the connection weight vector for neuron i, and P_i is the input of the neurons. In most cases, λ_i is considered to be the 0th input of the neuron. Thus, we can get a simplified equation of the above expression, which is shown in formula (1). In this equation, value $X_0 = -1$, and $a_{i0} = \lambda_i$.

$$P_i = \sum_{j=0}^{n} a_{ij}X_j. \quad (1)$$

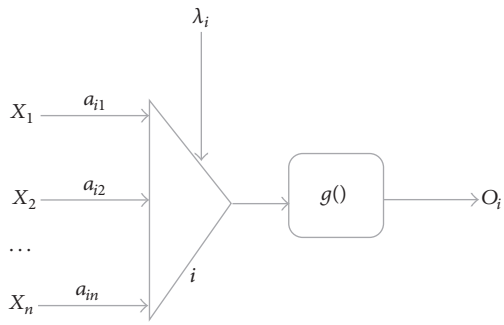


Figure 3: Neuron model.

5.3. An Improved BP Neural Network Algorithm for Prediction. BPNN can effectively extract the hidden nonlinear relationships among the Cloud workloads. However, BPNN with a fixed learning rate cannot extract this nonlinear relationships among the samples of large datasets, since BPNN has a slow convergence rate for large-scale datasets in the range of Big Data. A modified BP algorithm named VLBP (variable learning rate backpropagation) has been proposed to enhance this convergence rate [29]. In comparison with the BP algorithm, VLBP has a characteristic enhancement in both the computation speed and precision of the output. But the VLBP algorithm can be susceptible to several numbers of local minima resulting from the irregular shake surface error. This slows down the update process of Mean Square Error (MSE) and increases the presence of local minimum points. This results in a higher approximation precision despite the improvement in the convergence rate. VLBP exhibits a fluctuating and slower learning process and increases the length of the computation.

This necessitates further improvements in the BP Neural Networks for the purpose of enhancing its prediction efficiency whilst training large datasets. This paper proposes a novel prediction method using a modified BP algorithm by incorporating variant conceptions of a genetic algorithm. Our proposed prediction method effectively adjusts the learning rate of the neurons to a certain probability in accordance with the trend of the MSE during the execution of the VLBP algorithm. The learning rate may not be changed or multiplied by the factor ρ greater than 1 when MSE increases beyond the set threshold ζ.

Our proposed prediction algorithm is described as follows:

(1) Generate a random number rand(u) ($0 <$ rand(u) < 1).

(2) If rand(u) is less than a defined value (set as 0.8 in experiments), then execute VLBP algorithm.

(3) If rand(u) is greater than the defined value, else if MSE has increased, then the learning rate is multiplied by a factor greater than 1 despite MSE exceeding ζ or not; if MSE decreases after updating the connection weight, then the learning rate is multiplied by a factor between 0 and 1.

We named our proposed algorithm as RVLBPNN (Rand Variable Learning Rate Backpropagation Neural Network).

Through this method, the learning rate of the neurons will not be decreased at any time resulting from the slow renewal of MSE near the local minimum point. But, there is also a certain probability of increasing the learning rate of the neurons. RVLBP algorithm can identify the global minimum point by effectively avoiding the local minimum points. Thus our proposed algorithm reduces the presence of local minimum points during the learning process, thereby improving the learning efficiencies of the network neurons.

6. Performance Evaluation

6.1. Experiment Sample. This section demonstrates the efficiency of our proposed prediction model based on RVBLPNN. We train the input data sample to predict the anticipated values in the near future representing the future workloads expected to arrive at the datacentre. The experiment samples are trained in MATLAB 7.14 and the test datasets used are the publically available Google workload traces [23]. The datasets are a collection of 28 days of Google usage data workloads consisting of 46093201 tasks comprising CPU intensive workloads, memory-intensive workloads, and both CPU and memory-intensive workloads. The dataset parameters include time, job id, parent id, number of cores (CPU workloads), and memory tasks (memory workloads), respectively, to define the sample attributes. We compare the prediction efficiencies of our proposed prediction model against the efficiencies of Hidden Markov Model (HMM) and Naïve Bayes Classifier (NBC); both of them were evaluated in our earlier works [5]. All the three models are evaluated for their efficiencies in predicting memory and CPU intensive workloads accordingly. We train the prediction model with a set of 10 samples and contrast the prediction output with the actual set of successive 10 samples.

MATLAB simulation environment provides a built-in model for RVLBPNN technique, modelling RVLBPNN as a supervised learning. The Neural Network is comprised of a three-layer network structure. This three-layer Neural Network can approximate any type of nonlinear continuous function in theory. We ultimately use 10 input nodes, 12 hidden nodes, and 10 output nodes through a number of iterations for enhancing the prediction accuracy. The data samples are normalized and imploded in the interval (0, 1). "logsig" function is selected as the activation function of input layer, hidden layer, and the output layer, so that the algorithm exhibits a good convergence rate. Further, variable learning rate and random variable learning rate are adopted, respectively. This experiment uses 100,000 workload data samples as the training data and another 100000 data samples as the reference data. The prediction accuracy is the measure of correlations between the predicted and actual set of sample values.

6.2. Result Analysis and Performance Evaluation

6.2.1. Memory Workloads Estimation. Figure 4 depicts the estimation results of RVLBPNN, HMM, and NBC model, respectively, in terms of their prediction accuracy whilst predicting the memory-intensive workloads. The number of test

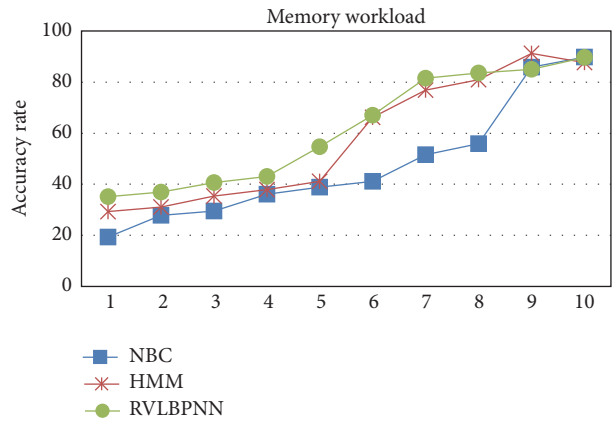


Figure 4: Prediction of memory-intensive workloads.

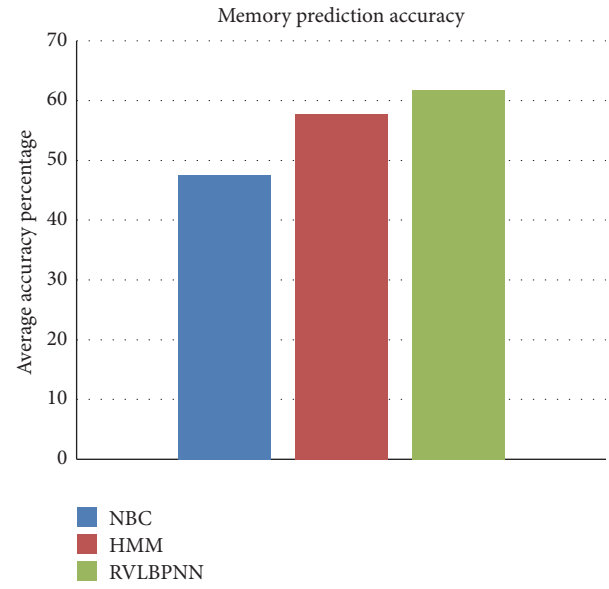


Figure 5: Prediction accuracy for memory-intensive workloads.

samples (x-axis) are plotted against the prediction accuracy (y-axis) for the three models; every set of sample consists of 10000 workload samples. For presenting the test results with a better interpretation, the sample results are sorted ascendingly from 1 to 10 based on the prediction results. The average accuracy percentage in estimating the memory-intensive workloads without considering the latency levels of individual workloads for NBC, HMM, and RVLBPNN are 47.69%, 57.77%, and 61.71%, respectively, as shown in Figures 4 and 5. It is evident from Figures 4 and 5 that the RVLBPNN exhibits a better prediction accuracy than both HMM and NBC techniques. It can be depicted from the estimation results that our proposed RVLBPNN model is demonstrating a minimum of 3% prediction accuracy better than HMM and 13% better than NBC, respectively.

This improved prediction accuracy of the RVLBPNN model is attributed to its ability of capturing the intrinsic relationship features among the arriving Cloud workloads. We further evaluate the efficiency of our proposed model in

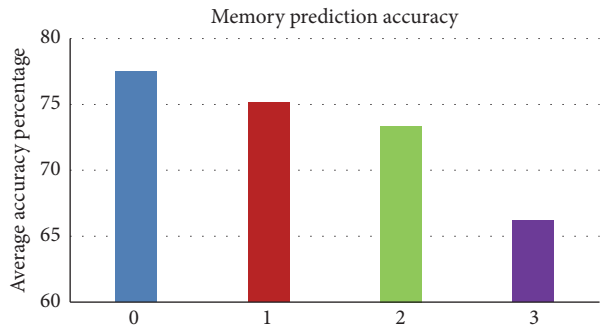


Figure 6: Latency-wise prediction accuracy for memory workloads.

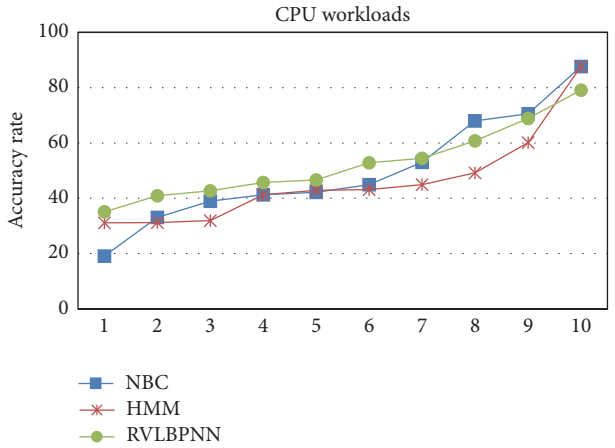


Figure 7: Prediction of CPU intensive workloads.

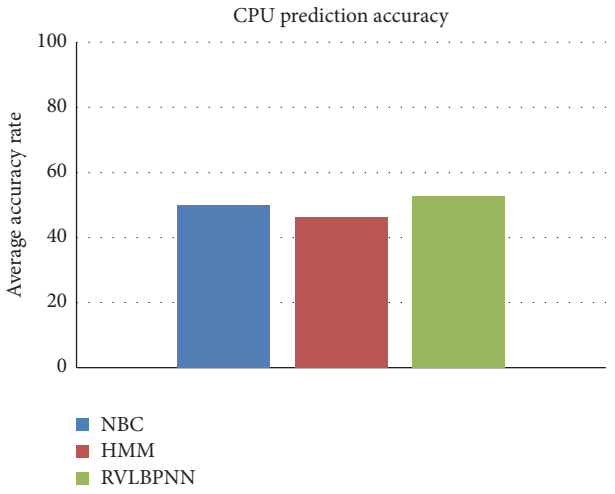


Figure 8: Prediction accuracy for CPU intensive workloads.

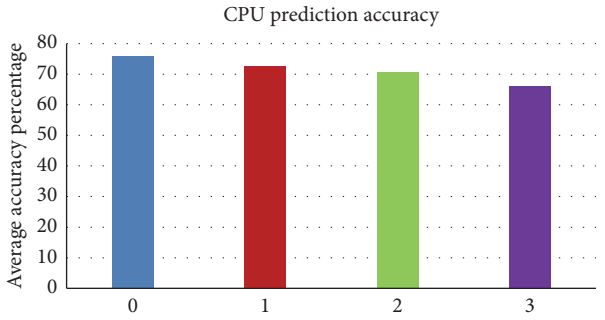


Figure 9: Latency-wise prediction accuracy for CPU workloads.

forecasting memory-intensive workloads of different latency sensitivity levels. Figure 6 depicts the estimation results of our proposed RVLBPNN model in terms of their prediction accuracy whilst predicting memory-intensive workloads of different latency sensitivity levels as described earlier in Section 4. It can be observed from Figure 6 that less latency sensitive memory workloads exhibit better predictability, with the prediction accuracy being 66.17% for level 3 workloads and 77.48% for level 0 workloads, respectively.

6.2.2. CPU Workloads Prediction. Similar to the memory-intensive workloads, the experiments are repeated for the CPU intensive workloads from the dataset. Figure 7 depicts the estimation results of RVLBPNN, HMM, and NBC whilst predicting the CPU intensive workloads. The average prediction accuracy of NBC, HMM, and RVLBPNN models is 49.87%, 46.36%, and 52.70%, respectively, whilst predicting CPU intensive workloads, as shown in Figure 8. It can be observed that RVLBPNN exhibits better prediction accuracy than both HMM and NBC models by a margin of around 3% and 6%, respectively.

We further evaluated the efficiency of our proposed prediction model in predicting the CPU intensive workloads of different latency levels. Figure 9 depicts the estimation results of our proposed RVLBPNN model whilst predicting the CPU intensive workloads of different latency sensitivity levels. We observe a similar trend of prediction accuracy between both memory and CPU workloads of different latency sensitivity levels. Again CPU intensive workloads of less latency levels are exhibiting better predictability, with the accuracy being 66.08% for level 3 workloads and 75.90% for level 0 workloads. This leads us to infer that least latency sensitivity level workloads exhibit a better rate of prediction accuracy for both CPU and memory-intensive workloads.

6.2.3. Interpretation and Discussion. From the experiment results, it is clearly evident that our proposed RVLBPNN model demonstrates better prediction accuracy than both HMM and NBC models by a considerable margin. Our proposed model outperforms the other two models whilst predicting both the CPU intensive and memory-intensive workloads. Meanwhile, we also observe that increasing levels of latency sensitivity of both CPU and memory-intensive workloads impose increasing error margin in the prediction results. Lower level of latency sensitivity exhibits better predictability. Since the majority of the Cloud workloads are of lower latency sensitivity levels, our proposed prediction model can accurately predict the trend of most of the arriving workloads. An increased level of intrinsic similarity among the arriving workloads facilitates a better learning rate of the neurons in the RVLBPNN model, which results in an increased prediction accuracy. From the experiments, we postulate that workloads should be treaded uniquely with

respect to their computational demand latency sensitivity and user requirements for achieving a reliable level of prediction accuracy. Furthermore, workload prediction analytics can be benefitted with better accuracy when the workloads are analyzed at the task level rather than at the job level.

7. Conclusion

Green computing has turned out to be one of the important characteristics for achieving sustainable smart world in the future. Resource management by the way of predicting the expected workloads facilitates optimum scaling of the server resources, reducing the presence of idle resources and allocating appropriate levels of server resources to execute the user requests. The reliability and accuracy levels of such prediction techniques directly impacts important decision making in large-scale Cloud datacentre environments. In this paper, we propose a novel workload prediction model for the purpose of predicting the future workloads in Cloud datacentres. Our proposed novel workload prediction model, called RVLBPNN, is based on BP Neural Network algorithm and predicts the future workloads by the way of exploiting the intrinsic relationships among the arriving workloads. The experimental results indicate that the proposed RVLBPNN model achieves better precision and efficiency than the HMM-based and NBC-based prediction techniques. As a future work, we plan to explore the possibilities of further improving the prediction accuracy of our proposed approach. For instance, incorporating the periodicity effects of the workload behavior into RVLBPNN can further enhance the prediction accuracy. Meanwhile, investigating the efficiencies of our novel prediction method in predicting the anticipated workloads in similar distributed environments will be one of our future research directions.

Competing Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grants nos. 61502209 and 61502207 and the Natural Science Foundation of Jiangsu Province under Grant no. BK20130528.

References

[1] H. Al-Aqrabi, L. Liu, R. Hill, and N. Antonopoulos, "Cloud BI: future of business intelligence in the cloud," *Journal of Computer and System Sciences*, vol. 81, no. 1, pp. 85–96, 2015.

[2] T. V. T. Duy, Y. Sato, and Y. Inoguchi, "Performance evaluation of a green scheduling algorithm for energy savings in cloud computing," in *Proceedings of the IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW '10)*, pp. 1–8, Atlanta, Ga, USA, April 2010.

[3] L. Ceuppens, A. Sardella, and D. Kharitonov, "Power saving strategies and technologies in network equipment opportunities and challenges, risk and rewards," in *Proceedings of the International Symposium on Applications and the Internet (SAINT '08)*, pp. 381–384, August 2008.

[4] J. Li, B. Li, T. Wo et al., "CyberGuarder: a virtualization security assurance architecture for green cloud computing," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 379–390, 2012.

[5] J. Panneerselvam, L. Liu, N. Antonopoulos, and Y. Bo, "Workload analysis for the scope of user demand prediction model evaluations in cloud environments," in *Proceedings of the 7th IEEE/ACM International Conference on Utility and Cloud Computing (UCC '14)*, pp. 883–889, December 2014.

[6] Z. Wan, "Sub-millisecond level latency sensitive cloud computing infrastructure," in *Proceedings of the 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT '10)*, pp. 1194–1197, Moscow, Russia, October 2010.

[7] H. Zhang, G. Jiang, K. Yoshihira, H. Chen, and A. Saxena, "Intelligent workload factoring for a hybrid cloud computing model," in *Proceedings of the Congress on Services—I (SERVICES '09)*, pp. 701–708, 2009.

[8] C. Glasner and J. Volkert, "Adaps—a three-phase adaptive prediction system for the run-time of jobs based on user behaviour," *Journal of Computer and System Sciences*, vol. 77, no. 2, pp. 244–261, 2011.

[9] C. A. L. Fehling, Frank, Retter et al., "CloudComputingPatterns2014," 2014.

[10] J. Wang, G. Jia, A. Li, G. Han, and L. Shu, "Behavior aware data placement for improving cache line level locality in cloud computing," *Journal of Internet Technology*, vol. 16, no. 4, pp. 705–716, 2015.

[11] G. Han, W. Que, G. Jia, and L. Shu, "An efficient virtual machine consolidation scheme for multimedia cloud computing," *Sensors*, vol. 16, no. 2, article 246, 2016.

[12] S. Mahambre, P. Kulkarni, U. Bellur, G. Chafle, and D. Deshpande, "Workload characterization for capacity planning and performance management in IaaS cloud," in *Proceedings of the 1st IEEE International Conference on Cloud Computing for Emerging Markets (CCEM '12)*, pp. 1–7, Bangalore, India, October 2012.

[13] A. Khan, X. Yan, S. Tao, and N. Anerousis, "Workload characterization and prediction in the cloud: a multiple time series approach," in *Proceedings of the IEEE Network Operations and Management Symposium (NOMS '12)*, pp. 1287–1294, IEEE, Maui, Hawaii, USA, April 2012.

[14] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards characterizing cloud backend workloads: insights from Google compute clusters," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 34–41, 2010.

[15] P. A. Dinda and D. R. O'Hallaron, "Host load prediction using linear models," *Cluster Computing*, vol. 3, no. 4, pp. 265–280, 2000.

[16] S. Di, D. Kondo, and W. Cirne, "Google hostload prediction based on Bayesian model with optimized feature combination," *Journal of Parallel and Distributed Computing*, vol. 74, no. 1, pp. 1820–1832, 2014.

[17] I. S. Moreno, P. Garraghan, P. Townend, and J. Xu, "An approach for characterizing workloads in google cloud to derive realistic resource utilization models," in *Proceedings of the IEEE 7th International Symposium on Service-Oriented System Engineering (SOSE '13)*, pp. 49–60, IEEE, Redwood City, Calif, USA, March 2013.

[18] N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," in

Proceedings of the IEEE 4th International Conference on Cloud Computing (CLOUD '11), pp. 500–507, Washington, DC, USA, July 2011.

[19] Z. Wan, "Cloud Computing infrastructure for latency sensitive applications," in *Proceedings of the IEEE 12th International Conference on Communication Technology (ICCT '10)*, pp. 1399–1402, November 2010.

[20] M. S. Bali and S. Khurana, "Effect of latency on network and end user domains in cloud computing," in *Proceedings of the 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE '13)*, pp. 777–782, Chennai, India, December 2013.

[21] Z. Wan, P. Wang, J. Liu, and W. Tang, "Power-aware cloud computing infrastructure for latency-sensitive internet-of-things services," in *Proceedings of the UKSim 15th International Conference on Computer Modelling and Simulation (UKSim '13)*, pp. 617–621, April 2013.

[22] C. Reiss, J. Wilkes, and J. L. Hellerstein, "Google cluster-usage traces: format + schema," Tech. Rep., Google Inc., Mountain View, Calif, USA, 2011.

[23] Google, "Google Cluster Data V1," 2011, https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md.

[24] J. Panneerselvam, L. Liu, N. Antonopoulos, and M. Trovati, "Latency-aware empirical analysis of the workloads for reducing excess energy consumptions at cloud datacentres," in *Proceedings of the IEEE 11th Symposium on Service-Oriented System Engineering (SOSE '16)*, pp. 62–70, Oxford, UK, March 2016.

[25] Z. Uykan, C. Güzeliş, and H. N. Koivo, "Analysis of input-output clustering for determining centers of RBFN," *IEEE Transactions on Neural Networks*, vol. 11, no. 4, pp. 851–858, 2000.

[26] X. Sun, Z. Yang, and Z. Wang, "The application of BP neutral network optimized by genetic algorithm in transportation data fusion," in *Proceedings of the IEEE 2nd International Conference on Advanced Computer Control (ICACC '10)*, pp. 560–563, Shenyang, China, March 2010.

[27] W. C. Wang, *BP Neural Network and Application in Automobile Engineering*, BeiJing Institute of Technology University, 1998.

[28] Z. Li, Q. Lei, X. Kouying, and Z. Xinyan, "A novel BP neural network model for traffic prediction of next generation network," in *Proceedings of the 5th International Conference on Natural Computation (ICNC '09)*, pp. 32–38, Tianjin, China, August 2009.

[29] M. T. Hagan, H. B. Demuth, and M. Beale, *Neural Network Design*, PWS Publishing, Boston, Mass, USA, 1996.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 7609460, 13 pages
http://dx.doi.org/10.1155/2016/7609460

Research Article

A Dynamic Pricing Reverse Auction-Based Resource Allocation Mechanism in Cloud Workflow Systems

Xuejun Li,[1,2] Ruimiao Ding,[1] Xiao Liu,[3] Xiangjun Liu,[1] Erzhou Zhu,[1] and Yunxiang Zhong[1]

[1]*School of Computer Science and Technology, Anhui University, Hefei, China*
[2]*School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Australia*
[3]*School of Information Technology, Deakin University, Melbourne, Australia*

Correspondence should be addressed to Xiao Liu; xiao.liu@deakin.edu.au

Received 22 July 2016; Accepted 3 October 2016

Academic Editor: Wenbing Zhao

Market-oriented reverse auction is an efficient and cost-effective method for resource allocation in cloud workflow systems since it can dynamically allocate resources depending on the supply-demand relationship of the cloud market. However, during the auction the price of cloud resource is usually fixed, and the current resource allocation mechanisms cannot adapt to the changeable market properly which results in the low efficiency of resource utilization. To address such a problem, a dynamic pricing reverse auction-based resource allocation mechanism is proposed. During the auction, resource providers can change prices according to the trading situation so that our novel mechanism can increase the chances of making a deal and improve efficiency of resource utilization. In addition, resource providers can improve their competitiveness in the market by lowering prices, and thus users can obtain cheaper resources in shorter time which would decrease monetary cost and completion time for workflow execution. Experiments with different situations and problem sizes are conducted for dynamic pricing-based allocation mechanism (DPAM) on resource utilization and the measurement of Time∗Cost (TC). The results show that our DPAM can outperform its representative in resource utilization, monetary cost, and completion time and also obtain the optimal price reduction rates.

1. Introduction

Workflow model is often used to manage complex business applications. A workflow is defined as a collection of tasks which are handled in a specific order [1, 2]. A workflow management system needs to allocate and execute tasks efficiently to meet users' needs. Cloud computing uses a pay-as-you-go model which provides virtually unlimited computational resources at lower costs with better reliability and delivers the resources by means of virtualization technologies [3]. Cloud workflow systems are workflow systems deployed on cloud computing environment to gain unlimited resources including computation, storage, and network [4].

Resource allocation for cloud workflow systems has received much attention. Allocating cloud resources to workflow is an NP-hard problem and needs to consider the overall performance of system especially monetary cost and completion time [5]. Resource allocation mechanism includes conventional methods and market-oriented methods [6, 7]. Conventional methods require global knowledge and complete information. Users pay for the resources based on reserved price. In contrast, market-oriented methods can offer incentives to participants and the methods decide the price based on the values that users can get from the resources [8].

Different from conventional counterparts, market-oriented methods assume that providers and users are rational and intelligent. And resource allocation depends on many factors including supply-demand relationship and resource price. Auction is a powerful tool to allocate resources in the market. Generally speaking, auction is a protocol that allows participants to indicate their interests in different resources and use these indications of interest to determine both resource allocation and price [9]. Reverse auction method is a typical auction. In conventional auction, there are one seller and multiple buyers. As for reverse auction, there are multiple sellers and only one buyer [10]. The user sends the specification of resource requirement to the cloud broker and requests

resources. The cloud broker transfers the specification to all cloud providers. The cloud providers sell resources with proper price and capabilities. And then the users select the optimal resources according to some criteria, for example, Quality of Service [11, 12]. In general, reverse auction is used to prevent the occurrence of trading fraud and achieve dynamic pricing and automatic procurement [11]. In the most relevant literature [12], the authors propose Biobjective Scheduling Strategy (BOSS) based on reverse auction to allocate resources for tasks of workflows and each task starts an auction and gets a resource to minimize the monetary cost and completion time. However, the resource price is fixed during the auction, so that some providers with weaker competitiveness may lose the auction all the time. This leads to low efficiency of resource utilization, which decreases the allocation equilibrium of tasks on resources in the cloud market.

In this paper, we firstly present a dynamic pricing strategy to change resource prices according to the trading situation and then present a DPAM mechanism to improve the efficiency of resource utilization. Changing resource price is a common and efficient way to increase competitiveness [13]. In the dynamic pricing strategy, the resource price changes according to the trading situation. Those providers who have strong competitiveness (i.e., higher chance of winning in historic auctions) in the market often keep resource prices unchanged during an auction. But for those who have weak competitiveness, decreasing the resource price with a certain rate is an effective way to increase the winning chances. Therefore, changing prices can increase the chance of winning an auction and gaining more revenue for providers with weaker competitiveness. In the meantime, the users can gain cheaper resources. In DPAM, many providers with weak competitiveness use dynamic pricing strategy to increase winning chances and revenue, so that resource utilization of the cloud market increases. Meanwhile, cloud workflows will be executed timely with less monetary cost. In our experiment, the measurement of TC is employed to evaluate the performance of our proposed strategy.

In our previous preliminary work, a dynamic pricing strategy in reserve auction was presented to change resource prices according to the trading situation, and then a novel DPAM was proposed to improve the efficiency of resource utilization [14]. Based on this work, we have further investigated the resource allocation based on reverse auction and made the following substantial extensions in this paper:

(i) In problem analysis, a real world stock trading workflow is given and resource allocation processes of BOSS and DPAM are described based on this example. The process illustrates the difference of pricing mechanism between BOSS and DPAM.

(ii) In evaluation, firstly the performance of BOSS and DPAM on resource utilization and the measurement of TC with different problem sizes are tested. Then simulation on DPAM with different price reduction rates evaluates its performance on resource utilization and TC.

Dynamic pricing based allocation mechanism is designed to improve resource utilization and decrease monetary cost and completion time. In summary, this mechanism has three major advantages:

(i) In reverse auction, dynamic pricing strategy is more effective to increase the revenue of providers with weak competitiveness and decrease users' cost than fixed pricing. Providers with weak competitiveness decrease the price to increase the chance of winning auction and gaining more revenue. Simultaneously, users who choose the resource with lower price will have less monetary cost.

(ii) Dynamic pricing based allocation mechanism can improve resource utilization for providers. Providers change resource prices to sell more resources according to dynamic pricing strategy especially for those with weak competitiveness. This brings higher resources utilization because more resources are chosen.

(iii) Dynamic pricing based allocation mechanism can decrease monetary cost and completion time for users. More price-competitive resources will appear in the market because of increasing competition among providers. So users can easily obtain cheaper and more resources and hence decrease monetary cost and completion time.

The rest of the paper is organized as follows. Section 2 describes some related research. In Section 3, we show an example to analyze the problem. Section 4 proposes a dynamic pricing strategy, and Section 5 presents dynamic pricing-based allocation mechanism. In Section 6, traditional mechanism and ours are evaluated. Finally, Section 7 concludes the paper and discusses our future work.

2. Related Work

Resource allocation has become an important task to provide efficient and economical resources in cloud computing environment [15]. Wood et al. [16] propose an approach for dynamic allocation of resources by defining a unique metric based on the consumption of the three resources: CPU, network, and memory. Görlach and Leymann propose a method for dynamic provisioning of services in clouds in order to optimize the distribution of services [17]. However, their proposed approaches are not efficient and economic because the allocation of resource does not consider market situation.

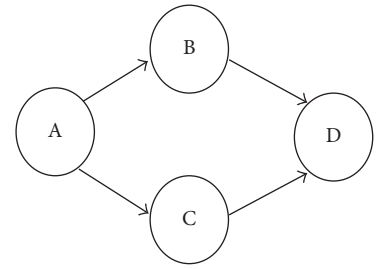
Market-oriented resource allocation has received much attention as it is a significant problem of large-scale distributed systems. In [18], authors present a model for resource allocation in grid using market-oriented concepts including commodity market, posted price modeling, and contract net models bargaining modeling. Mao and Humphrey explore the cloud autoscaling framework for resource allocation. The goal is to ensure that all jobs finish before their respective deadlines while running on these resources which consume the least amount of money [19]. In [20], in order to effectively manage resource allocation and workflow execution, Wang et al. design a mechanism which responds to the user's

continuous workflow requests and schedules their execution. In [21, 22], a lot of heuristic methods are presented to solve the problem in services systems. These heuristic methods consider the optimization algorithm from the aspects of cost, deadline, and reliability which can improve the performance of the algorithm. Ludwig presents a heuristic program for resource allocation on utility computing infrastructure. This heuristic program optimizes the number of resources allocated to tasks of workflow and speeds up the execution within a limitation of budget [23]. In [24], the authors propose a resource allocation approach to better match the resource allocated to the job with the cloud's residual resource.

Auction [25, 26] is a popular method to solve resource allocation problem. In [27, 28], the authors present auction-based mechanisms to determine optimal resource price, taking into account the user's budget and time constraints. Prasad G et al. present a combinatorial auction mechanism to allocate multiple resources in one auction [6]. However, they consider the pricing model of only one seller. Reverse auction is a popular auction in which the roles of buyer and seller are reversed [29]. In an ordinary auction buyers compete to obtain goods or service by offering increasingly higher prices [30]. In the reverse auction [31], the sellers typically decrease prices to compete against each other and obtain business from the buyer. The authors employ reverse auction to select the optimal resource based on available information to maximize their own profits [32]. A cloud resource allocation approach based reverse auction [33] is presented to select suitable cloud resource providers for users.

However, these methods do not focus on the pricing mechanism of the resource allocation. Resource pricing is an important aspect in resource allocation. In [34], authors mention Commodity Market which means that the sellers set the price for merchandise and the buyers pay money to get it. The price is predetermined by the seller and does not change over time based on supply-demand relationship. But fixed pricing is not suitable for the changeable market of cloud resources.

Dynamic pricing has gained wide attention from both industry and academia in the cloud computing. Amazon EC2 [35] has introduced a "spot pricing" scheme where the spot price is set according to resource supply and demand. Because fixed pricing does not reflect the dynamic changes of supply and demand, a dynamic scheme for allocation of multiple-type resources [36] is proposed to increase the percentages of successful buying and selling. In [12], authors introduce a pricing model and a truthful mechanism for scheduling single tasks considering two objectives: completion time and monetary cost based on reverse auction. However, they do not consider the competitiveness among providers which leads to the fact that the losers may always lose auction because they do not try to improve their competitiveness. To solve this problem, we propose dynamic pricing based scheduling mechanism for allocating resources efficiently, in which providers who lose an auction will decrease resource price in order to win so as to gain more revenue.



A: stock issuance C: continuous auction
B: fixed auction D: formation of price chart

Figure 1: Stock trading workflow.

Table 1: Characteristic of tasks.

Task	Workload
A	6
B	4
C	5
D	7

Table 2: Characteristics of resources.

RN	CA	RP	SP
1	1	0.09	0.14
2	1.4	0.14	0.20
3	1.2	0.12	0.17

RN: resource number, CA: computation ability, RP: reserve price, and SP: starting price.

3. Problem Analysis

In this section, a stock trading workflow is given to explain the difference of pricing mechanism between the BOSS mechanism and ours. Stock trading is a typical process in the market. At first, stock exchange starts with stock issuance and then price formation process follows. During this process, fixed auction and continuous auction happen simultaneously. At last, price chart is generated.

As shown in Figure 1, stock trading workflow contains four tasks. The execution sequence of tasks partially depends on relation. The succeeding tasks start only when their predecessor tasks finish. Table 1 indicates the workload of each task. Table 2 shows some characteristics of three resources including resource number, computation ability, reserve price, and starting price. Here, computation ability is the computation speed of CPUs, the reserve price is the lowest price of resource during the auction, and the starting price is the first resource price when auction starts.

3.1. Resource Allocation Process of DPAM. For the BOSS mechanism [12], tasks start an auction according to the specific order to select the resource with the minimum product of completion time and total monetary cost. And providers give their bids $bid_j = (\mathrm{CA}_j, \mathrm{SP}_j)$ which indicate computation ability and starting price of resource j. The workflow

Table 3: Resource allocation process of BOSS.

RN	ST	ET	CT	MC	TC	WR
(a) First auction						
1	0	6.00	6.00	0.84	5.04	
2	**0**	**4.29**	**4.29**	**0.86**	**3.67**	2
3	0	5.00	5.00	0.85	4.25	
(b) Second auction						
1	4.29	4.00	8.29	0.56	4.64	
2	**4.29**	**2.86**	**7.15**	**0.57**	**4.08**	2
3	4.29	3.33	7.62	0.57	4.32	
(c) Third auction						
1	4.29	5.00	9.29	0.70	6.50	
2	7.15	3.57	10.72	0.71	7.66	3
3	**4.29**	**4.17**	**8.46**	**0.71**	**5.99**	
(d) Last auction						
1	8.46	7.00	15.46	0.98	15.15	
2	**8.46**	**5.00**	**13.46**	**1.00**	**13.46**	2
3	8.46	5.83	14.29	0.99	14.17	

RN: resource number, ST: start time, ET: execution time, CT: completion time, MC: monetary cost, TC: Time ∗ cost, and WR: winner.

Table 4: Resource allocation process of DPAM.

RN	CP	ST	ET	CT	MC	TC	WR
(a) First auction							
1	0.14	0.00	6.00	6.00	0.84	5.04	
2	**0.20**	**0.00**	**4.29**	**4.29**	**0.86**	**3.67**	2
3	0.17	0.00	5.00	5.00	0.85	4.25	
(b) Second auction							
1	0.11	4.29	4.00	8.29	0.45	3.71	
2	0.20	4.29	2.86	7.15	0.57	4.08	3
3	**0.14**	**4.29**	**3.33**	**7.62**	**0.45**	**3.46**	
(c) Third auction							
1	**0.09**	**4.29**	**5.00**	**9.29**	**0.45**	**4.16**	
2	0.16	4.29	3.57	7.86	0.57	4.49	1
3	0.14	7.62	4.17	11.79	0.57	6.68	
(d) Last auction							
1	0.09	9.29	7.00	16.29	0.63	10.22	
2	**0.14**	**9.29**	**5.00**	**14.29**	**0.70**	**10.00**	2
3	0.12	9.29	5.83	15.12	0.70	10.59	

RN: resource number, CP: current price, ST: start time, ET: execution time, CT: completion time, MC: monetary cost, TC: Time ∗ cost, and WR: winner.

completion time is the time required for executing the whole workflow under the partially ordered relation of tasks. The total monetary cost of workflow is the sum of all tasks' monetary cost. The task's monetary cost only covers the execution cost on the allocated resource. It is assumed that there is no additional cost while moving execution from one cloud provider to another. Once a task is assigned to one resource, it will be executed on this resource until its completion and cannot be reallocated to a cheaper or better resource during its execution. The resource allocation process of BOSS is shown in Table 3.

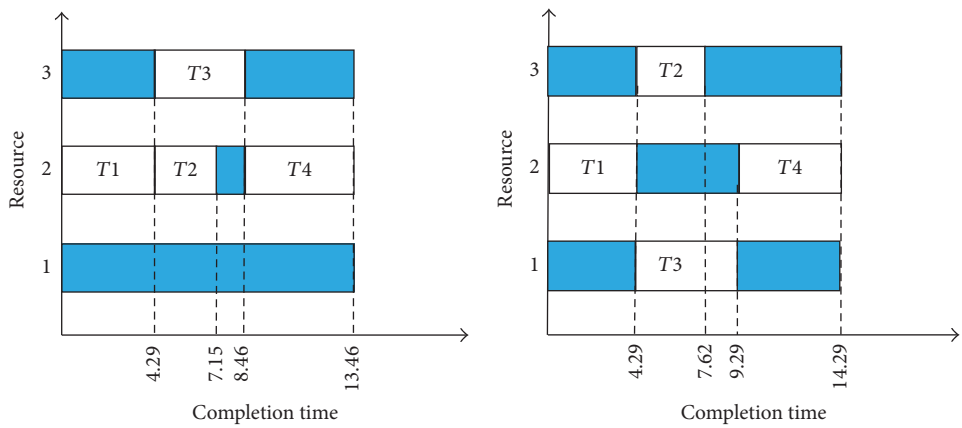
In Table 3, the start time is the larger one between the resource's free time and predecessor tasks' finish time. The execution time is calculated as the workload divided by computation ability. The completion time is the sum of the start time and the execution time. The measurement of TC is the product of completion time and monetary cost. Tasks select the resource with the minimum TC. In Table 3(a), task A selects resource 2 as the winner, because its TC is the minimum. Similarly, task B selects resource 2 as the winner in the second auction and task C selects resource 3 as the winner in the third auction. In the last auction, task D selects resource 2 as the winner. From the tables, the completion time is 13.46 which is the completion time of last task D on resource 2. The total monetary cost of all tasks is 0.86 + 0.57 + 0.71 + 1.00 = 3.14. The product of completion time and total monetary cost is 42.2644.

3.2. Resource Allocation Process of DPAM. In our mechanism, the resource price can be dynamically changed to improve providers' competitiveness. If one provider wins an auction, resource price will be kept unchanged. Otherwise, resource price will be decreased at a certain rate in the next auction. However, during any auction, the resource price cannot be lower than its reserve price. Resource allocation process of DPAM is shown in Table 4. Here the price reduction rate is set as 20%.

At first, task A starts an auction and three resources give their bids (CA, CP). As shown in Table 4(a), the current price is the resource price of current auction. Task A selects resource 2 as the winner, because its TC is the minimum. So resource 1 and resource 3 lose the auction and they decrease the current prices by 20% in Table 4(b). In the second auction, task B selects resource 3 as the winner. Then resource 1 and resource 2 decrease their current price by 20% as shown in Table 4(c). Similarly, task C selects resource 1 as the winner and task D selects resource 2. From Table 4, the completion time of the workflow is 14.29 which is the completion time of last task. The total monetary cost of all tasks is 0.86 + 0.45 + 0.45 + 0.70 = 2.46. The product of completion time and total monetary cost is 35.1534.

3.3. Comparison of BOSS and DPAM. In this subsection, resource prices and the winner of each auction are compared between BOSS and DPAM as shown in Table 5.

In Table 5, the winners of BOSS are 2, 2, 3, and 2. But the winners of DPAM are 2, 3, 1, and 2. In BOSS, resource price is always fixed and resource 1 with low competitiveness is never used. However, in DPAM resource 1 becomes new winner by dynamic pricing strategy which brings higher resource utilization. Resource utilization shows the allocation of tasks on resources (see Formula (2)). Therefore, resource utilization of DPAM is $1/[(2-4/3)^2 + (1-4/3)^2 + (1-4/3)^2]/3 = 9/2$, which is bigger than resource utilization of BOSS ($1/[(3-4/3)^2 + (1-4/3)^2 + (0-4/3)^2]/3 = 9/14$). The reason is that these providers which never become winner in BOSS may win the auction in DPAM. This means more providers win the auction and sell their resources.



(a) Gantt chart of BOSS (b) Gantt chart of DPAM

Figure 2: Gantt charts of resource allocation and task execution.

Table 5: Comparison of BOSS and DPAM.

RN	PB	PD	WB	WD
		(a) First auction		
1	0.14	0.14		
2	0.20	0.20	2	2
3	0.17	0.17		
		(b) Second auction		
1	0.14	0.11		
2	0.20	0.20	2	3
3	0.17	0.14		
		(c) Third auction		
1	0.14	0.09		
2	0.20	0.16	3	1
3	0.17	0.14		
		(d) Last auction		
1	0.14	0.09		
2	0.20	0.14	2	2
3	0.17	0.12		

RN: resource number, PB: price of BOSS, PD: price of DPAM, WB: winner of BOSS, and WD: winner of DPAM.

The Gantt charts of BOSS and DPAM are depicted in Figure 2. The charts show tasks execution order and the resource executes on which task. In Figure 2(a), only resources 2 and 3 are used. While in Figure 2(b) all resources are used. It is easy to draw that DPAM has higher resource utilization than BOSS. In DPAM resource price is dynamic, so resource with weak competitiveness decreases price to improve competitiveness until it wins one auction. However, in BOSS the resource with low competiveness may never win any auction.

4. Dynamic Pricing Strategy

As the number of cloud resource providers increases in reverse auction, they compete against each other to maximize their revenue. So an effective pricing strategy is necessary for providers to increase their competitiveness. Firstly, two propositions are described to prove that the dynamic pricing strategy can improve the revenue of providers and also decrease the monetary cost of users. Then dynamic pricing strategy is proposed.

Proposition 1. *Dynamic pricing strategy can increase the revenue of provider with weak competitiveness.*

Proof. Assume that provider A and provider B have the resources with same computation ability. Resource price of A is p_A and resource price of B is p_B, where $p_A < p_B$. So provider B will lose auction because his competitiveness is weaker than A. If the resource price is fixed, competitiveness of B is always weaker than A and then provider B would never win an auction. Otherwise, if resource price is dynamic, provider B can decrease the price from p_B to p'_B which is lower than p_A. Then B can win auction and hence increase revenue because its competitiveness is higher than that of A. Hence, the dynamic pricing strategy can increase the revenue of provider with weak competitiveness. □

Proposition 2. *Dynamic pricing strategy can decrease user's monetary cost.*

Proof. Assume that provider A and provider B have the resources with same computation ability. Resource price of A is p_A and resource price of B is p_B, where $p_A < p_B$. User will select A's resource because its resource price is lower. If resource price is fixed, user will always select A's resource and monetary cost is p_A. Otherwise, if resource price is dynamic, provider B must decrease the price from p_B to p'_B to win the auction. Here p'_B is lower than p_A. So user will select B's resource and monetary cost is p'_B. Hence, dynamic pricing strategy can decrease user's monetary cost.

Each provider sets reserve price, starting price, and price reduction rate for a resource. When one task starts auction, providers join the auction and give their bids with computation ability and price. After one auction finishes, providers change or do not change the resource price according to

transaction situation. If providers want to increase competitiveness and win auctions, they will change their price according to the dynamic pricing strategy in Formula (1):

$$p_{\mathrm{A}}^{\mathrm{cur}'} = \begin{cases} p_{\mathrm{A}}^{\mathrm{cur}}, & \text{if A is winner,} \\ p_{\mathrm{A}}^{\mathrm{cur}} \cdot (1-\gamma), & \text{if A is loser and } p_{\mathrm{A}}^{\mathrm{cur}} \cdot (1-\gamma) > p_{\mathrm{A}}^{\mathrm{res}}, \\ p_{\mathrm{A}}^{\mathrm{res}}, & \text{if A is loser and } p_{\mathrm{A}}^{\mathrm{cur}} \cdot (1-\gamma) < p_{\mathrm{A}}^{\mathrm{res}}, \end{cases} \tag{1}$$

where $p_{\mathrm{A}}^{\mathrm{cur}}$ refers to the current resource price of provider A. $p_{\mathrm{A}}^{\mathrm{res}}$ refers to the reserve price of resource. γ denotes the price reduction rate. In the strategy, if provider A is the winner, its resource price will still be $p_{\mathrm{A}}^{\mathrm{cur}}$ in the next auction. Otherwise, if provider A is the loser and $p_{\mathrm{A}}^{\mathrm{cur}} \cdot (1-\gamma) > p_{\mathrm{A}}^{\mathrm{res}}$, its resource price will be $p_{\mathrm{A}}^{\mathrm{cur}} \cdot (1-\gamma)$ in the next auction. The resource price will be $p_{\mathrm{A}}^{\mathrm{res}}$, if $p_{\mathrm{A}}^{\mathrm{cur}} \cdot (1-\gamma) < p_{\mathrm{A}}^{\mathrm{res}}$.

In conclusion, dynamic pricing strategy is efficient during the auction. Providers decrease the prices to increase the chance of winning auctions in order to gain more revenue. Simultaneously users choose the resources with lower prices and spend less monetary cost. □

5. Dynamic Pricing Based Allocation Mechanism (DPAM)

In this section, firstly resource utilization (Formula (2)) and evaluation value TC (Formula (3)) are defined and then novel dynamic pricing based allocation mechanism is proposed. In the auction-based cloud market, the purpose of providers is to sell resources at the most proper price so as to gain the highest revenue. And the purpose of users is to execute workflows with shortest completion time and lowest monetary cost. In this mechanism, users select the best resource according to the product of completion time and monetary cost. And the provider with the minimum product will be the winner. After each auction, providers change their resource price according to current trading situation. If their competitiveness is weak and loses the auction, they usually decrease the price in certain rate to increase competitiveness. Otherwise, if they win, it is effectively to keep the price unchanged or increased.

Resource utilization shows the allocation equilibrium of tasks on resources. It is described by variance of winning auction times for each provider. Resource utilization is inversely proportional to variance. Especially when the variance is zero, resource utilization is optimal.

$$ResourceUtilization = \frac{1/\sum_1^n \left(num_j - \overline{num}\right)^2}{n}, \tag{2}$$

where n is the amount of resources. num_j refers to winning times of resource j and $\overline{num}$ is the average value of winning auction times of all providers.

During auction, a task selects the resource with the minimum TC as the winner. TC_{ij} is the product of completion time and monetary cost of task i on resource j. In [12], authors use the measurement TC to measure the BOSS with other mechanisms. There are two reasons for using TC as a measurement: (1) it presents the whole evaluation of completion time and monetary cost for workflow execution; (2) the truthfulness of the BOSS mechanism depends on TC. So we use the measurement TC in order to make a more accurate comparison with BOSS.

$$\mathrm{TC}_{ij} = \left(t_{ij} + \frac{workload_i}{ability_j}\right) * \left(price_j * \frac{workload_i}{ability_j}\right). \tag{3}$$

Each task starts execution only when its predecessor tasks have finished according to the partially ordered relation of tasks. t_{ij} refers to the start time of task i executing on resource j. It equals the latest time when its predecessor tasks have finished and simultaneously resource j is idle. $workload_i$ is the workload of task i; $ability_j$ and $price_j$ are computation ability and price per time unit of resource j, respectively. So $workload_i/ability_j$ is the time required for task i on resource j. And $(t_{ij} + workload_i/ability_j)$ refers to the finishing time of task i on resource j. $(price_j * workload_i/ability_j)$ is the monetary cost required for task i.

In Algorithm 1, there are n tasks and m resources (lines (1)-(2)). Each user starts an auction in order and calculates the product of completion time and monetary cost for every resource (lines (3)–(10)) and then selects the resource with the minimum product (line (11)). Then user pays to the winner (line (12)). At last, all providers change price according to dynamic pricing strategy and join the next auction (line (13)).

When workflows are submitted and tasks start auctions, providers give their bids to compete for the opportunity of providing resources. In the auction, providers change price and increase chances of selling resource, so that they can gain more revenue and higher resource utilization. In addition, users always select the optimal resource, so the product of the completion time and monetary cost of executing tasks is the minimum.

6. Evaluation

In this section, experiments are conducted for evaluation of the performance of BOSS and DPAM on resource utilization and the measurement of TC with different situations and problem sizes. Moreover, the performance of DPAM on resource utilization (see Formula (2)) and TC with different price reduction rates is verified. Firstly, experiment setup is given (see Section 6.1). Secondly, simulation of specific workflow is described for evaluating BOSS and DPAM (see Section 6.2). Thirdly, we conduct experiments with the medium problem size and evaluate BOSS and DPAM with different situations and the performance of DPAM with different price reduction rates (see Section 6.3). At last, both from different situations and different problem sizes, simulation results show the performance of BOSS and DPAM and the performance of DPAM with different price reduction rates (see Section 6.4).

Input: workflows and resources
Output: allocation of tasks on resources
(1) *tasks* ← [*n*]; /* Assign the tasks to *tasks* list with partially relation */
(2) *resources* ← [*m*]; /* Assign the resources to *resources* list */
(3) *i* = 1;
(4) While *i* ≤ *n* do
(5) *tempTask* ← *ith task*;
(6) *j* = 1;
(7) While *j* ≤ *m* do
(8) *tempResource* ← *jth resources*;
(9) *TCs* ← *CalculateTCs*(*tempTask*, *tempResource*);
/* calculate TC of *tempTask* on *tempResource* (Formula (3)) */
(10) End
(11) *winner* ← *resourceWithMinTC*(*TCs*);
/* select the optimal resource with the minimum TC */
(12) *tempTask* pays to *winner*;
(13) all providers *changePrice*;
/* change resource price (Formula (1)) */
(14) End

Algorithm 1: Dynamic pricing based allocation mechanism.

Table 6: Problem size classification.

Small	Medium	Large
$1 \le n \le 40$	$50 \le n \le 100$	$200 \le n \le 300$
$1 \le m \le 10$	$10 \le m \le 50$	$80 \le m \le 120$

6.1. Experiment Setup. The simulation environment runs on a PC with the following configurations: 2 CPU cores, 4 GB RAM, and Microsoft Windows 7 OS. The workflows are classified into three situations: balanced, semibalanced, and unbalanced [12]. Task workload follows normal distribution $N(1000000, 1000)$. The resource ability is set from 200 to 1200 with an arithmetic sequence and the common difference is quotient of 1000 divided by task amount. The resource price is set from the real Amazon Web Services price (https://aws.amazon.com/). In BOSS resource price is set from 0.14 to 0.84 per time unit. In DPAM, to implement dynamic pricing strategy, all resources have starting prices and reserve prices. The starting price is set from 0.14 to 0.84 and reserve price is set from 0.1 to 0.6, respectively.

In simulation of specific cloud workflows, the workflow has 10 tasks and the amount of resources is 7. The price reduction rate for DPAM is 10%. In simulation of general cloud workflows, they are classified into small, medium, and large by problem size besides different situations. Problem size classification is shown in Table 6, where n is amount of tasks and m is amount of resources. In addition, the price reduction rate is set from 0 to 1 in step of 0.1.

6.2. Simulation of Specific Workflows. In specific experiment, specific workflows are used to verify whether DPAM performs better than BOSS on resource utilization and TC.

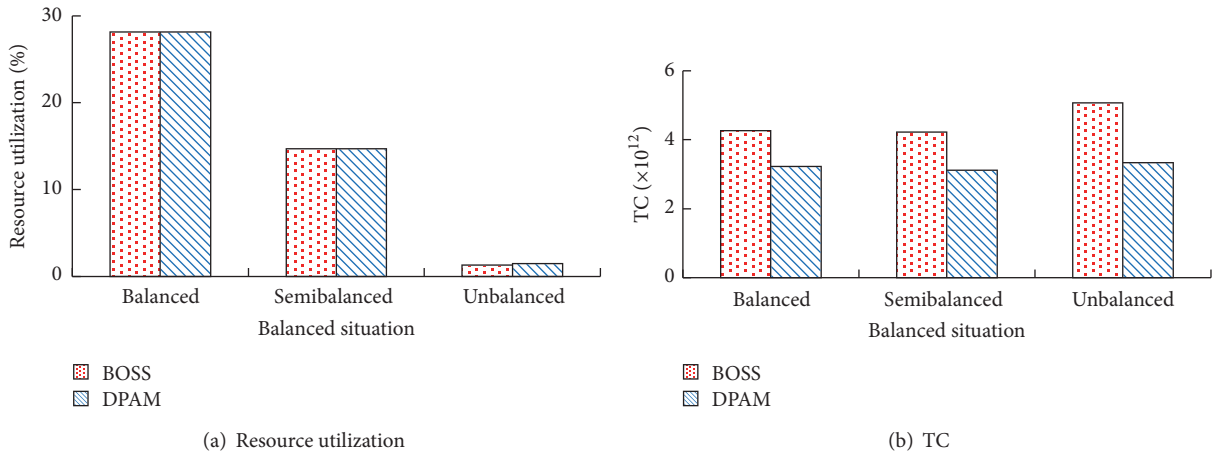
As shown in Figure 3(a), resource utilization of DPAM is always higher than that of BOSS. DPAM can improve resource utilization compared with BOSS. This is because providers with low competitiveness change their resource prices and then these resources have more chances to be sold.

In Figure 3(b), three different situations of TCs of DPAM are all lower than those of BOSS. This means that it takes shorter time and lower monetary cost for workflow execution. In DPAM, providers decrease their resource prices to improve the competitiveness. So users can get the resource with shorter completion time or lower monetary cost.

6.3. Simulation of General Workflows with Different Balanced Situations. In this section, two experiments are conducted on general workflows with different balanced situations. The problem size is medium. The first experiment simulates BOSS and DPAM to evaluate their performance on recourse utilization and TC (see Figure 4). The second experiment simulates DPAM with different price reduction rates to evaluate its performance on resource utilization and TC (see Figure 5).

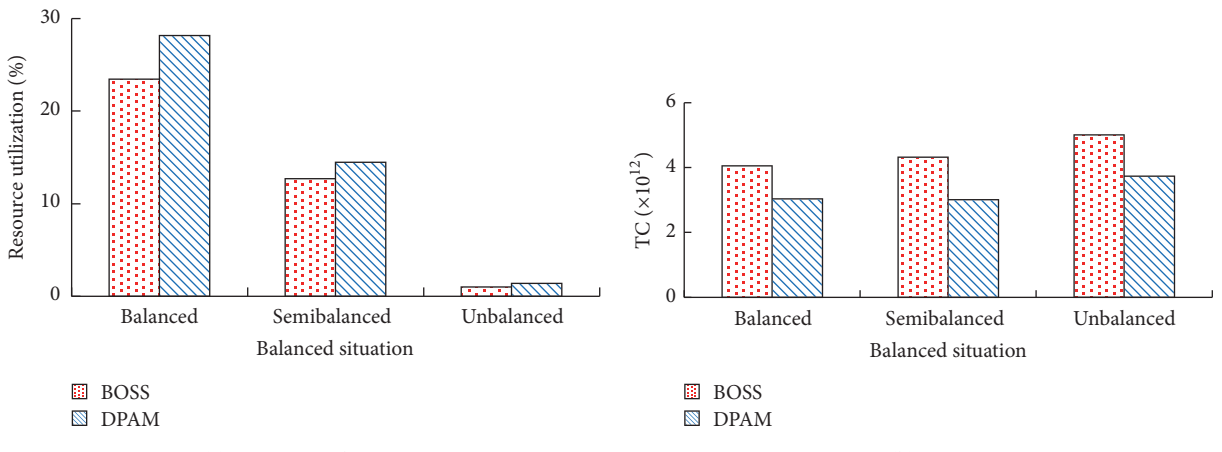
6.3.1. Resource Utilization and TC of BOSS versus DPAM. As shown in Figure 4(a), resource utilization of DPAM is always higher than BOSS. This indicates that DPAM performs better in resource utilization. In DPAM, more resources are sold by changing prices especially for resources with lower competitiveness. These resources are never sold in BOSS. Figure 4(b) shows that TC of DPAM is lower than that of BOSS. DPAM brings shorter completion time and lower monetary cost. The reason is that resource price is dynamic and then there are more resources with higher computation ability and lower price.

6.3.2. Resource Utilization and TC of DPAM with Different Price Reduction Rates. Figure 5 shows resource utilization and TC of DPAM with different price reduction rates. In Figure 5(a) resource utilization is constant when price



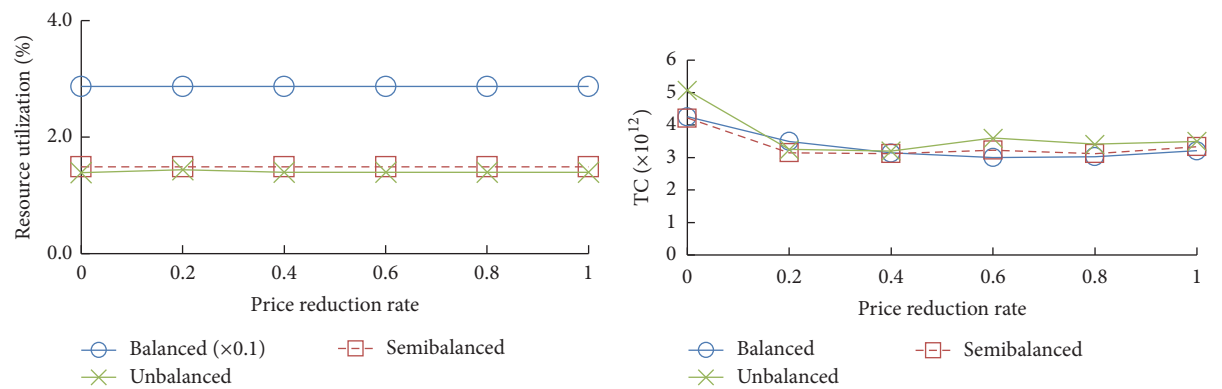
(a) Resource utilization (b) TC

Figure 3: Resource utilization and TC of BOSS versus DPAM for specific workflow.



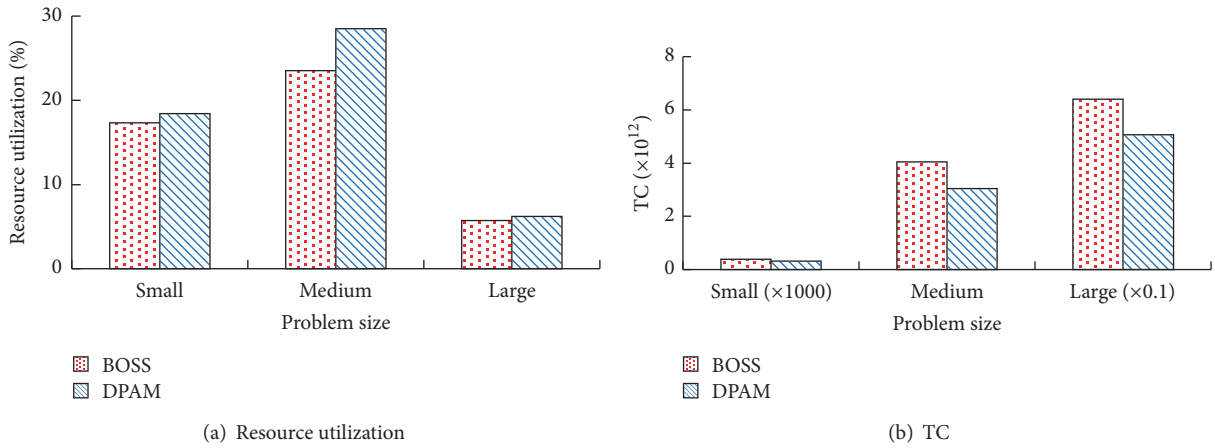
(a) Resource utilization (b) TC

Figure 4: Resource utilization and TC of BOSS versus DPAM for general workflows.



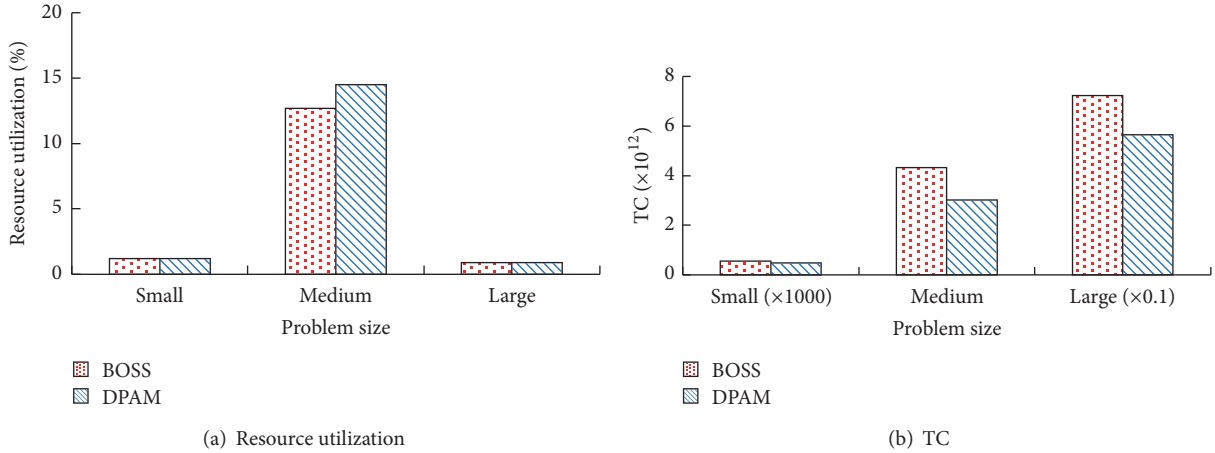
(a) Resource utilization (b) TC

Figure 5: Resource utilization and TC of DPAM with different price reduction rates.



(a) Resource utilization (b) TC

Figure 6: Resource utilization and TC of BOSS versus DPAM in balanced situation.



(a) Resource utilization (b) TC

Figure 7: Resource utilization and TC of BOSS versus DPAM in semibalanced situation.

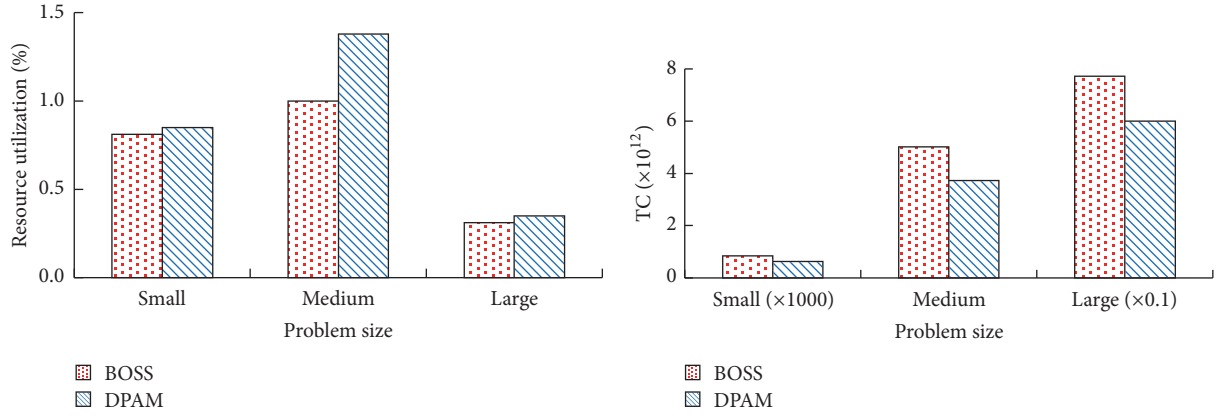
reduction rate is bigger than 0.2. This is because the resource price is equal to the reserve price when price reduction rate is high enough. As shown in Figure 5(b), TC of workflows with all situations decreases when price reduction rate is not zero. It is easy to draw that DPAM is better than BOSS.

6.4. Simulation of General Workflows with Different Problem Sizes. In this subsection, another two sets of experiments conducted on general workflows with different problem sizes and balanced situations are described. The first set of experiments simulates BOSS and DPAM to evaluate their performance on recourse utilization and TC (see Figures 6–8). The second set of experiments simulates DPAM with different price reduction rates to evaluate the performance on resource utilization and TC (see Figures 9–11).

6.4.1. Resource Utilization and TC of BOSS versus DPAM. Figures 6–8 present the performance of BOSS and DPAM on resource utilization and TC from different balanced situations and different problem size. In Figures 6(a), 7(a), and 8(a), resource utilization of balanced workflow is higher that of unbalanced workflow. This is because more tasks in balanced workflow are executed in parallel and many resources are used. In three situations, resource utilizations of DPAM are all higher than that of BOSS. Figures 6(b), 7(b), and 8(b) show that TC of DPAM is always lower than that of BOSS. The reason is that the resource with lower price or higher computation ability is selected as winner.

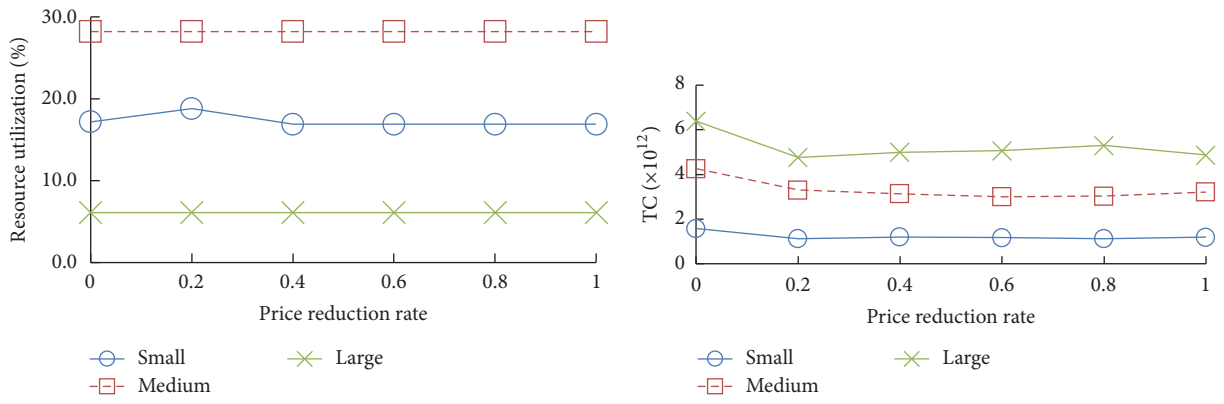
6.4.2. Resource Utilization and TC of DPAM with Different Price Reduction Rates. Figures 9(a), 10(a), and 11(a) show that resource utilization changes only when price reduction rate is lower than 0.3. This indicates that it is not necessary to make price reduction rate too high. The reason is that resource price cannot be smaller than reserve price. Figures 9(b), 10(b), and 11(b) show that TC decreases when resource price reduces in some rates. And TC of large problem size workflows decreases apparently than other sizes. This is because dynamic pricing brings more competitive resources with lower price and higher ability.

In overall terms, the performance of DPAM on resource utilization and TC with different situations is better than BOSS shown in Figure 4. The performance of DPAM on resource utilization and TC with different problem sizes is



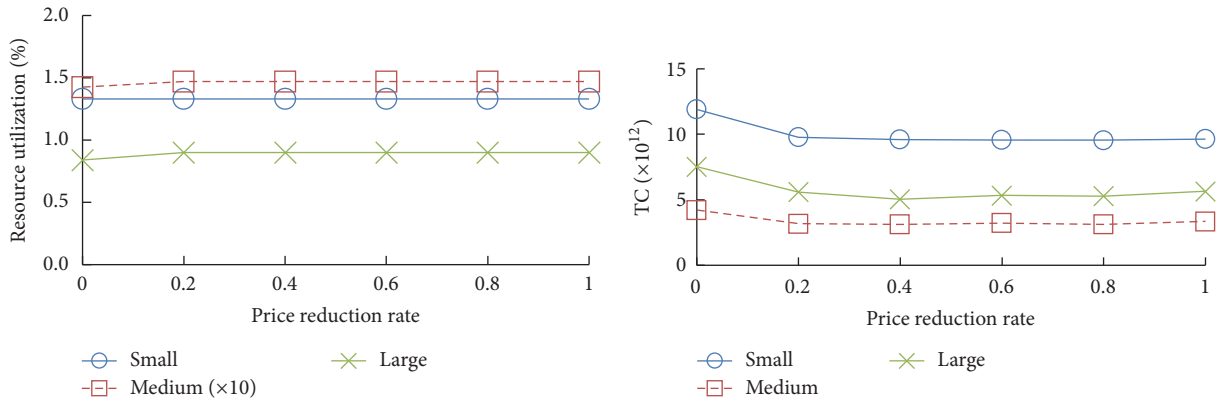
(a) Resource utilization (b) TC

Figure 8: Resource utilization and TC of BOSS versus DPAM in unbalanced situation.



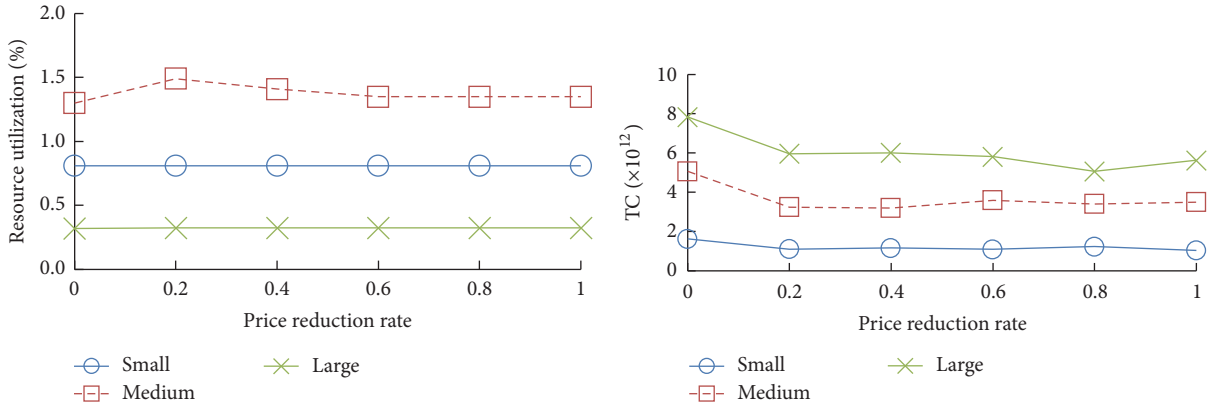
(a) Resource utilization (b) TC

Figure 9: Resource utilization and TC with different rates in balanced situation.



(a) Resource utilization (b) TC

Figure 10: Resource utilization and TC with different rates in semibalanced situation.



(a) Resource utilization

(b) TC

Figure 11: Resource utilization and TC with different rates in unbalanced situation.

shown in Figures 6–8. In DPAM, many providers with weak competitiveness use dynamic pricing strategy to increase chances of making a deal and gain more revenue, so resource utilization of market increases. Meanwhile, workflows can execute timely with less cost. So the performance of DPAM on resource utilization and TC is better than that of BOSS. Moreover, the performance of DPAM on resource utilization and TC with different price reduction rates is shown in Figures 5 and 9–11. Resource utilization and TC are invariant when price reduction rate is higher than 0.2. This is because resource price cannot be lower than the reserve price. In addition, performance of TC and resource utilization is always better when price reduction rate is bigger than zero.

7. Conclusion and Future Work

In this paper, we proposed a dynamic pricing strategy to improve resource providers' competitiveness in the cloud market. A novel dynamic pricing based allocation mechanism was presented to allocate resources for cloud workflows. With our mechanism, resource providers can change the price to increase the possibility of selling resources and gain more revenue, which improves resources utilization. The users select the best resource with the minimum TC (Time $*$ Cost), which ensures shorter completion time and lower monetary cost. Finally, we evaluated our mechanism and compared with the representative BOSS strategy. The results showed that our mechanism can achieve high resources utilization, shorter completion time, and lower monetary cost. With the dynamic pricing strategy, providers can decrease their resource price to improve competitiveness.

In future, increasing price will be involved in dynamic pricing strategy. It is a good way for those resource providers who have sharply higher competitiveness to increase price to gain more revenue. At the same time, we will use the standard scientific datasets to run experiments besides random data. This will increase the credibility of the results of the experiment and be more scientific to reflect the performance of the DPAM mechanism. In addition, besides completion time and monetary cost, we will consider adding other QoS criteria such as reliability, response time, and service providers' reputation.

Competing Interests

There is no conflict of interests related to this paper.

Acknowledgments

This work is partially supported by Natural Science Foundation of China under nos. 61672034, 61300042, and 61300169, MOE Project of Humanities and Social Sciences under no. 16YJCZH048, and the Key Natural Science Foundation of Education Bureau of Anhui Province Project KJ2016A024. The authors are grateful for Professor Yun Yang from Swinburne University of Technology, Australia, for providing constructive feedback to improve this paper. The price reduction rate is set by empirical knowledge. Therefore, the rational rate deserved to be researched.

References

[1] J. Wang, M. AbdelBaky, J. Diaz-Montes, S. Purawat, M. Parashar, and I. Altintas, "Kepler + cometcloud: dynamic scientific workflow execution on federated cloud resources," *Procedia Computer Science*, vol. 80, pp. 700–711, 2016.

[2] G. Juve and E. Deelman, "Scientific workflows and clouds," *Crossroads*, vol. 16, no. 3, pp. 14–18, 2010.

[3] A. Prasad, P. Green, and J. Heales, "On governance structures for the cloud computing services and assessing their effectiveness," *International Journal of Accounting Information Systems*, vol. 15, no. 4, pp. 335–356, 2014.

[4] C. Lin and S. Lu, "Scheduling scientific workflows elastically for cloud computing," in *Proceedings of the 2011 IEEE 4th International Conference on Cloud Computing (CLOUD '11)*, pp. 746–747, Washington, DC, USA, July 2011.

[5] T. T. Huu and C. K. Tham, "An auction-based resource allocation model for green cloud computing," in *Proceedings of the IEEE International Conference on Cloud Engineering (IC2E '13)*, pp. 269–278, San Francisco, Calif, USA, March 2013.

[6] V. Prasad G, S. Rao, and A. S. Prasad, "A combinatorial auction mechanism for multiple resource procurement in cloud computing," in *Proceedings of the 12th International Conference on Intelligent Systems Design and Applications (ISDA '12)*, pp. 337–344, Kochi, India, November 2012.

[7] M. A. Rahman and R. M. Rahman, "CAPMAuction: reputation indexed auction model for resource allocation in Grid computing," in *Proceedings of the 7th International Conference on Electrical and Computer Engineering (ICECE '12)*, pp. 651–654, IEEE, Dhaka, Bangladesh, December 2012.

[8] X. Weng, X. Wang, C.-L. Wang, K. Li, and M. Huang, "Resource allocation in cloud environment: a model based on double multi-attribute auction mechanism," in *Proceedings of the 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom '14)*, pp. 599–604, December 2014.

[9] C. N. Boyer and B. W. Brorsen, "Implications of a reserve price in an agent-based common-value auction," *Computational Economics*, vol. 43, no. 1, pp. 33–51, 2014.

[10] H. Qu, I. O. Ryzhov, and M. C. Fu, "Learning logistic demand curves in business-to-business pricing," in *Proceedings of the 43rd Winter Simulation Conference: Simulation: Making Decisions in a Complex World (WSC '13)*, pp. 29–40, Washington, DC, USA, December 2013.

[11] A. S. Prasad and S. Rao, "A mechanism design approach to resource procurement in cloud computing," *IEEE Transactions on Computers*, vol. 63, no. 1, pp. 17–30, 2014.

[12] H. M. Fard, R. Prodan, and T. Fahringer, "A truthful dynamic workflow scheduling mechanism for commercial multicloud environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1203–1212, 2013.

[13] B. Sharma, R. K. Thulasiram, P. Thulasiraman, S. K. Garg, and R. Buyya, "Pricing cloud compute commodities: a novel financial economic model," in *Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '12)*, pp. 451–457, IEEE, Ottawa, Canada, May 2012.

[14] X. Li, X. Liu, and E. Zhu, "An efficient resource allocation mechanism based on dynamic pricing reverse auction for cloud workflow systems," in *Proceedings of the Asia-Pacific Conference on Business Process Management*, pp. 59–69, 2015.

[15] H. Xu and B. Li, "Resource allocation with flexible channel cooperation in cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 5, pp. 957–970, 2013.

[16] T. Wood, P. J. Shenoy, A. Venkataramani, and M. S. Yousif, "Black-box and gray-box strategies for virtual machine migration," in *Proceedings of the 4th USENIX Conference on Networked Systems Design & Implementation*, pp. 229–242, 2007.

[17] K. Görlach and F. Leymann, "Dynamic service provisioning for the cloud," in *Proceedings of the IEEE 9th International Conference on Services Computing (SCC '12)*, pp. 555–561, June 2012.

[18] X. Shi and Y. Zhao, "Dynamic resource scheduling and workflow management in cloud computing," in *Proceedings of the International Conference on Web Information Systems Engineering*, pp. 440–448, 2010.

[19] M. Mao and M. Humphrey, "Auto-scaling to minimize cost and meet application deadlines in cloud workflows," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '11)*, pp. 1–12, ACM, Seattle, Wash, USA, November 2011.

[20] J. Wang, P. Korambath, I. Altintas, J. Davis, and D. Crawl, "Workflow as a service in the cloud: architecture and scheduling algorithms," *Procedia Computer Science*, vol. 29, pp. 546–556, 2014.

[21] L. Wang, J. Shen, and J. Yong, "A survey on bio-inspired algorithms for web service composition," in *Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD '12)*, pp. 569–574, Wuhan, China, May 2012.

[22] L. Wang and J. Shen, "Multi-phase ant colony system for multi-party data-intensive service provision," *IEEE Transactions on Services Computing*, vol. 9, no. 2, pp. 264–276, 2016.

[23] S. A. Ludwig, "Particle swarm optimization approach with parameter-wise hill-climbing heuristic for task allocation of workflow applications on the cloud," in *Proceedings of the 25th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '13)*, pp. 201–206, IEEE, Herndon, Va, USA, November 2013.

[24] D. Li, C. Chen, J. Guan, Y. Zhang, J. Zhu, and R. Yu, "DCloud: deadline-aware resource allocation for cloud computing jobs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 8, pp. 2248–2260, 2016.

[25] H. Wang, Z. Kang, and L. Wang, "Performance-aware cloud resource allocation via fitness-enabled auction," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 4, pp. 1160–1173, 2016.

[26] M. M. Nejad, L. Mashayekhy, and D. Grosu, "Truthful greedy mechanisms for dynamic virtual machine provisioning and allocation in clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 2, pp. 594–603, 2015.

[27] F. Teng and F. Magoules, "Resource pricing and equilibrium allocation policy in cloud computing," in *Proceedings of the 10th IEEE International Conference on Computer and Information Technology*, pp. 195–202, 2010.

[28] M. Mihailescu and Y. M. Teo, "On economic and computational-efficient resource pricing in large distributed systems," in *Proceedings of the 10th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, pp. 838–843, Melbourne, Australia, May 2010.

[29] L. Pham, J. Teich, H. Wallenius, and J. Wallenius, "Multi-attribute online reverse auctions: recent research trends," *European Journal of Operational Research*, vol. 242, no. 1, pp. 1–9, 2015.

[30] M. Takeda, D. Takahashi, and M. Shobayashi, "Collective action vs. conservation auction: lessons from a social experiment of a collective auction of water conservation contracts in Japan," *Land Use Policy*, vol. 46, pp. 189–200, 2015.

[31] C. Xu, L. Song, Z. Han et al., "Efficiency resource allocation for device-to-device underlay communication systems: a reverse iterative combinatorial auction based approach," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 348–358, 2013.

[32] P. Setia and C. Speier-Pero, "Reverse auctions to innovate procurement processes: effects of bid information presentation design on a supplier's bidding outcome," *Decision Sciences*, vol. 46, no. 2, pp. 333–366, 2015.

[33] J. R. Fooks, K. D. Messer, and J. M. Duke, "Dynamic entry, reverse auctions, and the purchase of environmental services," *Land Economics*, vol. 91, no. 1, pp. 57–75, 2015.

[34] W. Depoorter, K. Vanmechelen, and J. Broeckhove, "Advance reservation, co-allocation and pricing of network and computational resources in grids," *Future Generation Computer Systems*, vol. 41, pp. 1–15, 2014.

[35] Y. Zhao, Y. Li, I. Raicu, S. Lu, W. Tian, and H. Liu, "Enabling scalable scientific workflow management in the Cloud," *Future Generation Computer Systems*, vol. 46, pp. 3–16, 2015.

[36] M. Mihailescu and Y. M. Teo, "Strategy-proof dynamic resource pricing of multiple resource types on federated clouds," in *Algorithms and Architectures for Parallel Processing*, C.-H. Hsu, L. T. Yang, J. H. Park, and S.-S. Yeo, Eds., vol. 6081 of *Lecture Notes in Computer Science*, pp. 337–350, Springer, Berlin, Germany, 2010.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 6907231, 21 pages
http://dx.doi.org/10.1155/2016/6907231

Research Article

Fast Program Codes Dissemination for Smart Wireless Software Defined Networks

Xiao Liu, Tianyi Wei, and Anfeng Liu

School of Information Science and Engineering, Central South University, Changsha 410083, China

Correspondence should be addressed to Anfeng Liu; afengliu@mail.csu.edu.cn

Received 17 July 2016; Accepted 25 August 2016

Academic Editor: Xiong Luo

In smart wireless software defined networks (WSDNs), sensor nodes are deployed in the monitored area to sense data. In order to increase the flexibility of WSDNs configuration, sensor nodes use programmable technology. Thus, programming and software engineering that integrate Internet of Things (IoT) lead to a smart world. Due to the large capacity of program codes and the limited energy of wireless network, only a subset of nodes is selected to spread program codes, and the remaining nodes are in sleep status to save energy. In this paper, a fast program codes dissemination (FPCD) scheme for smart wireless software defined networking is proposed; many nodes in the area far from the sink will be selected to spread program codes; those areas have much energy left, while the area near the sink chooses less number of active nodes to spread program codes to save energy. Thus, FPCD scheme can reduce delay for spreading program codes while retaining network lifetime. The theoretical analysis and experimental results show that our approach can reduce transmission delay by 10.76%–105.791% while retaining network lifetime compares with previous broadcast schemes.

1. Introduction

In the Internet of Things (IoT), millions of objects with sensors collect data and send the data to control center servers that analyze, manage, and use the data in order to construct smart world applications, such as smart manufacturing, smart grid, smart and connected health, smart home, and intelligent transportation systems and even smart city [1–6]. What is more, programming and software engineering are integrated into the Internet of Things and cloud computing, which works towards a smart world.

WSDNs can configure and update program codes in sensor nodes so that IoT has new features, because of the flexibility of configuration and updates of this software, which attracted wide attention of researchers. In WSDNs, program codes in sensor nodes are spread to all network by broadcasting; the program codes have new features after sensor nodes recompiled this program code. Program codes dissemination is a fundamental operation in WSDN [7]. Broadcast is one of the main transmission modes for program codes in WSDNs. The aim of broadcast is to propagate program codes to all nodes in a minimum latency. This problem has been studied extensively and has been shown to be NP-hard [8]. Some urgent applications such as life monitoring, intelligent transportation systems, and industrial automation often have very stringent requirements on program codes dissemination delay [9], and the system does not work properly by a lack of program codes, which could lead to serious accidents, such as the loss of personnel and property. Nevertheless, the harshness of wireless environments and the characteristics of WSDNs pose severe challenges on designing efficient program codes dissemination scheme. Energy of sensor nodes is limited, and thus sensor nodes usually select a small part of nodes to transmit program codes [10–13]. If the number of active nodes is small, the speed for spreading program codes is slow, and the energy consumption in this time is low; thus, network lifetime is higher. Otherwise, if the larger number of active nodes is selected to transmit program codes, the transmission delay is less, but the network lifetime is low. The issue of tradeoff between lifetime and delay is a challenging issue. There are some researches about program codes (or bulk data) dissemination in wireless sensing networks, those researches mainly focus on how to select the optimal scheduling policy

to minimize energy consumption and broadcasts delay. To the best of our knowledge, these studies only achieved tradeoff between network lifetime and propagation delay, but not optimized network lifetime and the propagation delay at the same time. A fast program codes dissemination (FPCD) scheme is proposed to prolong lifetime and reduce broadcast delay at same time toward smart Internet of Things. The main contributions of the FPCD scheme can be listed as follows: (1) FPCD scheme uses a method of global vision to optimize program codes dissemination and thus reduce program codes dissemination delay while retaining origin broadcasting agreement, which has broad applicability.

In this paper, we not only consider the energy consumption of program codes dissemination, but also the energy consumption of other operations in the network, which can take full advantage of energy to improve broadcast performance. WSDNs not only have broadcast operation of program codes, but also data collection operation. Data operation is a regular behavior, but broadcasting and updating program codes are temporary and infrequent operations [14]. For broadcasting program codes, the energy consumption is lower in the area near the sink. But in the process of data collection, the energy consumption in the area far away from the sink is lower, and the energy consumption in the area near the sink is higher. Thus, the FPCD scheme adopts a different strategy, the way is to increase the number of active nodes in the area far away from the sink. Because the transmission delay is related to the number of active nodes, increasing the number of nodes involved in broadcasting can effectively reduce the disseminate delay of program codes.

(2) The optimization parameters and the number of active nodes in different areas are discussed and calculated in the FPCD scheme. The performance of the FPCD scheme outperforms existing ones. Through our simulation study, we demonstrate that, for the proposed scheme, broadcast delay can be reduced and energy efficiency can be enhanced simultaneously. Compared with the former scheme, the broadcast delay can be reduced by as much as 10.76%. More importantly, the proposed scheme improves the performances without harming network lifetime, which is difficult to achieve in the previous schemes.

The rest of this paper is organized as follows. In Section 2, the related works are reviewed. The system model is described in Section 3. In Section 4, a novel FPCD scheme is presented for WSDNs. The performance analysis for FPCD scheme is provided in Section 5. Section 6 shows experimental results and comparison. We conclude in Section 7.

2. Related Work

Program codes (or bulk data) dissemination has been formulated and investigated in [15]. It can be divided into following types depending on different performance requirements of broadcast operation for different application.

(1) The minimum transmission broadcast (MTB) problem: in such studies, the objective of this study is how to reduce broadcast times. In previous schemes, the node is in an active state. Thus, the method for reducing broadcast times is to find a Minimum Connected Dominating Set (MCDS) of the network [16]. In a network as a tree, nodes in the tree can cover the entire network. Thus, as long as one node in the tree broadcasts data packet once, all nodes in the tree can receive the data packet. In [17], authors proved that building a minimum flooding tree is identical to finding an MCDS.

In fact, one goal of reducing broadcast times is to reduce energy consumption. The broadcast method about reducing energy consumption has received a lot of researches. Lim and Kim [18] proposed a heuristic algorithm broadcast BIP (broadcast incremental power). It builds a broadcast tree whose root is source node; in the process of building tree, one node uncovered by the tree will be added to the broadcast tree, and the cost paid is minimized. Wieselthier et al. [19] optimized BIP algorithm; it proposed an effective search algorithm r-shrink; the energy consumption can be reduced by reschedule the nonleaf nodes of broadcast tree.

In most WSDNs, in order to save energy, a node alternates between dormant and active states, this method is developed and applied to WSNs for energy conservation [20–23]. In broadcast, due to this mechanism that a node is required to transmit broadcast message multiple times to all its neighbor nodes at different moments. As a result, the MTB problem in duty cycled networks (MTB-DC problem) needs to be investigated for solutions in which both the set of forwarding nodes and their broadcast schedules are identified.

Liu [24] proposed Level-Based Approximation Scheme to identify the forwarding nodes and their corresponding receivers for all time slots and then constructed a broadcast backbone by connecting these forwarding nodes to the broadcast source.

(2) The minimum latency broadcast scheduling (MLBS): in this scheme, it reduces not only the energy consumption of node, but also the transmission time for broadcasting [25, 26].

Le Duc et al. [25] considered MLBS in duty cycled WSNs and presented two approximation algorithms, BS-1 and BS-2, that produce a maximum latency of at most $((\Delta - 1)TH)$ and $13TH$, respectively. Here, Δ is the maximum degree of nodes, T denotes the number of time slots in a scheduling period, and H is the broadcast latency lower bound obtained from the shortest path algorithm.

Zhao et al. [26] proposed a Broadcast over Duty Cycle and LEACH (BOD-LEACH) protocol, takes advantage of the LEACH's energy efficient clustering. The proposed protocol added new common static and dynamic broadcast periods to support and accelerate broadcasting. The dynamic periods are scheduled following the past arrivals of messages and using a Markov chain model.

Khiati and Djenouri demonstrated that epidemic theory is suitable for data dissemination by using broadcasting protocols in wireless sensor networks [27]. Khiati and Djenouri found that if the number of selected active nodes is large, the energy consumption is higher, the speed of data dissemination is fast, and converge time is lower. Conversely, if the number of selected active nodes is small, the energy

consumption is less; thus, network lifetime is high. Previous schemes adopt fixed parameters to ensure network obtains desired level of performance under dynamic environments. Thus, Khiati and Djenouri proposed an epidemic-based algorithm by dynamically adjusting the number of selected active nodes to meet delay requirements while minimizing energy consumption.

The scheme in [27] is the most similar with the scheme about this paper. But the ratio of active node is equal in the whole network for their scheme. The energy left of network cannot be full used; thus, data dissemination converge time is higher. In the FPCD scheme, the active node ratio is determined according to the adequacy degree of energy which can make full use of energy and improve the level of performance for WSDNs.

3. System Model

3.1. Network Model. (1) We consider WSDNs consisting of a large number of homogenous sensor nodes and one sink node. The sink node is regarded as the control center. Those sensor nodes are deployed in the target area such as environment monitor, industrial field, and smart field [27–30]. The network radius is R, and the transmission radius of sensor nodes is r. The sink node is in the center of a network. The density of sensor nodes in the network is ρ. The sensor nodes are powered by battery whose energy is limited. The energy of the sink is unlimited.

(2) Duty cycling is adopted by nodes to reduce energy consumption. A node has two statues: awake and sleep. The duty cycle is the ratio of time in awake status to the cycle time. When node is in an awake status, the node can spread data. But the energy consumption in this time is higher. And when node is in a sleep status, its wireless transmission device is turned off, and thus the data cannot be transmitted to other nodes, but their energy consumption in this time is lower 100–1000 times than the energy consumption of node when node is in an awake state. Thus, the smaller duty cycle is, the lower the energy consumption is. But, if duty cycle is low, the ratio of active node is small, the delay of data dissemination will be slower, but network lifetime is longer. On the other hand, if duty cycle is high, the ratio of active node is higher, and the delay of data dissemination will be faster at cost of low lifetime.

(3) Senor node senses surrounding environment; when sensor nodes sense event that interests user, the sensed data can be sent to the sink by multiple-hop method. Considering the probability of occurrence of an event is λ. But when the operating system of sensor node upgrade, or the debug in software of sensor node need to be fixed, the software codes of operating system need to be spread to all nodes in the network. The sink collects the data which are sensed by node, this process is called data collection, and the process for disseminating program codes to each node is called data dissemination.

3.2. Energy Consumption Model. In this paper, we adopt the topical energy consumption model in [28–33], where the transmission energy consumption, ω_t, follows (1) and energy consumption, ω_r, for receiving the following (2):

$$\begin{aligned} \omega_{t,1}(d) &= lE_{\text{elec}} + l\varepsilon_{\text{fs}}d^2 \quad \text{if } d < d_0 \\ \omega_{t,2}(d) &= lE_{\text{elec}} + l\varepsilon_{\text{amp}}d^4 \quad \text{if } d \geqq d_0, \end{aligned} \tag{1}$$

$$\omega_r = lE_{\text{elec}}, \tag{2}$$

where E_{elec} represents transmitting circuit loss. Both the free space (d^2 power loss) and the multipath fading (d^4 power loss) channel models are used. If the transmission distance is less than the threshold d_0, the power amplifier loss is based on free space model; if the transmission distance is larger than or equal to the threshold d_0, respectively, the multipath attenuation model is used. ε_{fs} and ε_{amp} are the energy required by power amplification in the two models.

3.3. Problem Statement. In a WSDN, the sink disseminates program codes to sensor nodes; then, the sensor nodes diffuse those program codes to the nodes far from the sink. The aim of this paper is to reduce the dissemination delay [34] while minimizing energy consumption. Thus, the aims of this paper are as follows.

(1) Minimization Program Codes Dissemination Delay. In this paper, the dissemination (or broadcast) delay refers to the time between the sink (control center) disseminates program codes and all nodes receive the program codes except for the nodes that never receive program codes through broadcast. Let T_{init} be the time before broadcast program codes, and let T_{end} be the time when all nodes receive the program codes. Thus, the broadcast delay T can be expressed as

$$\min(T) = \min\left(T_{\text{init}} - T_{\text{end}}\right). \tag{3}$$

(2) Maximization Network Lifetime. Network lifetime Γ refers to the elapsed time when the first node dies in the network. Because the death of first node can affect the connection of network, the program codes sent by the sink cannot broadcast to all nodes. Thus, the aim of this paper is to prolong network lifetime. Let E_{init} stand for the initial energy of nodes in the network, and let e_i be the consumption energy of node in the network. Thus, network lifetime can be expressed as follows:

$$\max(\Gamma) = \max\left(\sum_{i=1}^{n} \frac{E_{\text{init}}}{e_i}\right). \tag{4}$$

(3) Maximization Effective Energy Utilization Rate (Denoted as δ). δ refers to the ratio of energy consumption of network to the total energy in the network. e_i is the energy consumption of node in the network; E_{init} is the initial energy of node in the network. Our target is to maximize energy utilization of whole the network. Thus, the effective energy utilization can be expressed as follows:

$$\max(\delta) = \max\left(\frac{\sum_{i=1}^{n} e_i}{\sum_{i=1}^{n} E_{\text{init}}}\right). \tag{5}$$

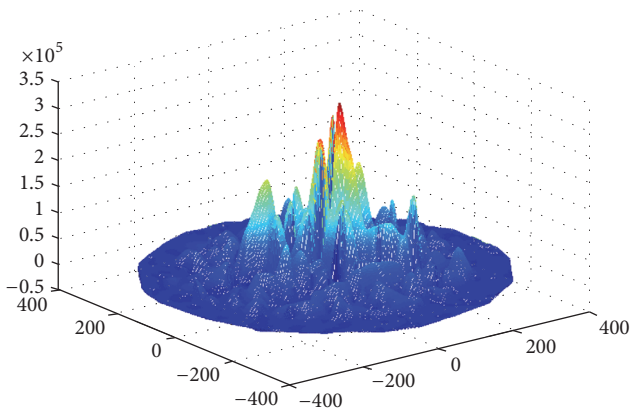


Figure 1: The energy consumption of WSDN.

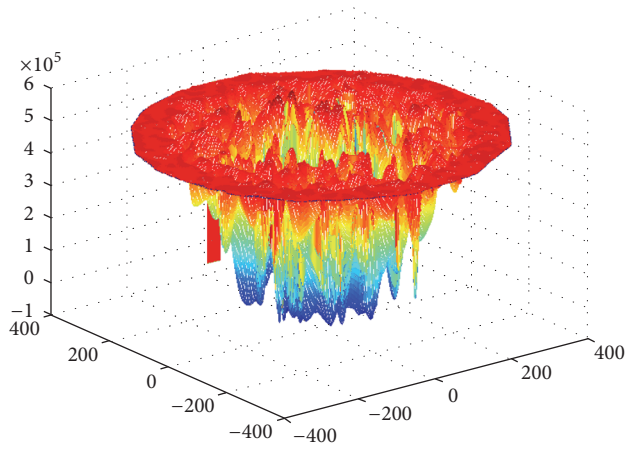


Figure 2: The left energy of WSDN.

Obviously, the goal of the FPCD is to minimize delay, T, maximize the network life, Γ, and make effective energy utilization, δ, which can be summarized as follows:

$$\begin{aligned} \min(T) &= \min\left(T_{\text{init}} - T_{\text{end}}\right) \\ \max(\Gamma) &= \max\left(\sum_{i=1}^{n} \frac{E_{\text{init}}}{e_i}\right) \\ \max(\delta) &= \max\left(\frac{\sum_{i=1}^{n} e_i}{\sum_{i=1}^{n} E_{\text{init}}}\right). \end{aligned} \tag{6}$$

4. Main Design of FPCD

4.1. Research Motivation. (1) The energy left in non-hotspots area is abundant but not in use.

Besides program codes dissemination, WSDNs can collect data from sensor nodes. In the process of data collection, sensor nodes in the network send data to the sink, nodes in the area near the sink transmit much more data to the sink, and thus the energy consumption of nodes in the area near the sink is more than the energy consumption of nodes in the area far away from the sink. It is shown in Figure 1 that the energy consumption of node is unbalanced. The left energy of node is given in Figure 2. As can be seen, many nodes in the network have much energy left. According to relevant research, when the network dies from hotspots problem, the energy left in the network is up to 90%. Thus, the main idea of proposed scheme is to make full use of the energy left in the nonhotspots area to increase the number of active nodes to transmit program codes to reduce dissemination delay.

(2) The number of active nodes is large and the dissemination delay is small, but the energy consumption is higher.

When the sink spreads program codes to all nodes except the nodes that never receive program codes in the network, if the dissemination delay of program codes is smaller, the performance in the network is better. However, in WSDNs, nodes use the way of duty cycling to save energy. When duty cycle of network is smaller, the ratio of active node in the network is lower, and the number of nodes in an awake status is smaller, resulting in the speed for spreading program codes being slower; thus, the transmission delay is bigger. Figure 3 shows the relationship between delivery delay and different number of active nodes. As can be seen, the more the number of active nodes is, the less the dissemination delay is. For some emergency event, the number of active nodes should be increased to reduce transmission delay. The reason is that if there are many active nodes, the sink will have many neighbor nodes in the range of transmission radius. When the sink node wants to spread program codes to the network. In program codes transmission, the program codes can be spread from one node to many nodes, the time for spreading program codes to the network is short, and the transmission delay is less. Otherwise, the transmission delay is higher.

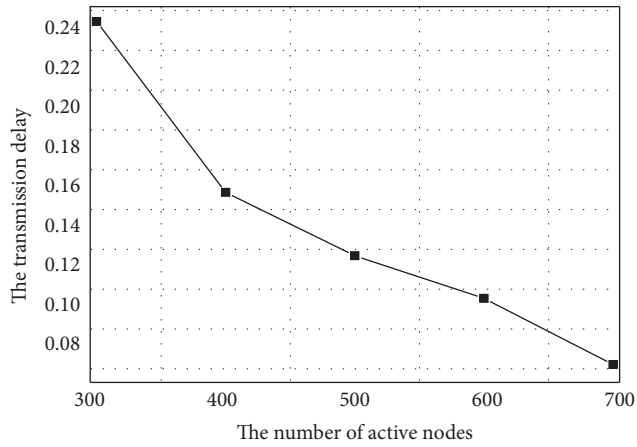


Figure 3: The dissemination delay under different number of active nodes.

(3) If using energy left in the network to increase the number of active nodes, the dissemination delay can be reduced while retaining network lifetime.

In previous schemes, the system cannot reduce dissemination delay under the condition that network lifetime cannot be reduced. The reason is that the larger the number of active nodes is, the less the transmission delay is, but the higher the energy consumption of node is. Thus, increasing the

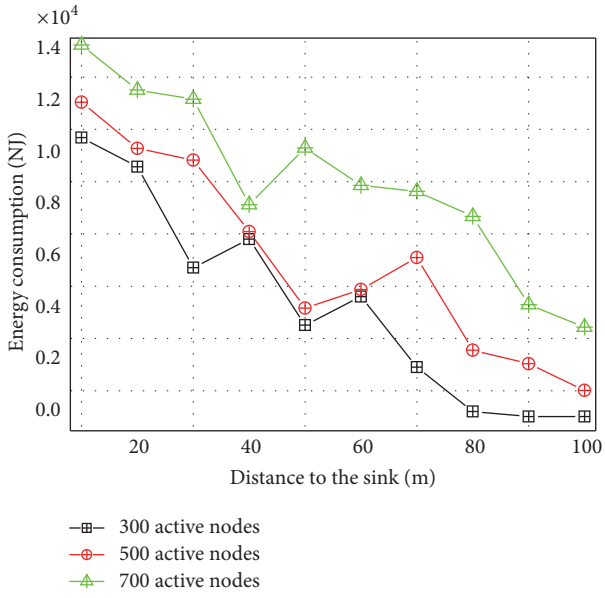


FIGURE 4: The energy consumption in different areas.

number of active nodes may affect the energy consumption to damage network lifetime. The relationship between the energy consumption and the number of active nodes is shown in Figure 4. From Figure 4, with the number of active nodes growing, the energy consumption grows. The main idea of the proposed scheme is to make full use of the energy left in the network to increase the number of active nodes to reduce the transmission delay. In the FPCD scheme, the number of active nodes is adjusted dynamically according to the energy left in the network while retaining network lifetime.

From the above analysis, the broadcast quality of overall network is based on the residual energy. Adjusting the number of active nodes dynamically can achieve better comprehensive broadcast performances. For this reason, a fast program codes dissemination (FPCD) scheme is proposed.

4.2. The Calculation of Energy Consumption of the Network. The nodes at different distances from the sink have definite energy consumption. The value is given by theoretical analysis. Considering that all nodes in network adopt the same duty cycle. The energy consumption of nodes at different areas is calculated.

Firstly, because nodes in different areas use same duty cycle, the number of active nodes in different areas is the same. Due to the same broadcast operation in different areas, the energy consumption in different areas is the same. But the main operation of WSDN is data collection. In the process of data collection, the energy consumption is unbalanced, the energy consumption for data collection is given in Theorem 1.

Theorem 1. *Considering network radius is R, transmission radius of node is r. The probability of occurrence of an event is λ, each data packet is transmitted to the sink using shortest routing path. Considering node i whose distance from the sink is l, $l = zr + x$, ω_t is the energy consumption for transmission of one-bit data. ω_r is the energy consumption for receiving one-bit data. After one round of data collection, the energy consumption E^l_{data} for data operation of node i is*

$$E^l_{data} = \left((z+1) + \frac{z(z+1)r}{2l}\right)\lambda\omega_t + \left(z + \frac{z(z+1)r}{2l}\right)\lambda\omega_r \quad z = \left\lfloor \frac{R-l}{r} \right\rfloor. \tag{7}$$

Proof. Considering network radius is R, the node transmission radius is r, the generation rate of event is λ. The data packet is routed based on the shortest routing. Considering node whose distance from sink is l, $l = hr + x$, where h is the hop counts from the node to the sink. The number of transmitting data loaded by this node is

$$\left((z+1) + \frac{z(z+1)r}{2l}\right)\lambda \mid z = \left\lfloor \frac{R-l}{r} \right\rfloor. \tag{8}$$

The number of receiving data loaded by this node is

$$\left(z + \frac{z(z+1)r}{2l}\right)\lambda \mid z = \left\lfloor \frac{R-l}{r} \right\rfloor. \tag{9}$$

And ω_t is the energy consumption for transmitting data. ω_r is the energy consumption for receiving data. So the energy consumption of node i is

$$\left((z+1) + \frac{z(z+1)r}{2l}\right)\lambda\omega_t + \left(z + \frac{z(z+1)r}{2l}\right)\lambda\omega_r. \tag{10}$$

□

4.3. Adjustable Calculation of the Number of Active Nodes. Compared with previous scheme, the FPCD scheme can adjust the number of active nodes based on the remaining energy. Therefore, the calculating method of the number of active nodes in different regions of the network is given in this section.

Theorem 2. *Considering node i whose distance from the sink is l, $l = hr + x$, ω_t is the energy consumption for transmission one bit data. ω_r is the energy consumption for receiving one bit data. d is the length of program codes. The number of active nodes in the area at l m distance from the sink is n_l in one duty cycle, after one round of program code transmission, the total energy consumption E^l_{pro} of the area whose distance from the sink is l is*

$$E^l_{pro} = n_l\omega_r d + n_{l+r}\omega_t d. \tag{11}$$

Proof. In the FPCD scheme, for node i whose distance from the sink is l, $l = hr + x$. When the sink spreads program codes to all nodes except the nodes that never receive the packet, the total energy consumption of the area whose distance from the sink is l include two kinds of energy: one is the energy consumption for receiving program codes from the last hop node and the other energy consumption is the consumed energy for spreading program codes to the next hop nodes.

There are n_l active nodes in this area, this area will receive n_l program codes from the last node, and thus the energy consumption for receiving those program codes is

$$E_{\text{re}}^{l} = n_l \omega_r d. \tag{12}$$

But those program codes should be spread to next nodes; in the next areas, there are n_{l+r} active nodes, each active node should receive those program codes once. Thus, the energy consumption for transmitting program codes is

$$E_{\text{se}}^{l} = n_{l+r} \omega_t d. \tag{13}$$

The total energy consumption E_{pro}^{l} of the area whose distance from the sink is l is

$$E_{\text{pro}}^{l} = n_l \omega_r d + n_{l+r} \omega_t d. \tag{14}$$

□

Theorem 3. *Considering node i whose distance from the sink is l, $l = hr + x$. $E_{\max}$ is the maximum energy consumption for data collection and spread program code in the network. The number of nodes in the area at l distance from the sink is n_{total}^{l}. In the FPCD scheme, considering the optimized number of active nodes in the area at $l = hr + x$ distance from the sink is $n_l^{optimize}$, it should meet*

$$n_l^{optimize} \alpha + n_{l+r}^{optimize} = \frac{\left(E_{\max} - n_{total}^{l} E_{data}^{l}\right)}{\omega_t d}. \tag{15}$$

Proof. In the FPCD scheme, the aim is to make full use of the energy left to increase the number of active nodes to reduce the transmission delay. For node i whose distance from the sink is l, $l = hr + x$. According to Theorem 2, the total energy consumption of the area at l m distance from the sink is E_{pro}^{l}, the energy consumption of nodes at l m distance from the sink is E_{data}^{l} according to Theorem 1. In order to increase the number of active nodes to reduce transmission delay while retaining network lifetime, the total energy consumption in this area at l m distance from the sink is $E_{\text{pro}}^{l} + n_{\text{total}}^{l} E_{\text{data}}^{l}$. Thus, the difference between the maximum energy consumption of the network and the total energy consumption in this area is $\Delta E = E_{\max} - (E_{\text{pro}}^{l} + n_{\text{total}}^{l} E_{\text{data}}^{l})$. The total energy consumption of this area should be close to the maximum energy consumption to make full use of the energy left in this area. That is,

$$E_{\max} = n_l^{\text{optimize}} \omega_r d + n_{l+r}^{\text{optimize}} \omega_t d + n_{\text{total}}^{l} E_{\text{data}}^{l}. \tag{16}$$

It can be seen that ω_r can be instead as $\alpha\omega_t$, and α is a constant. Equation (16) is

$$E_{\max} = n_l^{\text{optimize}} \alpha\omega_t d + n_{l+r}^{\text{optimize}} \omega_t d + n_{\text{total}}^{l} E_{\text{data}}^{l}. \tag{17}$$

The optimizable number of active nodes in different areas is

$$n_l^{\text{optimize}} \alpha + n_{l+r}^{\text{optimize}} = \frac{E_{\max} - n_{\text{total}}^{l} E_{\text{data}}^{l}}{\omega_t d}. \tag{18}$$

□

4.4. Scheduling Method. The number of active nodes is adjusted adaptively by the FPCD scheme according to the energy left of node. The main goal of FPCD scheme is to increase the number of active nodes in the area far away from the sink to reduce the transmission delay while retaining network lifetime.

In this paper, program codes are transmitted to all nodes after the sink receives the data from all sensor nodes. First, the sink collects data from sensor nodes, and the nodes consume energy. Second, the sink spreads program codes. This scheme mainly studies the program codes transmission. In the beginning of program codes transmission, the energy of nodes is the energy left of node in data collection. In the process of data collection, sensor nodes sense data from the surrounding environment, and then the sensed data can be transmitted to the sink. When those data packets are transmitted from the area at l distance to the sink, the message about energy consumption for transmitting data packet is recorded in the area at l distance from the sink; it is E^l. In the FPCD scheme, when the sink spreads program codes to nodes in the network, the program codes can be spread to their neighbor nodes, until the program codes to all nodes except the nodes that never receive the program codes. Before the sink spreads the program codes, the maximum energy consumption $E_{\max}$ is set as 0, the message about $E_{\max}$ is stored in the program codes. When one node receives the program codes, the energy consumption of node i for transmitting the program codes is E_i, the area where the node stays will record the energy consumption

$$E^l = E^l + E_i. \tag{19}$$

If $E^l > E_{\max}$, the maximum energy consumption can be instead as E^l; that is, $E_{\max} = E^l$. After a round of program codes transmission, the message about the maximum energy consumption of nodes will be returned to the sink. Considering the requirements transmission delay is ζ_r, the transmission delay in this transmission is ζ_c. In the FPCD scheme, its aim is that the number of active nodes can be adjusted dynamically to ensure the transmission delay ζ_c in program codes dissemination smaller than the requirement transmission delay ζ_r, and network lifetime is higher as soon as possible.

The sink will send messages including ($E_{\max}$, ζ_r, ζ_c) to nodes at l distance from the sink. For node i at l distance from the sink, when node i receives the message from the sink, if $\zeta_c > \zeta_r$, it shows that the transmission delay in this time is bigger than the requirement transmission delay, the number of active node should be increased to reduce transmission delay. But if $\zeta_c < \zeta_r$, it shows that the transmission delay in this time is smaller than the requirement transmission delay, the number of active node should be reduced to improve network lifetime.

When $\zeta_c > \zeta_r$, the transmission delay is higher than the requirement transmission delay, the number of active nodes should be increased, and the method for increasing the number of active node is as follows.

(1) When node i receives the feedback message from the sink, it first computes the energy difference between

the maximum energy consumption $E_{\max}$ and the energy consumption E^l; it is as follows:

$$\Delta E = E_{\max} - E^l. \tag{20}$$

When the area at l distance from the sink increase one active node, the energy consumption in the area can increase E^{Δ}, the number of increased active nodes is

$$\Delta n = \frac{\Delta E}{E^{\Delta}}. \tag{21}$$

If $\Delta n < 1$, the number of active nodes will add 1 in the next round of program codes dissemination, if the energy consumption of this area is lower than $E_{\max}$, the number of active nodes is correct in this area. But when the energy consumption of this area is higher than $E_{\max}$, the number of active nodes does not need to be adjusted.

If $\Delta n > 1$, the areas at l distance from the sink will increase by Δn active nodes, the number of adjustable active nodes in this area is as follows:

$$n_l = n_l + \Delta n. \tag{22}$$

The number of active nodes will be instead with the value $n_l + \Delta n$ if the transmission delay is smaller than the requirement transmission delay in the next process of program codes transmission. But if the transmission delay is higher than the requirement transmission delay in the network, the number of active nodes will be instead with $n_l + \Delta n$, and the number of active nodes in the areas nearest the sink adds one.

When $\zeta_c < \zeta_r$, the transmission delay is smaller than the requirement transmission delay; in order to improve the network lifetime, the number of active nodes in the area near the sink is reduced by 1.

Each area at l distance from the sink has marking information $\mathscr{B}^l$ which is used to express whether the number of active node in this area should be adjusted. $\mathscr{B}^l = 1$ shows the number of active nodes in this area should be adjusted in the next process of program codes transmission, $\mathscr{B}^l = 0$ shows the number of active node in this area is not adjusted in the next process of program codes transmission, continuing the process of adjusting the number of active nodes in different areas, until all $\mathscr{B}^l$ are 0 in the network. It shows the number of active nodes in this area is optimized.

When nodes are in an active status, the nodes will spread program codes in the duty cycle time, until the program code reaches all nodes except the nodes that never receive program codes. The number of active nodes in the area will be adjusted in the next cycle time, until the system reaches stable status.

The FPCD scheme algorithm is given in Algorithm 1.

5. Performance Analysis in Theory

5.1. Broadcast Delay Analysis. This section analyzes the transmission delay in the FPCD scheme. Program codes are generated by the sink and then are spread to the area far away from the sink. In this paper, the broadcast method for program code is that the sink sends program codes to nodes in the transmission range, the nodes that receive the program codes continue transmitting program codes until reach the network boundary. In this process, each node only receives program codes once when the program codes are transmitted by the sink once. Program codes dissemination can be divided into 2 stages; in the first stage, program codes are spread to the area far away from the sink, until program codes reach the network boundary. In the second stage, though program codes reach the network boundary, some nodes cannot receive program codes; program codes are broadcast several times until the broadcast is over. The first stage is a stage of expansion; the area for broadcasting program codes expands outward; the second stage is the stabilization phase. At this stage, the area of program codes dissemination is not expanded, but the number of broadcast nodes is increased. Thus, the delay of program codes dissemination can be divided into two parts: (1) the delay in the expansion phase; (2) the delay in stabilization phase. The delay in the expansion phase mainly depends on the time for spreading program codes to the network boundary. However, the requirement time for spreading program codes to the network boundary is very complex; it depends on several factors: (1) network size: the larger the network is, the longer the transmission delay is; (2) the transmission radius of node: the bigger the transmission radius is, the farther the transmission distance is when program codes are transmitted once, and thus the lower the transmission delay is; (3) the ratio of active node: apparently, the larger the number of active nodes is, the farther the distance traveled is, and the smaller the delay is. In the following, we analyze the expected distance for spreading program codes to the area far away from the sink, so that the expected hop for spreading program codes to the network boundary can be got in theory, which can obtain the minimum expected propagation delay in the first stage.

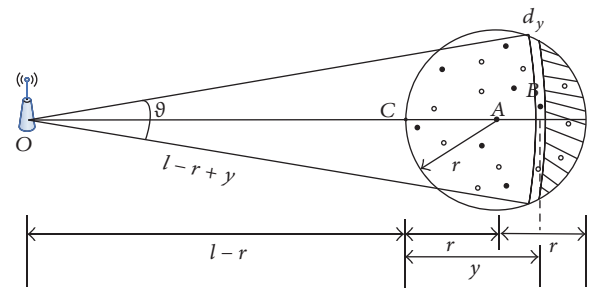


Figure 5: The illustration of the broadcast forwarding distance.

Definition 4 (dissemination forwarding distance y). As shown in Figure 5, node A is broadcast node, node B is in the transmission range of node A, and the distance from node B to the sink is farther than the distance from nodes in the transmission range of node A to the sink. A line OA is used to connect the sink O and node A, and C is the intersection of broadcast radius of node A and line OA. The length from longitudinal projection point of node B in line OA to C is broadcast forwarding distance y.

/** the requirement minimum transmission delay is ζ_r, the energy consumption of the area at l distance from the sink is E^l after the data collection is over. **/
Initialize: each area at l distance from the sink records the energy consumption in the last program code spread period: E^l. The maximum energy consumption $E_{\max} = 0$. $\Delta\zeta$ is a delay threshold
Part I: The action of node in message report to sink:
(1) **For** program code P which reach node i at l distance from the sink **Do**
(2) $E^l = E^l + E_i$; // the energy consumption will be record in
// this area
(3) **If** ($E^l > E_{\max}$)
(4) $E_{\max} = E^l$;
(5) **End if**
(6) send program code P with $E_{\max}$ to next node
(7) **If** the number of active nodes in this area changes
(8) $\mathscr{B}^l = 1$;
(9) **Else**
(10) $\mathscr{B}^l = 0$;
(11) **End IF**
(12) **End For**
Part II: The action of sink:
(13) $\zeta_u = \zeta_c - \zeta_r$,
(14) **For** each program code flow where $\zeta_u < 0$ and $|\zeta_u| > \Delta\zeta$
// the current transmission delay is lower than that the requirement
// transmission delay of application, the number of active nodes
// should be reduced.
(15) The last node in the program code path send tuple ($E_{\max}, \zeta_r, \zeta_c$) along the reverse routing directions to different areas in this program code flow path;
// feedback control the data flow which do not meet the
// requirement of application
(16) Sink get ζ_c, $E_{\max}$ from program code P;
(17) $E_{\max}$ can be updated by sink after a round of program codes transmission
(18) Sink send ($E_{\max}, \zeta_r, \zeta_c$) to each area in the network;
(19) Each area receives the tuple do action as steps (20)–(27);
(20) **If** $\mathscr{B}^l = 1$
// The number of active nodes in this area is reduced to reach
// the stable state
(21) The number of active nodes in this area should be adjust according to steps (45)–(51);
(22) update $E_{\max}, \zeta_c$;
(23) send tuple ($E_{\max}, \zeta_r, \zeta_c$) to each node;
(24) **Else**
(25) **If** (the current reliability do not meet the requirement of application)
// the number of active nodes in the area near sink can be reduced.
(26) Let $n_l = n_l - 1$
// The number of active nodes in the area with largest energy consumption minus 1
End if
(27) **End if**
(28) **End for**
(29) **For** each data flow where $\zeta_u > 0$ and $|\zeta_u| > \Delta\zeta$
// the current transmission delay is bigger than that the requirement
// transmission delay of application, the number of active nodes
// should be increased.
(30) Continue steps (15), (16), (17), and (18)
(31) **If** the area at l distance from the sink receives the tuples;
(32) **If** $\mathscr{B}^l = 1$
// The number of active nodes in this area is increased
(33) The number of active nodes in this area should be adjust according to steps (45)–(51);
(34) **Else**
(35) **If** $|\zeta_u| < \Delta\zeta$
// It shows that the system is in a balanced state.
(36) Stop the process for adjust the number of active nodes in different areas
(37) **Else**
// The number of active nodes in the area nearest sink can be
// increased.
(38) Let $n_l = n_l + 1$
(39) **End if**
(40) **End if**

Algorithm 1: Continued.

(41) update $E_{\max}$, ζ_c;
(42) **End for**
(43) Continue the steps (13)–(42) for adjusting the number of active nodes, until $|\zeta_u| < \Delta\zeta$;
(44) *The method for adjusting the number of active nodes in the area at l distance from the sink*
(45) Let $\Delta E = E_{\max} - E^l$, $E^{\Delta} = \Delta E\vartheta$;
(46) Compute Δn_l base Eq. (12) according ΔE;
(47) **If** $\Delta n_l > 1$
(48) $n_l = n_l + \Delta n_l$;
(49) **Else**
(50) $n_l = n_l + 1$;
(51) **End if**

Algorithm 1: A fast program codes dissemination (FPCD) scheme.

Because the transmission range is r, broadcast forwarding distance y is in $[0, 2r]$. If $y > r$, this shows that program codes can be transmitted to nodes whose distance from the sink is longer than the distance from node A to the sink. But if $y < r$, this shows that program codes cannot be transmitted to nodes whose distance from the sink is longer than the distance form node A to the sink. The reason is that when node A broadcasts program codes, all nodes in the broadcast radius range of node A, whose distance from the sink is longer than the distance from node A to the sink, are in sleep states, and areas whose distance from the sink is shorter than the distance from node A to the sink have active nodes, and thus $y < r$. Such a broadcast is meaningful, and the program codes can be spread in the next time.

Theorem 5. *Considering the distance from node A to the sink is l m, the transmission radius is r and the density of network is ρ. The ratio of active node is ε. When program codes are spread once by node A, the expected traveled propagation distance is $\mathcal{G}_l$:*

$$\mathcal{G}_l = \int_0^{2r} (y - r)\left(1 - (1 - \varepsilon)^{\vartheta(l-r+y)d_y\rho}\right)(1 - \varepsilon)^{A_l^y \rho}, \quad (23)$$

where $\vartheta = 2\cos^{-1}((l^2 + (l + y - r)^2 - r^2)/2l(l + y - r))$,

$$\begin{aligned} A_l^y = {} & 2r^2\cos^{-1}\xi - \xi\sqrt{(l + y - r)^2 - \xi^2} \\ & - 2\cos^{-1}\frac{\xi}{(l + y - r)}(l + y - r)^2 \\ & - \xi\sqrt{(l + y - r)^2 - \xi^2}, \end{aligned} \quad (24)$$

$$\xi = \frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}.$$

Proof. As shown in Figure 5, when the maximum distance for spreading program codes is $l - r + y$, it shows that all nodes in the shaded areas are in sleep status, but some nodes in the range of $l - r + y$ distance are in active status. The area of the black shaded area in Figure 6 is calculated as follows: the sector area $\triangleleft CAB$ as the center of A minus arcuate area $\widehat{COB}$

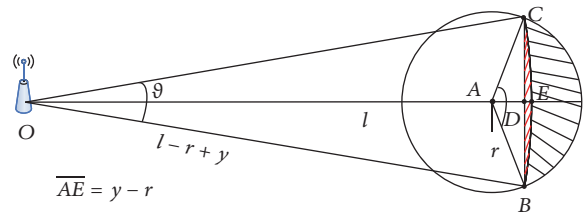


Figure 6: The calculation of the broadcast forwarding distance.

which is the intersecting area of the arc as the center of the sink O and the arc as the center of A. The arcuate area $\widehat{COB}$ is divided into two parts: one is the area of the red shaded part, and the other is the area of triangle $\triangle CAB$. The area of the red shaded part is the sector area $\triangleleft COB$ minus the area of triangle $\triangle COB$. Due to $\vartheta = \angle COB = 2\cos^{-1}((l^2 + (l + y - r)^2 - r^2)/2l(l + y - r))$. Thus, the area of sector area $\triangleleft COB$ is $2\cos^{-1}((l^2 + (l + y - r)^2 - r^2)/2l(l + y - r))(l + y - r)^2$, the length of line OD is $(l^2 + (l + y - r)^2 - r^2)/2l$, and the length of line DB is $\sqrt{(l + y - r)^2 - ((l^2 + (l + y - r)^2 - r^2)/2l)^2}$. The area of triangle $\triangle COB$ is as follows:

$$\sqrt{(l + y - r)^2 - \left(\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}\right)^2} \cdot \frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}. \quad (25)$$

Thus, the area of red shaded area is

$$2\cos^{-1}\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l\,(l + y - r)}(l + y - r)^2 - \sqrt{(l + y - r)^2 - \left(\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}\right)^2} \cdot \frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}. \quad (26)$$

The length of line AD is $(l^2 + (l + y - r)^2 - r^2)/2l - l = ((l + y - r)^2 - r^2 - l^2)/2l$. The area of triangle $\triangle CAB$ is

$$\frac{(l + y - r)^2 - r^2 - l^2}{2l} \cdot \sqrt{(l + y - r)^2 - \left(\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}\right)^2}. \quad (27)$$

$\angle CAB = 2\cos^{-1}(((l + y - r)^2 - r^2 - l^2)/2lr)$. Thus, the area of sector area $\triangleleft CAB$ is $2r^2\cos^{-1}(((l + y - r)^2 - r^2 - l^2)/2lr)$, the area of the black shaded area in Figure 6 is

$$\begin{aligned} A_l^y = {} & 2r^2\cos^{-1}\frac{(l + y - r)^2 - r^2 - l^2}{2lr} \\ & - \frac{(l + y - r)^2 - r^2 - l^2}{2l} \\ & \cdot \sqrt{(l + y - r)^2 - \left(\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}\right)^2} - 2 \\ & \cdot \cos^{-1}\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l\,(l + y - r)}(l + y - r)^2 \\ & - \sqrt{(l + y - r)^2 - \left(\frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}\right)^2} \\ & \cdot \frac{\left(l^2 + (l + y - r)^2 - r^2\right)}{2l}. \end{aligned} \quad (28)$$

However, the number of nodes in shaded area is $A_l^y\rho$, the probability of which all nodes in shaded area are in sleep status is $(1 - \varepsilon)^{A_l^y\rho}$.

But the probability of which area within the range of distance $l - r + y$ from the sink has nodes in awake status is as follows: considering the width of this area is $d_y \mid d_y \to 0$, thus the area of this area is $\vartheta(l - r + y)d_y$.

The number of nodes in this area is $\vartheta(l - r + y)d_y\rho$; thus, the probability of which area has one node at least in awake status is as follows:

$$1 - (1 - \varepsilon)^{\vartheta(l-r+y)d_y\rho}. \quad (29)$$

The probability of spreading program codes to the area at $l - r + y$ distance from the sink is as follows:

$$\left(1 - (1 - \varepsilon)^{\vartheta(l-r+y)d_y\rho}\right)(1 - \varepsilon)^{A_l^y\rho}. \quad (30)$$

In this time, the forward propagation distance is $y - r$. Therefore, the total expected forwarding distance is as follows:

$$\mathscr{G}_l = \int_0^{2r} (y - r)\left(1 - (1 - \varepsilon)^{\vartheta(l-r+y)d_y\rho}\right)(1 - \varepsilon)^{A_l^y\rho}. \quad (31)$$

□

Since the expected forwarding distance is shown in Theorem 5 when program codes is spread in one hop, the following will estimate broadcast times of spreading program codes from the sink to the network boundary. $\mathscr{G}_l$ in Theorem 5 is the expected forwarding distance of spreading program codes from nodes at l m distance from the sink to other node. Because the $\mathscr{G}_l$ in different distance has little difference, using $\mathscr{G}$ to express the forwarding distance in the network with transmission radius is r. If the distance from the sink to the network boundary is R, the hop count can be calculated as

$$\mathscr{H}^r = \left\lceil \frac{R}{\mathscr{G}} \right\rceil. \quad (32)$$

Since the delay for spreading program codes in each hop is τ, the time for spreading program codes from the sink to the network boundary is $\tau\mathscr{H}^r$. When program codes is spread to the network boundary, the network will be into the second stage. In the process of spreading program codes to network boundary, nodes that receive program codes will broadcast the program codes to neighbor nodes, nodes in the network boundary will be the last to receive the program codes. Thus, the time for spreading program codes in second stage is the time of spreading program codes to nodes in the network boundary. The required time in second stage is regarded as the time in which all nodes in the broadcast cycle receive program codes, considering the duty length of node is τ, its duty cycle ε can ensure the ratio ε of active node. Because program codes can be broadcast once in one τ, thus after $(1/\varepsilon)\tau$ cycle, each node will be in awake status once, all nodes can receive the program codes. From the above analysis, it can be obtained that the propagation delay of the entire network is as follows:

$$\mathfrak{D} = \left(\left\lceil \frac{R}{\mathscr{G}} \right\rceil + \frac{1}{\varepsilon}\right)\tau. \quad (33)$$

Theorem 5 gives the dissemination forwarding distance. Many factors affect dissemination forwarding distance; Figure 7 shows the relationship between the active node ratio and forwarding distance. In Figure 7, the forwarding distance refers to the transmission distance for spreading program codes to the area far away from the sink; it is expressed as $\vec{y}$. As can be seen from Figure 5, if most areas in the transmission radius of node A do not have nodes in active status, then $\vec{y} \le 0$. The probability of this situation is $(1 - \varepsilon)^{\rho\pi r^2/2}$, it is small. Thus, generally speaking, $\vec{y} > 0$. The bigger ε is, the probability that there are active node in the area far away from the sink in the range of broadcast of node A is bigger, and $\vec{y}$ is larger. The results are shown in Figure 7; the bigger ε is, the bigger $\vec{y}$ is, which shows that the forwarding distance of broadcast once is bigger, and the delay of spreading program codes to entire network is smaller.

Figure 8 shows the $\vec{y}$ in different ρ. Its trend is the greater node density is, the greater its $\vec{y}$ is. If the number of active nodes is large, the probability that the node far away from the sink in the range of broadcast of node A receive program codes is larger, the longer the forwarding distance is. At the same time, when the density of node is increased, program

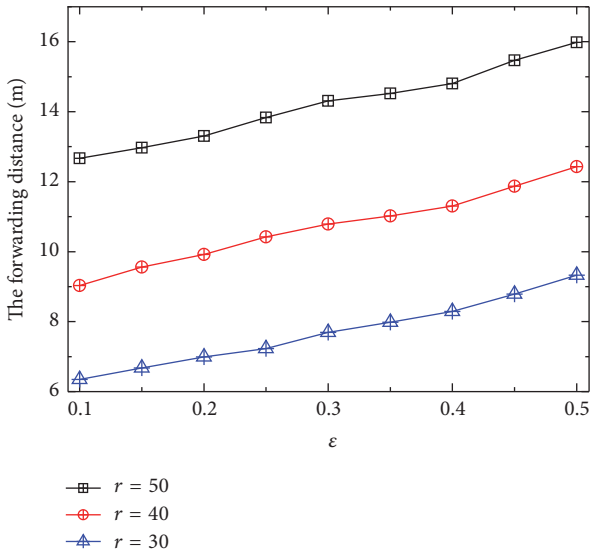


Figure 7: The relation of ε and forwarding distance.

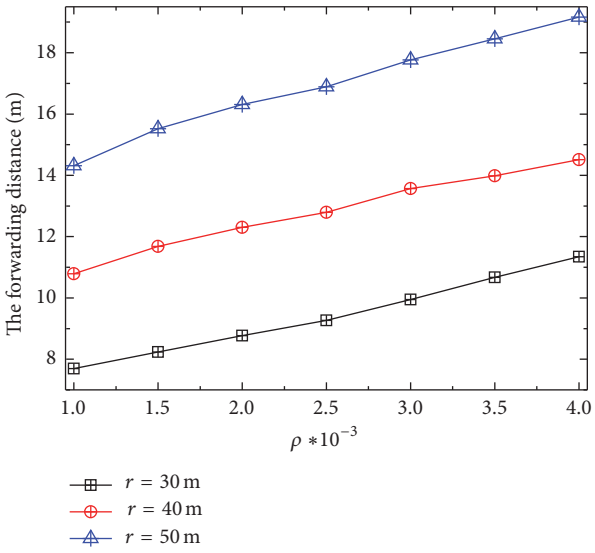


Figure 8: The relation of ρ and forwarding distance.

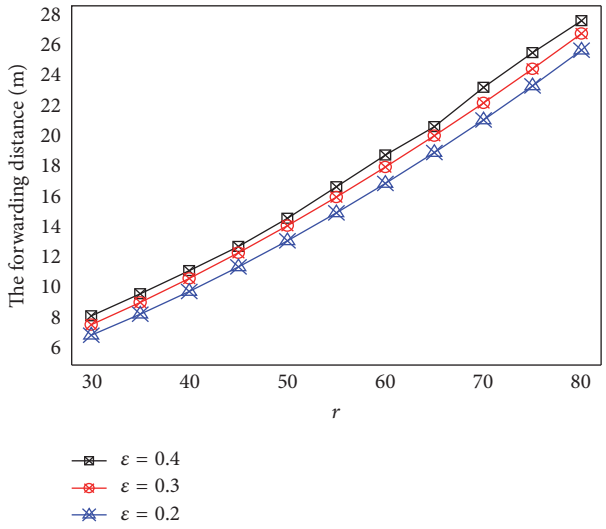


Figure 9: The forwarding distance under different ε and r.

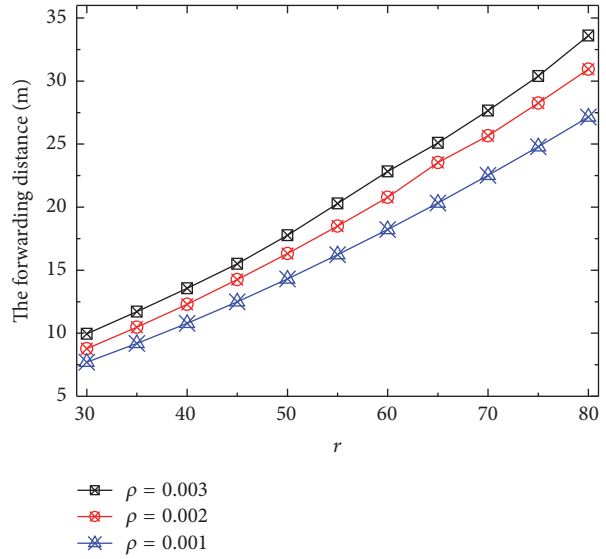


Figure 10: The forwarding distance under different ρ and r.

codes are easily diffused and the delay for spreading program codes is smaller.

Figure 9 shows the forwarding distance under different transmission radius r. As can be seen, the bigger r is, the bigger $\vec{y}$ is, and the smaller the transmission delay is. But when $r = R$, the delay of network in the first stage is 0. Figure 10 gives $\vec{y}$ under different r and ρ. The rule is that the bigger r is and the bigger ρ is, the bigger $\vec{y}$ is.

Figures 11 and 12 give the transmission delay under different node density ρ and different active node ratio ε. The transmission delay in the figure is referred to the expected delay of network. The rule is that when the node density and the active node ratio ε are increased, the dissemination delay is reduced. Its reason is that when ρ and ε are increased, large number of nodes can receive the program codes in program codes dissemination, the data can be broadcast to other nodes by those nodes, the speed for broadcasting data is fast. When ρ and ε are increased, the forwarding distance is longer when program codes broadcast once according to Theorem 5, the speed of spreading program codes to network boundary is fast, and the delay is small. The bigger r is, the longer forwarding distance is, and the smaller the delay is.

Based on Section 4, Figure 13 shows the value of ε in different areas. In the FPCD scheme, the area near the sink adopts the same ε with the previous scheme; it is the initial values 0.2 and 0.3 of ε in Figure 13. Then, the FPCD scheme increases the ε in the area far from the sink according to the energy consumption. It can be seen that ε in the area far away from the sink is higher than the initial ε.

Since different areas have different ε in the FPCD scheme, the forwarding distance in different areas is different. The

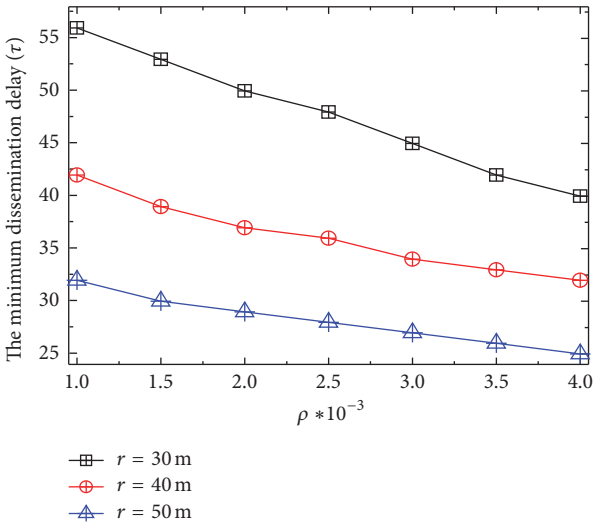


Figure 11: The dissemination delay under different r and ρ.

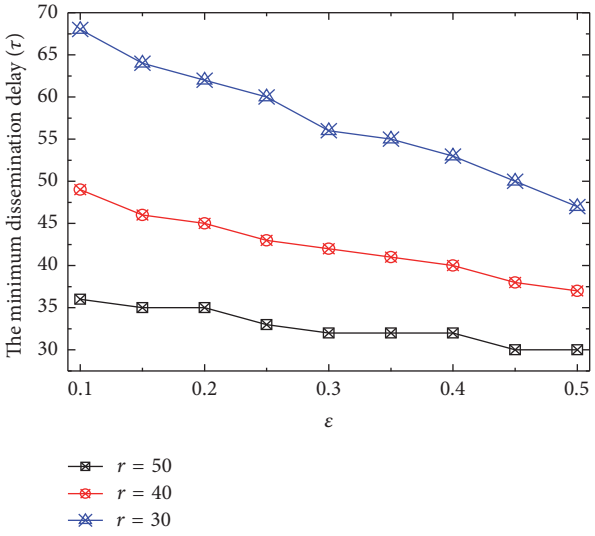


Figure 12: The dissemination delay under different r and ε.

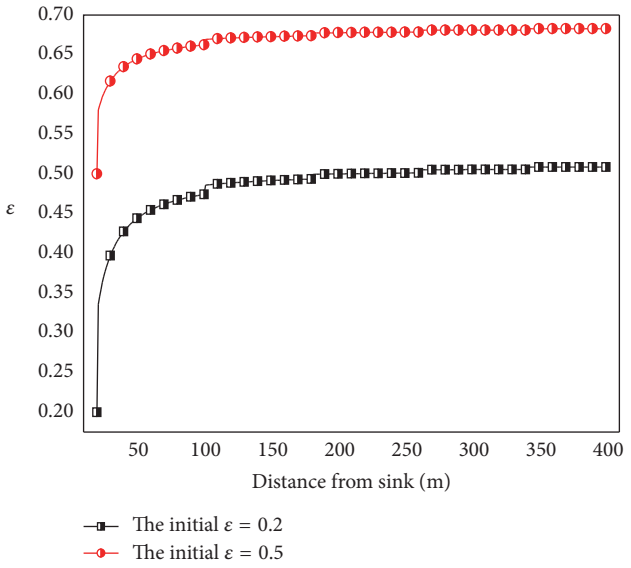


Figure 13: The value of ε in a FPCD scheme.

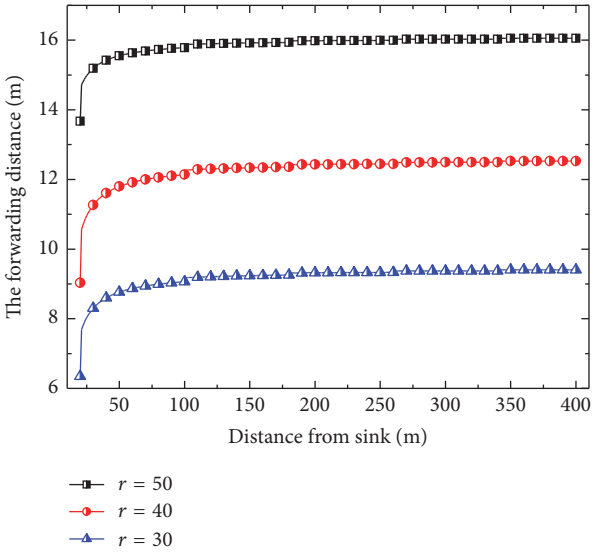


Figure 14: The forwarding distance in a FPCD scheme.

bigger ε is, the longer the forwarding distance is, and the delay is smaller; it is shown in Figure 14. Figure 15 gives the dissemination delay in the FPCD scheme and other schemes. Compared to the scheme with the same ε, the dissemination delay in the FPCD scheme is reduced by 13.25%–27.57%.

5.2. Broadcast Times Analysis. In program codes dissemination, broadcasting program codes is the most energy-consuming operation; therefore, the best method to save energy is to reduce broadcast times. There is an example to illustrate the relationship of ε and broadcast times. The bigger ε is, the smaller the requirement broadcast times are. The most special case is that when $r = R$, it shows that the sink broadcasts program codes once, and all nodes in the network can receive program codes. If $\varepsilon = 1$, all nodes in the network can receive data when the sink spread program codes once. But if ε is smaller, the number of nodes that receive data is less; thus, the requirement broadcast times are long. Theorem 6 gives the requirement broadcast times in a certain ε.

Figure 16 shows the relationship between ε and broadcast times. With the increase of ε, the requirement broadcast times of program codes dissemination are reduced. The broadcast times can be obtained according to the different ε in different areas in Figure 13 and Theorem 6 (see Figure 17). Finally, Figure 18 shows the total broadcast times in the FPCD scheme and other schemes; it can be seen that the FPCD scheme can reduce the requirement broadcast times greatly.

Theorem 6. *Considering a network with network radius R, the transmission radius is r and node density of network is ρ. The*

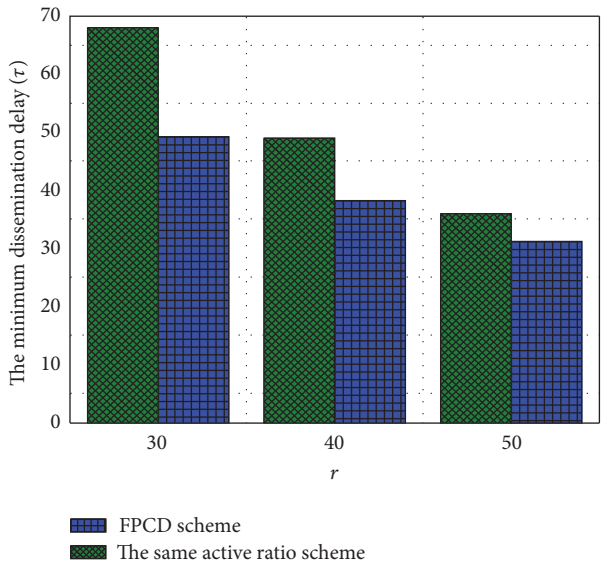


Figure 15: The dissemination delay in FPCD scheme and other schemes under different r.

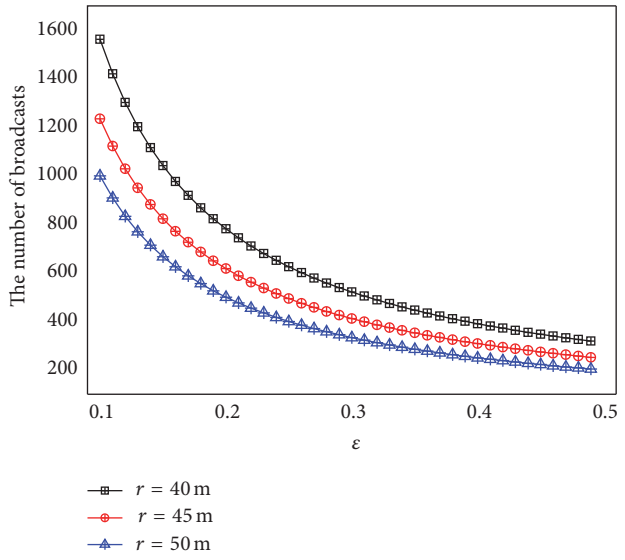


Figure 16: The number of broadcasts under the different r.

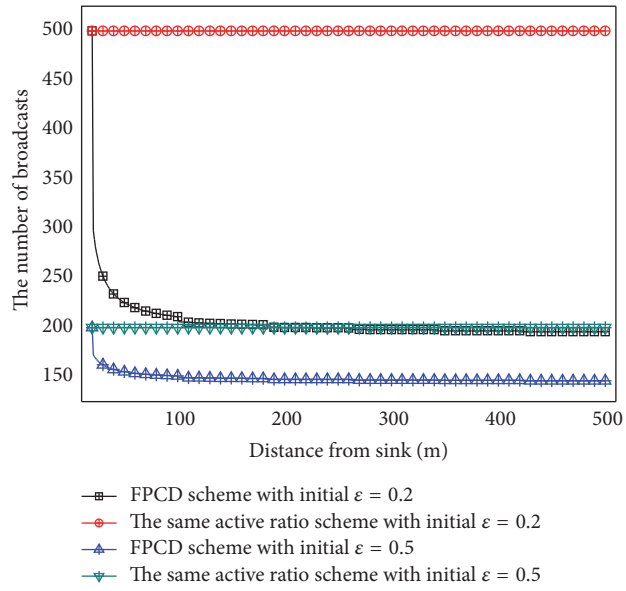


Figure 17: The broadcast times in different areas in FPCD scheme and other schemes.

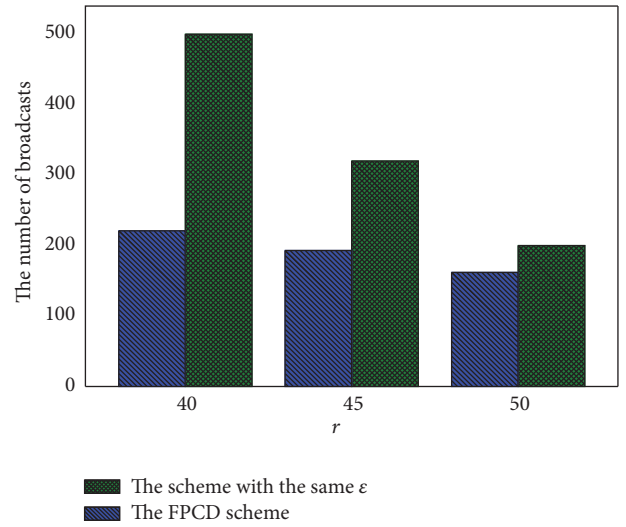


Figure 18: The broadcast times in FPCD scheme and the scheme with same ε under different r.

ratio of active node is ε. The requirement broadcast times $\mathfrak{B}$ of all the network and the number of receiving data $\mathscr{R}$ are

$$\mathfrak{B} = \frac{R^2}{r^2}\frac{1}{\varepsilon}$$
$$\mathscr{R} = \pi R^2\rho - \frac{R^2}{r^2}\frac{1}{\varepsilon}. \tag{34}$$

Proof. The number of nodes within the communication range is $n_r = \pi r^2\rho$. When one node broadcasts program codes, $\varepsilon n_r - 1$ nodes will receive the data. Because the status of node will change in the $(1/\varepsilon)\tau$ cycle, thus the number of nodes updated is εn_r each cycle. That is to say, in the $(1/\varepsilon)\tau$ cycle, the status of each node can be updated once, and there are $1/\varepsilon$ nodes to spread program codes; $\varepsilon n_r - 1$ nodes receive data packet in one broadcast.

The number of broadcast packets in the network is $n_R = \pi R^2\rho$. A total of communication cell is $m = \pi R^2\rho/\pi r^2\rho = R^2/r^2$; thus, the amount of broadcast in entire network is as follows: broadcast time is $\mathfrak{B} = m(1/\varepsilon) = (R^2/r^2)(1/\varepsilon)$; the times for receiving data are as follows:

$$\mathscr{R} = m\left(n_r - \frac{1}{\varepsilon}\right) = \frac{R^2}{r^2}\left(\pi r^2\rho - \frac{1}{\varepsilon}\right)$$
$$= \pi R^2\rho - \frac{R^2}{r^2}\frac{1}{\varepsilon}. \tag{35}$$

□

5.3. Energy Consumption and Network Lifetime Analysis. This section analyzes the energy consumption in the FPCD scheme and previous schemes. Firstly, as long as the network has energy left, the FPCD scheme can make full use of left energy to increase the number of active nodes, thereby improving the performance of program codes dissemination. Thus, energy efficiency in the FPCD scheme is 100% in theory.

Considering the energy efficiency in other schemes. From the above analysis, each node receives program codes once. If one node receives new data, the node can spread the data once. Thus, the energy consumption of program codes dissemination is balanced. But in the process of data collection, the energy consumption of network is not balanced. The energy left of network is given.

In the process of data collection, the energy consumption of node nearest the sink is

$$E_{\max} = \left((z_0 + 1) + \frac{z_0 (z_0 + 1) r}{2l_0}\right) \lambda\omega_t + \left(z_0 + \frac{z_0 (z_0 + 1) r}{2l_0}\right) \lambda\omega_r \mid z_0 = \left\lfloor \frac{R - l_0}{r} \right\rfloor. \tag{36}$$

The energy consumption of the area at l distance from the sink

$$E_l = \left((z + 1) + \frac{z (z + 1) r}{2l}\right) \lambda\omega_t + \left(z + \frac{z (z + 1) r}{2l}\right) \lambda\omega_r \mid z = \left\lfloor \frac{R - l}{r} \right\rfloor. \tag{37}$$

Thus, the energy left in the area at l distance from the sink is

$$E_l^s = E_{\max} - E_l. \tag{38}$$

The remaining energy of entire network is that considering a small arc area at the l distance from the sink, the width of the arc area is d_l; when $d_l \to 0$, the area of this arc area is $2\pi l d_l$; thus, the remaining energy of this area is $2\pi l d_l \rho E_l^s$. So the energy left of all network is as follows:

$$E_{\text{total}}^s = \int_{l_0}^{R} 2\pi l \rho E_l^s d_l = 2\pi\rho \int_{l_0}^{R} l \left(E_{\max} - E_l\right) d_l. \tag{39}$$

The available energy consumption of the network is $\pi R^2 \rho E_{\max}$.

Thus, the energy efficient is as follows:

$$\mu_e = \frac{E_{\text{total}}^s}{\pi R^2 \rho E_{\max}} = \frac{2 \int_{l_0}^{R} l \left(E_{\max} - E_l\right) d_l}{R^2 E_{\max}}. \tag{40}$$

Based on the above analysis, Figure 19 gives the energy consumption of data collection in different areas. As can be seen, the energy consumption in the area near the sink is higher than the energy consumption in the area far away from the sink; thus, the area far away from the sink has much energy left. It can be seen from Figure 20 that most areas have 80% energy left; the average energy left reaches 90%. Figure 21 shows the energy consumption in the FPCD scheme and the previous scheme. It shows that the remaining energy network can fully be used for program dissemination, the FPCD scheme has a very high energy efficiency, and it is more than 4 times that in other strategies.

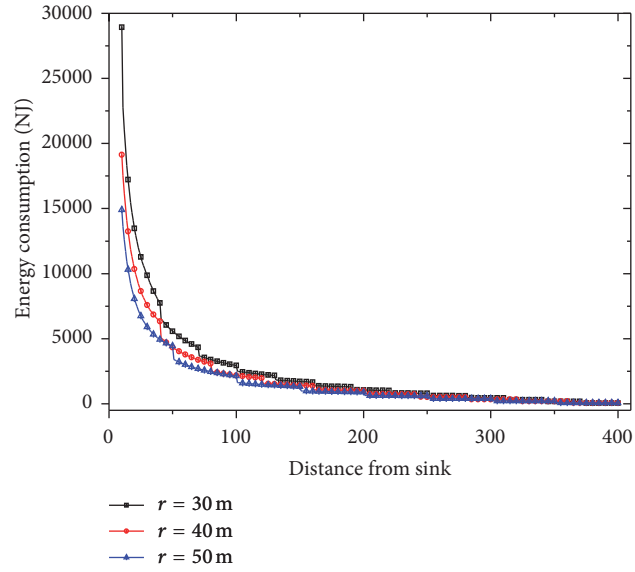


Figure 19: The energy consumption for data collection in different areas.

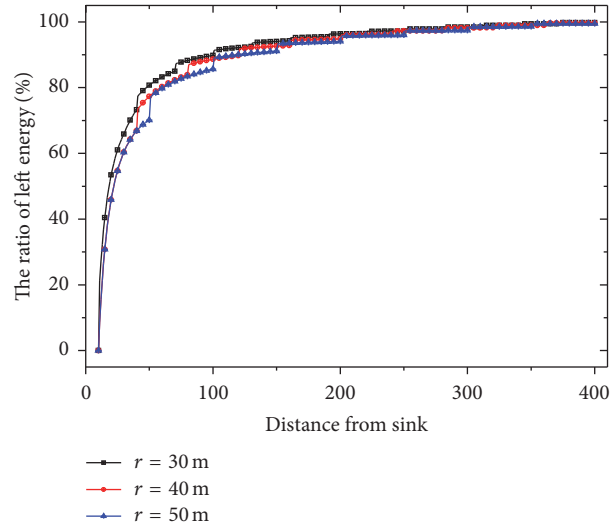


Figure 20: The ratio of energy left of node in different areas in the network to total energy of nodes.

6. Experimental Results and Analyses

OMNET++ is employed for experimental verification [35]. Without loss of generality, the network parameters are as follows: network radius R = 400 m, packet transfer radius r = 50 m, event generation probability of node in one round λ = 0.1, and 200 nodes are deployed. Other parameters settings refer to Table 1. We compare FPCD scheme against the same active ratio scheme. In the same active ratio scheme, the

Table 1: Network parameters.

Symbol	Description	Value
d_0	Threshold distance (m)	87
r_s	Sensing range (m)	15
E_{elec}	Transmitting circuit loss (nJ/bit)	50
e_{fs}	Power amplification for the free space (pJ/bit/m^2)	10
e_{amp}	Power amplification for the multipath fading (pJ/bit/m^4)	0.0013
E_{init}	Initial energy (J)	0.5

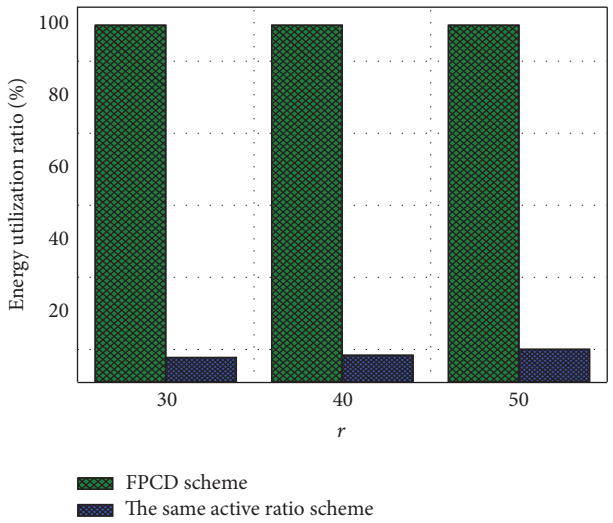


Figure 21: The energy efficiency in the FPCD scheme and other schemes.

number of active nodes in different areas is the same. That is, the number of active nodes in the area far from the sink is the same as the number of active nodes in the area near the sink. Program codes not only are transmitted to the network boundary but also are transmitted to the neighbor nodes. The aim is that all nodes in the network receive the program codes produced by the sink.

6.1. Comparative Analysis of Broadcast Times. The number of broadcast nodes is shown in Figure 22. As can be seen, the number of broadcast nodes in the FPCD scheme is higher than that in the same active ratio scheme. When the ratio of active nodes in the area near the sink is 30%, the number of broadcast nodes in those two schemes is higher than the number of broadcast nodes in the network with 20% active nodes in the area near to the sink. The ratio of the number of broadcast nodes in the FPCD scheme to the same active ratio scheme is shown in Figure 23. From Figure 23, the number of broadcast nodes in the FPCD scheme is 1.08889–2 times that in the same active ratio scheme when the ratio of active nodes in the area near the sink is 30%. The number of broadcast nodes in the FPCD scheme is 1.08824–2.5 times that in the same active ratio scheme when the ratio of active nodes in

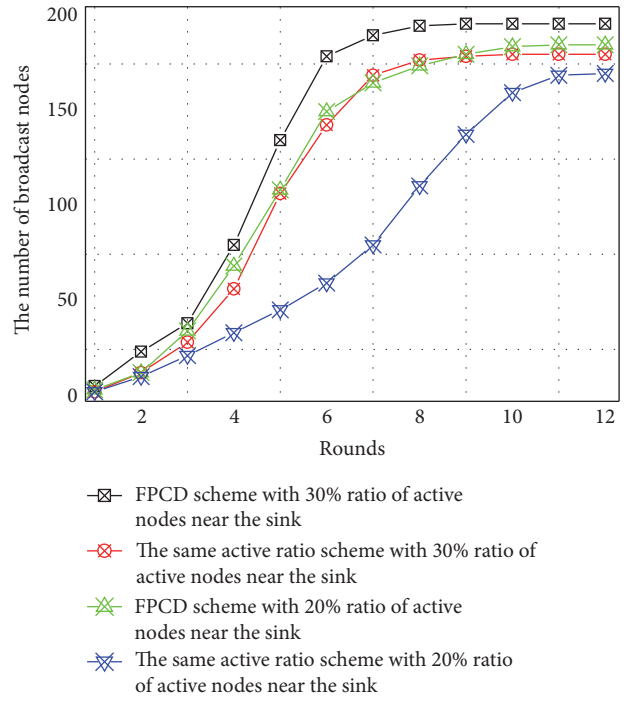


Figure 22: The number of broadcast nodes in FPCD scheme and the same active ratio scheme.

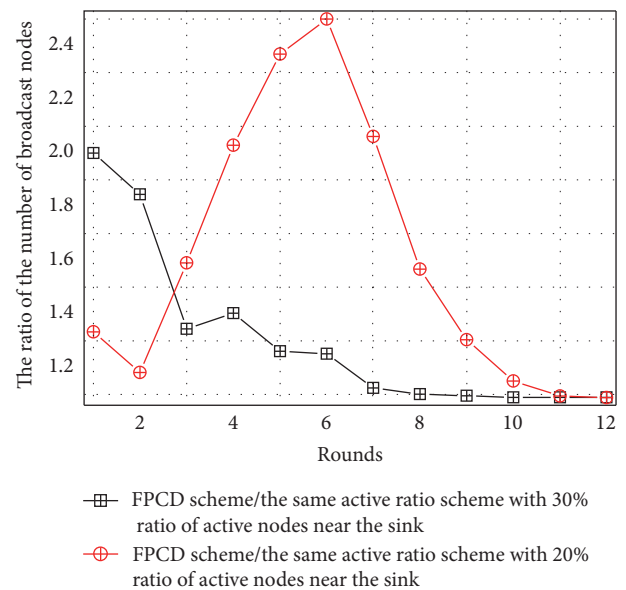


Figure 23: The ratio of number of broadcast nodes in FPCD scheme to the same active ratio scheme.

the area near to the sink is 20%. It shows that FPCD scheme has better performance.

6.2. Comparative Analysis of Broadcast Delay. The transmission delay of network under different proportion of active nodes in the area near the sink in the FPCD scheme and the same active ratio scheme is given in Figure 24. As can be seen, (1) with the increase of the ratio of active nodes

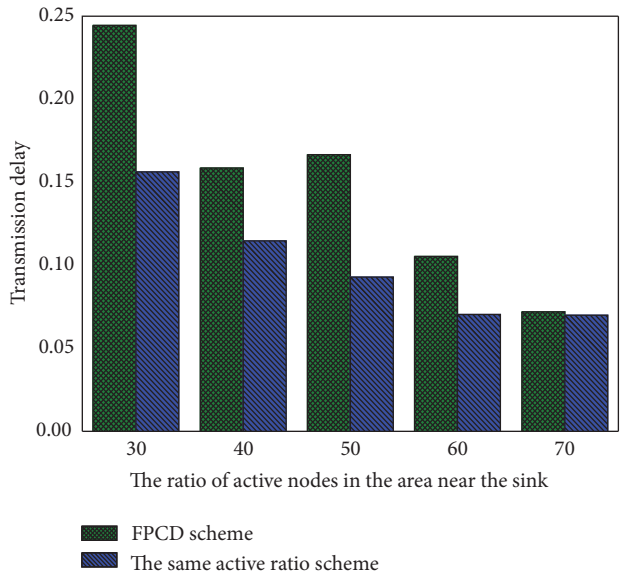


Figure 24: The transmission delay under difference ratio of active nodes in the area near the sink.

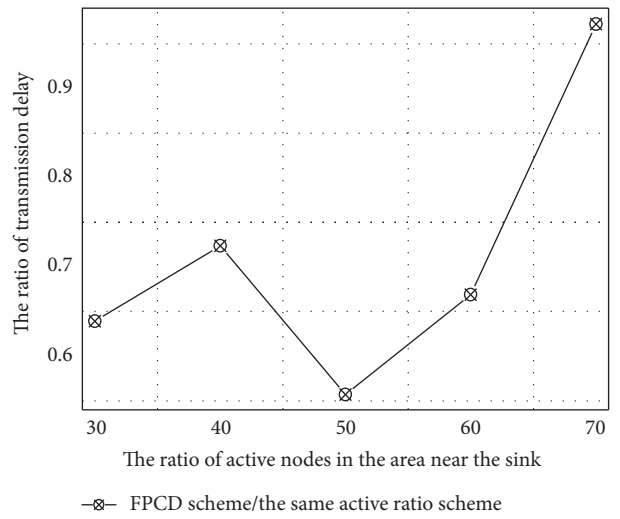


Figure 25: The ratio of transmission delay under difference ratio of active nodes in the area near the sink in FPCD scheme to PCDE scheme.

in the area near the sink, the transmission delay of network is reduced in those two schemes. The reason is that with the increase of ratio of active nodes in the area near to the sink, the energy consumption of the nodes nearest the sink is increased. Thus, there is more energy left in the areas far away from the sink, the energy left in the areas far away from the sink should be used to increase the number of active nodes while retaining the network lifetime. Thus, the program codes can be transmitted to all nodes except the nodes that never receive the program codes with a short time. (2) The transmission relay of network in the FPCD scheme is smaller than that of the same active ratio scheme.

The ratio of transmission delay of network under different ratio of active nodes in the area near the sink in the FPCD scheme to the same active ratio scheme is given in Figure 25. As can be seen, the transmission delay of network in FPCD scheme is 0.63921 times that in the same active ratio scheme when the ratio of active nodes in the area near the sink is 30%. It shows that the FPCD scheme has better performance.

The transmission delay of network under different transmission radius r in the FPCD scheme and the same active ratio scheme is given in Figure 26. It can be seen that no matter what the transmission radius r is, the transmission delay of network in the FPCD scheme is smaller than that of the same active ratio scheme. The transmission delay of network is reduced with the increase of transmission radius r. The reasons are that (1) in the FPCD scheme, the number of active nodes in the area far away from the sink is increased, when the sink spreads program codes to nodes; there are many nodes in the range of transmission radius; the program codes can be spread to all nodes quickly; thus, the program codes can be spread to all nodes except the nodes that never receive the program codes with a very short time. In the FPCD scheme, the number of active nodes in the area far away from the sink is increased. Thus, the transmission delay

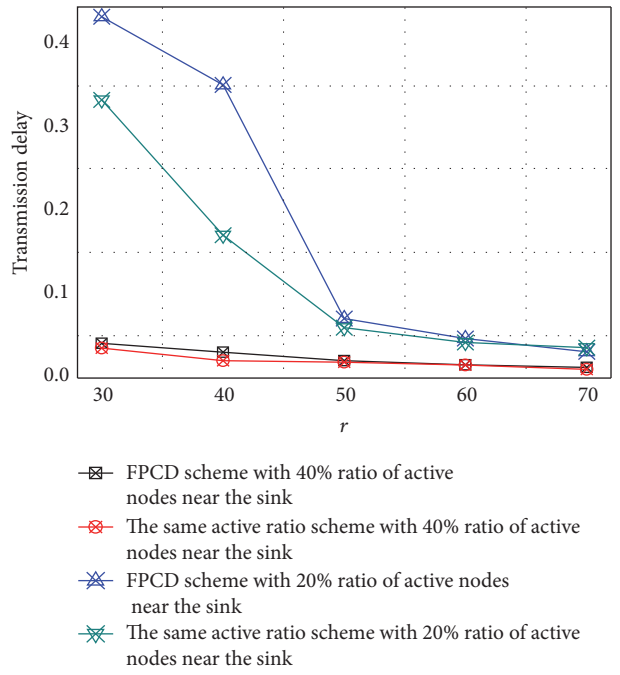


Figure 26: The transmission delay under different transmission radius r in FPCD scheme and the same active ratio scheme.

is reduced. (2) When the r increase, the hop count from the sink to nodes which are farthest from the sink is reduced. When the sink spread the program codes to the nodes, the program codes can reach the nodes far away from the sink in a small hop count, so the transmission delay is reduced.

The ratio of transmission delay of network under different transmission radius r in the FPCD scheme to the same active ratio scheme is given in Figure 27. The transmission delay in FPCD scheme is 1.1076–2.05791 times that in the same active

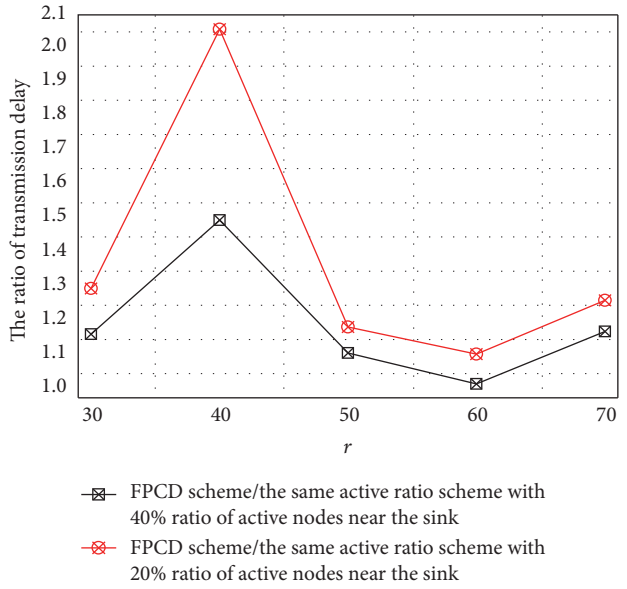


Figure 27: The ratio transmission delay under different transmission radius r in FPCD scheme to the same active ratio scheme.

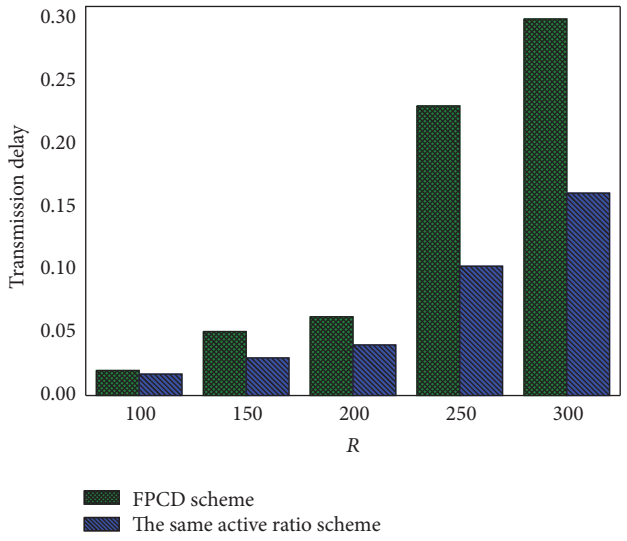


Figure 28: The transmission delay under different transmission radius R in FPCD scheme and the same active ratio scheme.

ratio scheme when the ratio of active nodes in the area near the sink is 20%.

The transmission delay of network under different network radius R in the FPCD scheme and the same active ratio scheme is shown in Figure 28. It can be seen that the transmission delay of network in the FPCD scheme is smaller than that in the same active ratio scheme. The transmission delay of network is increased with the increase of network radius R. The larger the network radius R is, the hop counts from the sink to the nodes which farthest from the sink is increased due to the fixed transmission radius. Thus, when the sink sends program codes, the nodes which are farthest

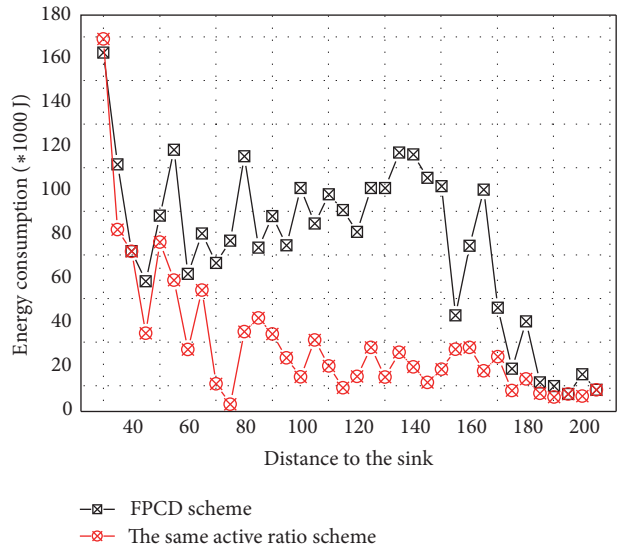


Figure 29: The energy consumption in different areas.

from the sink receive the program codes at more hop count, resulting in bigger transmission delay.

6.3. Comparative Analysis of Energy Consumption and Network Lifetime. The energy consumption of nodes in different areas in the FPCD scheme and the same active ratio scheme is given in Figure 29. As can be seen, (1) when the nodes' distance from the sink is 10, the energy consumption of nodes in the area in the FPCD scheme is lower than that in the same active ratio scheme. (2) But when the nodes' distance from the sink is bigger than 20, the energy consumption of nodes in those areas in the FPCD scheme is higher than the energy consumption of nodes in those areas in the same active ratio scheme, and the energy consumption in those areas is not bigger than the maximum energy consumption in the network, which shows that the FPCD scheme cannot affect the network lifetime. The reason is that, in the FPCD scheme, the number of active nodes in the area far away from the sink is increased. The number of receiving program codes is more, and the number of transmitting program codes is more. Thus, the energy consumption in this area in FPCD scheme is much higher than that of the same active ratio scheme.

The ratio of energy consumption of nodes in different areas in the FPCD scheme to the same active ratio scheme is shown in Figure 30. It can be seen that (1) when the nodes' distance from the sink is 30, the energy consumption of this area in FPCD scheme is 0.96263 times that of the same active ratio scheme. In the other areas, the energy consumption in the FPCD scheme is 1.00148–10.23938 times that in the same active ratio scheme. It shows that the FPCD scheme can make full use of the energy left in the area far from the sink while retaining network lifetime.

The total of energy consumption under different ratios of active nodes in the area near the sink in FPCD scheme and the same active ratio scheme is given in Figure 31. As can be seen, (1) with the increase of the ratio of active nodes

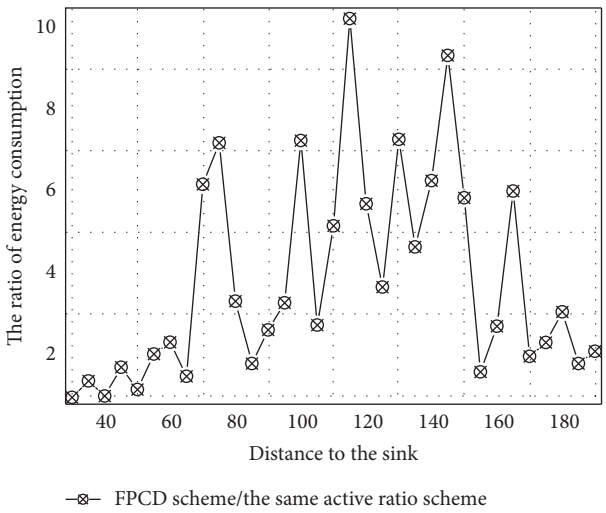


Figure 30: The ratio of energy consumption in different areas in FPCD scheme to PCDE scheme.

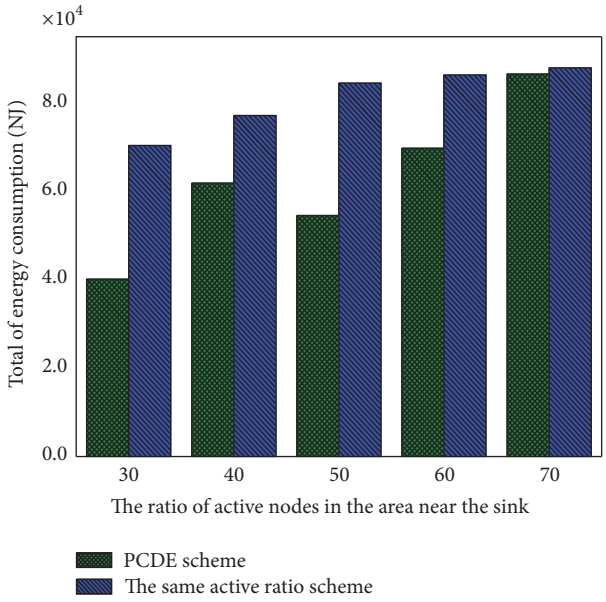


Figure 31: The total of energy consumption under different ratio of active nodes in the area near the sink in FPCD scheme and the same active ratio scheme.

in the area near the sink, the total of energy consumption is increased. (2) The total of energy consumption in the FPCD scheme is higher than that in the same active ratio scheme. In the FPCD scheme, the number of active nodes in the area far from the sink is increased. When the proportion of active nodes nearest the sink is increased, the energy consumption of nodes near the sink is higher. Thus, there is more energy left to increase the number of active nodes in the areas far away from the sink. Thus, the number of active nodes in the areas far away from the sink is more. Nodes in areas far from the sink consume more energy, and thus the total of energy consumption in the network increased.

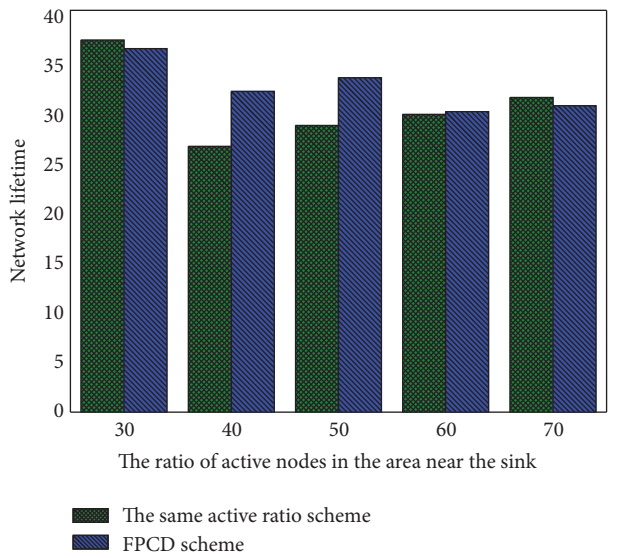


Figure 32: Network lifetime under different ratio of active nodes in the area near the sink in FPCD scheme and the same active ratio scheme.

Network lifetime under different proportion of active nodes nearest to the sink in the FPCD scheme and the same active ratio scheme is given in Figure 32. As can be seen, the network lifetime in the FPCD scheme is not lower than that in PCDE scheme. Though the total of energy consumption in the FPCD scheme is higher than that in the same active ratio scheme, the FPCD scheme cannot affect the network lifetime.

The total of energy consumption under different network radius r in the FPCD scheme and the same active ratio scheme is given in Figure 33. From Figure 33, the total of energy consumption in the FPCD scheme is higher than that of the same active ratio scheme. In the FPCD scheme, the number of active nodes in the area far from the sink is increased. When the sink spreads program codes, nodes far from the sink receive large number of program codes from the last nodes, and the energy consumption in this areas for receiving the program codes is increased. But at the same time, those nodes will transmit those program codes to many active nodes, the energy consumption for transmitting program codes is increased. In the same active ratio scheme, the number of active nodes is the same in different areas. Thus, the total of energy consumption in the FPCD scheme is higher than that in the same active ratio scheme. It can be seen from Figure 29. The energy consumption in the area near to the sink in the FPCD scheme is smaller than that in the same active ratio scheme and the energy consumption in other areas in the FPCD scheme is higher than that of the same active ratio scheme. The total of energy consumption in the FPCD scheme is higher than that of the same active ratio scheme. It shows that the FPCD scheme can improve energy efficient.

Network lifetime under different network radius r in the FPCD scheme and the same active ratio scheme is given in Figure 34. As can be seen, the network lifetime in the FPCD scheme is not lower than that in the same active ratio scheme.

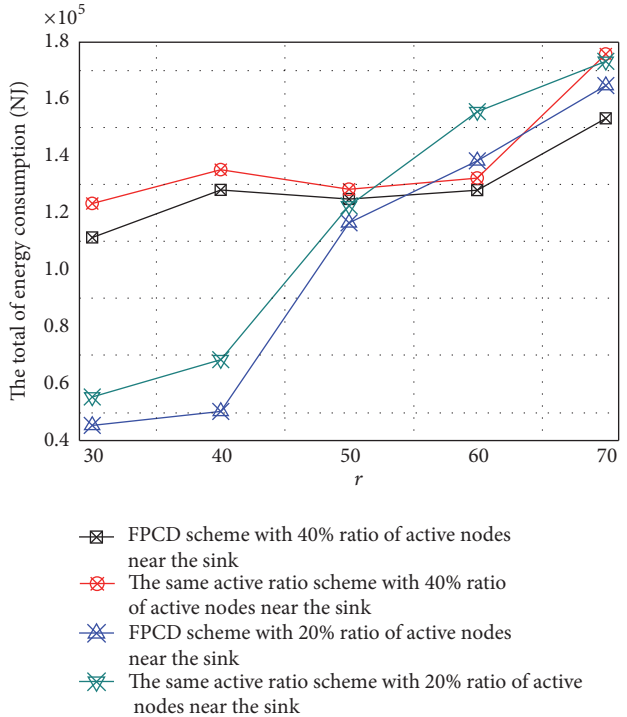


Figure 33: The total of energy consumption under different transmission radius r in FPCD scheme and the same active ratio scheme.

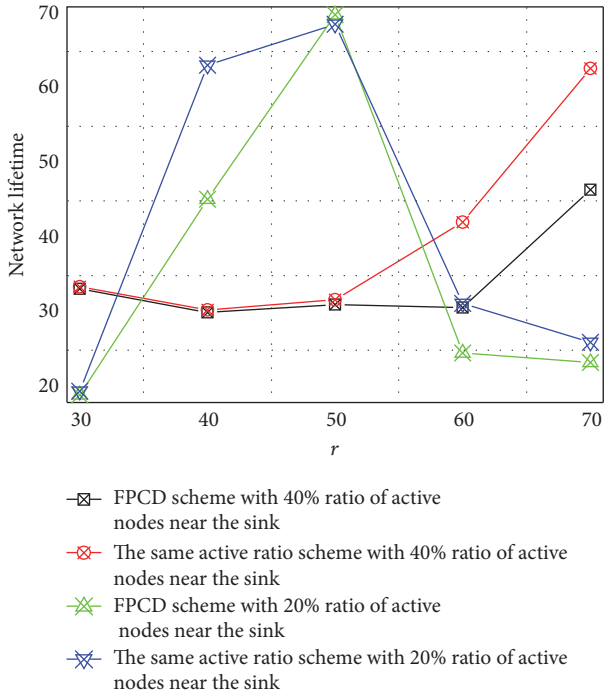


Figure 34: Network lifetime under different transmission radius r in FPCD scheme and the same active ratio scheme.

Though the number of active nodes in the area far away from the sink is increased, the energy consumption of this area is not higher than the energy consumption of nodes near to the sink. Thus, the network lifetime in the FPCD scheme is not lower than that of the same active ratio scheme.

The total of energy consumption under different network radius R in the FPCD scheme and the same active ratio scheme is given in Figure 35. From Figure 35, the total of energy consumption in the FPCD scheme is higher than that in the same active ratio scheme. With the increase of the network radius R, due to the fixed transmission radius r, the hop counts from the sink to nodes farthest from the sink are increased. The energy consumption for spreading the program codes from the sink to nodes farthest from the sink is higher. So the total of energy consumption of network in the FPCD scheme is higher than that in the same active ratio scheme.

Network lifetime under different network radius R in the FPCD scheme and the same active ratio scheme is given in Figure 36. As can be seen, network lifetime in the FPCD scheme is not lower than that of the same active ratio scheme. It shows that the FPCD scheme cannot damage the network lifetime. Tough the FPCD scheme has better performance from Figure 36, it makes full use of the energy to improve the number of active nodes far from the sink.

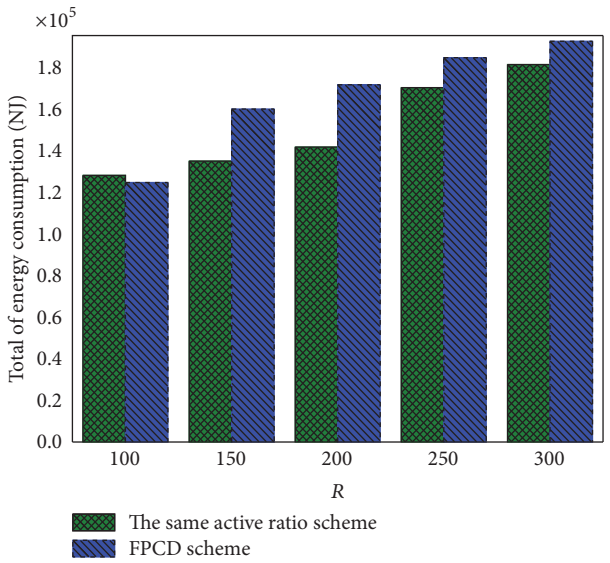


Figure 35: The total of energy consumption under different transmission radius R in FPCD scheme and the same active ratio scheme.

7. Conclusion

WSDNs have the important function for propagating program codes, but for some urgent applications, they often have very stringent requirements on program codes dissemination delay. So it is important to minimum transmission delay for spreading program codes while maximizing the network lifetime. In previous schemes, the transmission delay and network lifetime cannot be optimized at the same time. In this paper, a fast program codes dissemination (FPCD) scheme for wireless software defined networking is proposed to reduce the transmission delay while maximum network lifetime. The innovation of FPCD scheme is that the number

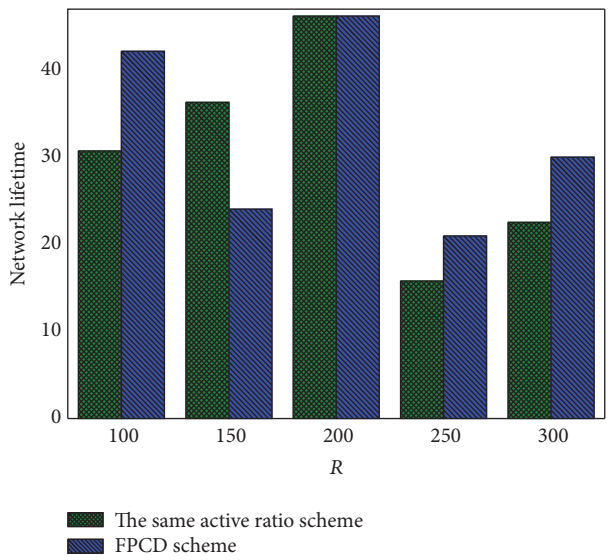


Figure 36: Network lifetime under different transmission radius R in FPCD scheme and the same active ratio scheme.

of active nodes in different areas can be adjusted dynamically according to the energy difference between the energy consumption in this area and the maximum energy consumption in the network. The energy left can be full used to increase the number of active nodes to reduce the transmission delay while maximum the network lifetime. Thus, FPCD scheme has better performances.

Competing Interests

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

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61379110, 61073104, 61370229, and 61370178), the National Basic Research Program of China (973 Program) (2014CB046305), and the S&T Projects of Guangdong Province under Grant nos. 2014B010103004, 2014B010117007, 2015A030401087, 2015B010110002, and 2016B010109008.

References

[1] X. Luo, J. Liu, D. D. Zhang, and X. Chang, "A large-scale web QoS prediction scheme for the Industrial Internet of Things based on a kernel machine learning algorithm," *Computer Networks*, vol. 101, pp. 81–89, 2016.

[2] S. He, J. Chen, X. Li, X. S. Shen, and Y. Sun, "Mobility and intruder prior information improving the barrier coverage of sparse sensor networks," *IEEE Transactions on Mobile Computing*, vol. 13, no. 6, pp. 1268–1282, 2014.

[3] W. Zhao, "Performance optimization for state machine replication based on application semantics: a review," *Journal of Systems and Software*, vol. 112, pp. 96–109, 2016.

[4] H. Liu, L. Yu, W. Wang, and F. Sun, "Extreme learning machine for time sequence classification," *Neurocomputing*, vol. 174, pp. 322–330, 2016.

[5] L. Yang, J. Cao, W. Zhu, and S. Tang, "Accurate and efficient object tracking based on passive RFID," *IEEE Transactions on Mobile Computing*, vol. 14, no. 11, pp. 2188–2200, 2015.

[6] K. Hua and W. Wang, "High security self-encoded spread spectrum watermarking using genetic algorithms," *Telecommunication Systems*, vol. 60, no. 1, pp. 143–148, 2014.

[7] A. Liu, Y. Hu, and Z. Chen, "An energy-efficient mobile target detection scheme with adjustable duty cycles in wireless sensor networks," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 22, no. 4, pp. 203–225, 2016.

[8] Y. Liu, A. Liu, Y. Hu et al., "FFSC: an energy efficiency communications approach for delay minimizing in internet of things," *IEEE Access*, vol. 4, pp. 3775–3793, 2016.

[9] H. Dai, G. Chen, C. Wang, S. Wang, X. Wu, and F. Wu, "Quality of energy provisioning for wireless power transfer," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 2, pp. 527–537, 2015.

[10] H. Li, D. Liu, Y. Dai, and T. H. Luan, "Engineering searchable encryption of mobile cloud networks: when QoE meets QoP," *IEEE Wireless Communications*, vol. 22, no. 4, pp. 74–80, 2015.

[11] R. Xie, A. Liu, and J. Gao, "A residual energy aware schedule scheme for WSNs employing adjustable awake/sleep duty cycle," *Wireless Personal Communications*, 2016.

[12] Y. Hu and A. Liu, "Improvement the quality of mobile target detection through portion of node with fully duty cycle in WSNs," *Computer Systems Science and Engineering*, vol. 31, no. 1, pp. 5–17, 2016.

[13] Y. Liu, M. Dong, K. Ota, and A. Liu, "ActiveTrust: secure and trustable routing in wireless sensor networks," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 2013–2027, 2016.

[14] X. Liu, K. Ota, A. Liu, and Z. Chen, "An incentive game based evolutionary model for crowd sensing networks," *Peer-to-Peer Networking and Applications*, vol. 9, no. 4, pp. 692–711, 2016.

[15] X. Zheng, J. Wang, W. Dong, Y. He, and Y. Liu, "Bulk data dissemination in wireless sensor networks: analysis, implications and improvement," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1428–1439, 2016.

[16] S. He, D.-H. Shin, J. Zhang, J. Chen, and Y. Sun, "Full-view area coverage in camera sensor networks: dimension reduction and near-optimal solutions," *IEEE Transactions on Vehicular Technology*, 2015.

[17] S. Guha and S. Khuller, "Approximation algorithms for connected dominating sets," *Algorithmica*, vol. 20, no. 4, pp. 374–387, 1998.

[18] H. Lim and C. Kim, "Flooding in wireless ad hoc networks," *Computer Communications*, vol. 24, no. 3-4, pp. 353–363, 2001.

[19] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, "On the construction of energy-efficient broadcast and multicast trees in wireless networks," in *Proceedings of the IEEE 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, vol. 2, pp. 585–594, Tel Aviv, Israel, March 2000.

[20] A. K. Das, R. J. Marks, M. El-Sharkawi, P. Arabshahi, and A. Gray, "r-shrink: a heuristic for improving minimum power broadcast trees in wireless networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '03)*, pp. 523–527, December 2003.

[21] H. Dai, X. Wu, L. Xu, F. Wu, S. He, and G. Chen, "Practical scheduling for stochastic event capture in energy harvesting sensor networks," *International Journal of Sensor Networks*, vol. 18, no. 1-2, pp. 85–100, 2015.

[22] M. Dong, K. Ota, L. T. Yang, A. Liu, and M. Guo, "LSCD: a low-storage clone detection protocol for cyber-physical systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 5, pp. 712–723, 2016.

[23] H. Li, X. Lin, H. Yang, X. Liang, R. Lu, and X. Shen, "EPPDR: an efficient privacy-preserving demand response scheme with adaptive key evolution in smart grid," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 8, pp. 2053–2064, 2014.

[24] X. Liu, "A novel transmission range adjustment strategy for energy hole avoiding in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 67, pp. 43–52, 2016.

[25] T. Le Duc, D. T. Le, V. V. Zalyubovskiy, D. S. Kim, and H. Choo, "Level-based approach for minimum-transmission broadcast in duty-cycled wireless sensor networks," *Pervasive and Mobile Computing*, vol. 27, pp. 116–132, 2016.

[26] D. Zhao, K.-W. Chin, and R. Raad, "Approximation algorithms for broadcasting in duty cycled wireless sensor networks," *Wireless Networks*, vol. 20, no. 8, pp. 2219–2236, 2014.

[27] M. Khiati and D. Djenouri PhD, "BOD-LEACH: broadcasting over duty-cycled radio using LEACH clustering for delay/power efficient dissimilation in wireless sensor networks," *International Journal of Communication Systems*, vol. 28, no. 2, pp. 296–308, 2015.

[28] H. Byun and J. So, "Node scheduling control inspired by epidemic theory for data dissemination in wireless sensor-actuator networks with delay constraints," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1794–1807, 2016.

[29] S. He, J. Chen, F. Jiang, D. K. Y. Yau, G. Xing, and Y. Sun, "Energy provisioning in wireless rechargeable sensor networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 10, pp. 1931–1942, 2013.

[30] X. Luo, D. Zhang, L. T. Yang, J. Liu, X. Chang, and H. Ning, "A kernel machine-based secure data sensing and fusion scheme in wireless sensor networks for the cyber-physical systems," *Future Generation Computer Systems*, vol. 61, pp. 85–96, 2016.

[31] X. Liu, "A deployment strategy for multiple types of requirements in wireless sensor networks," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2364–2376, 2015.

[32] A. Liu, X. Liu, and Y. Liu, "A comprehensive analysis for fair probability marking based traceback approach in WSNs," *Security and Communication Networks*, vol. 9, no. 14, pp. 2448–2475, 2016.

[33] H. Li, Y. Yang, T. H. Luan, X. Liang, L. Zhou, and X. S. Shen, "Enabling fine-grained multi-keyword search supporting classified sub-dictionaries over encrypted cloud data," *IEEE Transactions on Dependable and Secure Computing*, vol. 13, no. 3, pp. 312–325, 2016.

[34] M. Malawski, K. Figiela, M. Bubak, E. Deelman, and J. Nabrzyski, "Scheduling multilevel deadline-constrained scientific workflows on clouds based on cost optimization," *Scientific Programming*, vol. 2015, Article ID 680271, 13 pages, 2015.

[35] OMNet++ Network Simulation Framework, http://www.omnetpp.org/.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 5796907, 10 pages
http://dx.doi.org/10.1155/2016/5796907

Research Article

Infinite Queue Management via Cascade Control for Industrial Routers in Smart Grid IP Networks

Ku-Hwan Kim,[1] Hoang-Linh To,[2] Won-Joo Hwang,[3] and Jung-Tae Lee[1]

[1]*Department of Electrical and Computer Engineering, Pusan National University, Pusan 46241, Republic of Korea*
[2]*Department of Information and Communication System, Inje University, Gyeongnam 50834, Republic of Korea*
[3]*Department of Information and Communications Engineering, HSV-TRC, Inje University, Gyeongnam 50834, Republic of Korea*

Correspondence should be addressed to Won-Joo Hwang; ichwang@inje.ac.kr

Received 28 April 2016; Accepted 17 August 2016

Academic Editor: Kun Hua

Smart grid applications experience an extremely wide range of communication delay. Data flows of those applications are normally aggregated at industrial network routers in substations, form infinite (long) queues termed bufferbloat issue, and might damage the operation of transmission control protocol. The default queue management scheme, DropTail, in such routers just drops packets if queue is full while the others in literature are mostly based on one-loop feedback control where an optimal point of performance between queue length and drop rate is limited. In this paper, we study the problem of managing a long queue of industrial router at substation under heterogeneous smart grid networks. Specifically, we propose an enqueue-dequeue dropping cascade control using a two-loop design method to control both window size and queue length. Moreover, our proposal can be easily implemented into router firmware with provided discrete expressions. Finally, our simulation results are presented to validate the possible benefits that can be gained from cascade control and compare the existing queue management methods as well.

1. Introduction

1.1. Overview. Smart grid is going to be a geographically widespread system soon, which integrated many diverse real-time applications (e.g., sensor) together with energy management applications. Smart grid, which is a digitally-enhanced version of the traditional electric grid, provides infrastructure to support diverse services such as finance, information, and electrical delivery among consumers, assets, and those users who have authorized access. A conceptual model for smart grid communication is presented in [1]. At the customer house, energy consumption information is monitored and then sent to service provider for further analysis and better support. At the operator side, data is collected to make decision according to demand response or energy saving.

With an emerging smart grid, various applications can be enabled to improve quality of service and consumer satisfaction. Some popular examples are building automation, automated meter reading, outage and restoration management, and electric vehicles. Most of the smart grid applications have a strict latency requirement in the range of 100 milliseconds to 5 seconds [2]. In [3], the delay ranges of different smart grid applications were approximately reported. For example, teleprotection applications allow traffic delay from 8 to 10 ms while they allow traffic delay for more than one second for interval measurement from smart meters. To respect it, we also need a fast communication infrastructure that can handle a huge amount of exchanging data and is able to provide a near real-time response. Latency is defined as the time interval between when the state occurred and when it was acted upon by a smart grid application [2]. Various applications own different latency requirements that depend on the kind of system response it is dealing with. In such scenario, network equipment including switches, Internet wired/wireless routers, and operating systems (O/S) are the fundamental components to provide a fast and reliable smart grid communication. Table 1 lists latency and bandwidth requirements for some specific applications in smart grid networks [4].

In such a heterogeneous environment with multiple types of applications, special equipment like industrial routers

Table 1: A wide range of delays for some smart grid applications.

Application	Latency	Bandwidth
Metering	0–15 min	10–100 kbps
Information exchange	5–30 s	14–100 kbps
Electric transportation	2 s–5 min	100 kbps

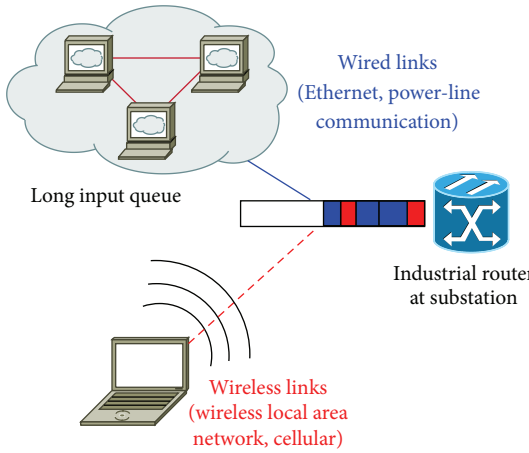


Figure 1: Long input queue at industrial router at substation.

at substations are necessary to handle. Commercially, we have seen an example product from the company, Virtual Access, that is, GW2027, which is an industrial router with applications like machine-to-machine (M2M), telemetry, supervisory control and data acquisition (SCADA), roadside, and wireless. Figure 1 illustrates such smart grid network situation in detail, where industrial router has to process data from both wired and wireless links. Among several great software features like remote configuration and fault reporting and so forth we, however, realize that a scheme to manage the input queue of the industrial router is not specified in the data sheet [5] and is usually Drop Tail, a simple scheme which drops packets in case of full buffer and allows packets otherwise.

1.2. Motivation. In general, each part of every network equipment or substation industrial router is usually preinstalled or configured with some amount of buffering, whether it is enough or not, to handle bursts of arriving packets and departing packets to the next link. It is important to ensure good utilization of the network link, especially in cases where arriving rate is greater than departing rate causing bottleneck point to be built up. The buffer then absorbs high-rate traffic packets which wait and are later served on the slower outgoing links [7]. Not enough buffering results in high dropping rate of most of the packets and low egress link utilization. For example, if a user transfers files using transmission control protocol (TCP), the user satisfaction is measured by how quickly the file transfer can complete, which is directly related to how effectively the protocol can utilize network links, that is, more buffering is better. Moreover, as the cost of buffers/memory keeps reducing, it is foreseen that large buffering would be put into every piece of network equipment and even preconfigured in operating systems. However, large buffering results in long latency experience for users as well. Besides, most of TCP schemes (Reno, New Reno, CUBIC, etc.) use loss-based congestion control. It means that a loss-based TCP sender continues increasing its congestion window size (i.e., sending rate) until a loss signal is detected by time-out or incorrect acknowledgment packets. In [8], it is suggested that "the widely used rule of thumb leads to buffers which are much larger than they need to be" and could be reduced by scarifying a small amount of bandwidth utilization. Using too much buffering, it obviously takes longer time for loss detection and therefore TCP may function incorrectly. The excessive long latency and the damage of TCP due to too much buffering are two major consequences of a problem recently termed as *bufferbloat*.

The up-to-date solutions to the bufferbloat issue, or debloat, consist of traffic shaping, TCP window size modification, and active queue management (AQM). In the scope of this paper, we concentrate on AQM algorithms because they can manage queuing delay efficiently, by either a direct or an indirect way. Most of Linux kernels and some open source routers (CeroWRT) have been integrated with some modern AQM algorithms [9, 10]. For example, depending on where the bottlenecks are, a Linux user can change the computer to use a different AQM qdisc (queuing discipline) via a `sysctl` or the `tc` (traffic control) command. Literature has witnessed several coexisting AQM candidate algorithms including Controlled Delay (CoDel) [11], Stochastic Flow Queuing CoDel (sfqCoDel) [12], and Proportional Integral Enhanced (PIE) [13] that have been proposed using varied theory tools (e.g., optimization, queuing, and control theory). Among them, the control theory approach owns some advantages to be implemented in smart grid network equipments, for example, parameterization controllers and well reference tracking. Theoretically, control-based AQM algorithms exploit the additive increase multiplicative decrease (AIMD) model of TCP and the continuous fluid-flow queue approximation [14, 15]. Then the specific controller's parameters are designed according to a closed-loop transfer function of the whole system. We, however, realize that these models almost design one-loop control for queue length only which has some problems of large overshoots and unacceptable lags (delay) in mixed-traffic scenarios. More loops operating in different network layers should improve the whole system performance, especially the input delay transient (overshoot) behaviour of the aforementioned debloat schemes.

1.3. Our Contributions. In this paper, we develop a novel active queue management using cascade control framework in control theory that is able to reduce large overshoot and therefore improve delay transient behaviour. To the best of our knowledge, our work is the first attempt to adapt cascade control method to bufferbloat research field. Using the well-known AIMD and fluid queue model [14], we decompose transfer functions into an inner and an outer loop. The inner one adjusts window size based on changing of traffic and

feedback window size and the outer one mission adjusts queue length based on feedback queue length value at time instant. Each loop's parameters are tuned using optimal gain and phase margin method [16]. This scheme operates at both the network transport layer (adjusting TCP congestion window size) and the network link layer (adjusting queue length). A difficulty when considering this cascade design is the interaction in time scale between two loops. The other side effect is complexity due to two additional controllers. Nevertheless, we show that better delay performance results can be achieved with EDDC via both numerical and simulation results (Section 5.2). Our primary contributions are summarized as follows:

(i) We develop a cascade control active queue management (EDDC) scheme by decomposing the TCP/AQM fluid-flow model into two cascaded loops for two independent controllers.

(ii) By applying the reliable optimal gain and phase margin tuning method, we derive a set of parameters for these controllers based on a given example from single-loop AQM design.

(iii) Our simulations are conducted in network simulator ns-2 using different smart grid application models, including file data transmission (FTP), voice over IP (VoIP), and video conference, which are close to the realistic smart grid applications. In most of cases, the related-delay results using EDDC are well controlled around the target delay (10 ms) while the large overshoot phenomenon disappears.

2. Related Works

In this section, we discuss the features of the famous existing AQM schemes: one mature (RED) and two modern AQM algorithms (CoDel, sfqCoDel) which are designed specifically for bufferbloat mitigation.

Random Early Detection (RED) is one of the first viable AQM algorithms [17]. However, it is proved to be so difficult to configure properly that hardly anybody uses it, even though many carrier grade routers implement it. With an appropriate set of parameters, RED is an effective algorithm. However, dynamically predicting this set of parameters was found to be difficult. As a result, RED has not been enabled by default, and its present use in the Internet is limited [18]. Other AQM algorithms have been developed since RFC2309 was published, some of which are self-tuning within a range of applicability. Hence, while this memo continues to recommend to deploy AQM, it no longer recommends that RED or any other specific algorithm is used as a default; instead it provides recommendation on how to select appropriate algorithms and that a recommended algorithm is able to automate any required tuning for common deployment scenarios.

"Controlled Delay" (CoDel) AQM algorithm [11] tries to address the problems that RED could not. First, the input signal into the algorithm (sojourn time versus average queue length) is of a different quality; second, CoDel (in its plain form) does drop/mark on dequeue (departure queue), while RED drops/marks on enqueue (arrival queue). Therefore TCP congestion control loop using CoDel responds much quicker than using RED; thus the reaction by the sender will probably be timely and relevant for that congestion epoch. With RED, the congestion signal (lost packet) has to traverse the filled-up buffer first; thus, the control loop time is much larger (includes the instantaneous queue length of the buffer) and is further delayed by the averaging going on. By design, CoDel does not need to be tuned specifically for one particular drain rate (bandwidth) of the queue unlike RED; so it adjusts much better to variable bandwidth MACs (e.g., Wi-Fi and DOCSIS link).

CoDel itself, however, drops packet of a group during each interval without considering the packet's priority level or user side. Recently, Stochastic Flow Queue CoDel (sfqCoDel) [12] is a promising AQM algorithm which demonstrates a satisfied performance to mitigate bufferbloat for users, for example, increase 10x speed of a network under load [9], and it has been implemented in open source routers (i.e., OpenWRT [10]). The sfqCoDel is a hybrid of deficit round-robin scheduling (DRR) and CoDel. It is renamed from "Fair Queuing" to "Flow Queuing" because flows that build a queue are treated differently than flows that do not. It stochastically classifies incoming packets into different queues; each queue is managed by the CoDel AQM algorithm. Packet ordering within a queue is preserved since queues have FIFO ordering [19].

3. System Model

3.1. TCP State-Space Model. We exploit the TCP fluid-flow model described by nonlinear differential equations that has been extensively studied in network routers interacting with TCP sources (e.g., [14, 20, 21]). This model captures the additive increase multiplicative decrease (AIMD) feature from TCP [22], without slow start and time-out mechanisms. However, this lack only affects initial start-up of the TCP system. Once the system reaches the stable point, the differential equations solver is able to track changes in the network well [23].

$$\begin{aligned}\dot{w}(t) &= f(w, w_R, q, p)\\ &= \frac{1}{R(t)} - \frac{w(t)}{2}\frac{w(t-R(t))}{R(t-R(t))}p(t-R(t));\end{aligned} \tag{1}$$

$$\dot{q}(t) = g(w, q) = N\frac{w(t)}{R(t)} - C_l, \tag{2}$$

where w is average TCP window size (packets); q is queue length at cable modems (packets); $p(\cdot)$ is packet dropping probability function ($0 \le p \le 1$); C_l is transmission capacity of link l (packets/sec); R is round-trip time (sec); $R(t) = T_p + q(t)/C_l$ with T_p being propagation delay; and N is number of TCP sessions.

The first equation (1) describes the TCP-Reno-sender behaviour based on AIMD while the second one (2) models a fluid queue, which allows traffic arrivals to be continuous rather than discrete, as in a classic queuing model like M/M/1

queue. The operating point (w_0, q_0, p_0) of model (1) can be derived at with $\dot{w} = 0$ and $\dot{q} = 0$ as follows

$$\begin{aligned} w_0^2 p_0 &= 2, \\ w_0 &= \frac{R_0 C_l}{N}, \\ R_0 &= T_p + \frac{q_0}{C_l}. \end{aligned} \tag{3}$$

By doing linearization around operating points (w_0, q_0, p_0) [14], with $\delta w \doteq w - w_0$; $\delta q \doteq q - q_0$; and $\delta p \doteq p - p_0$, a linearized small signal model is obtained as:

$$\begin{aligned} \delta \dot{w}(t) &= \frac{\partial f}{\partial w}\delta w + \frac{\partial f}{\partial w_R}\delta w_R + \frac{\partial f}{\partial q}\delta q + \frac{\partial f}{\partial p}\delta p, \\ \delta \dot{q}(t) &= \frac{\partial g}{\partial w}\delta w + \frac{\partial g}{\partial q}\delta q. \end{aligned} \tag{4}$$

By defining system state vectors as $x(t) = \left[\delta w(t) \quad \delta q(t)\right]^T$; control input vector as $u(t) = \delta p(t)$; external disturbance as $\omega(t)$; and system output vector as $y(t) = \delta q(t)$, we then rewrite the TCP system as a continuous time-invariant type using state-space modeling approach.

$$\begin{aligned} \dot{x}(t) &= Ax(t) + A_d x\left(t - R_0\right) + Bu\left(t - R_0\right) + G\omega(t), \\ y(t) &= Cx(t), \end{aligned} \tag{5}$$

where matrices A, A_d, B, C, and G represent state (system), input, output, and disturbance matrix, respectively, as follows:

$$\begin{aligned} A &= \begin{bmatrix} \frac{-N}{R_0^2 C_l} & 0 \\ \frac{N}{R_0} & \frac{-1}{R_0} \end{bmatrix}, \\ A_d &= \begin{bmatrix} \frac{-N}{R_0^2 C_l} & 0 \\ 0 & 0 \end{bmatrix}; \\ B &= \begin{bmatrix} \frac{-R_0 C_l^2}{2N^2} \\ 0 \end{bmatrix}, \\ C &= \begin{bmatrix} 0 & 1 \end{bmatrix} \\ G &= \begin{bmatrix} 0 & 0 \end{bmatrix}. \end{aligned} \tag{6}$$

The final system model equations which map the input (traffic flow) to the output (queue length) are

$$\begin{aligned} P(s) &= P_{\text{TCPwin}}(s) \cdot P_{\text{queue}}(s) \cdot e^{-sR_0} \\ &= \left[\frac{w(s)}{p(s)}\right] \cdot \left[\frac{q(s)}{w(s)}\right] = \left[\frac{A}{s+B}\right] \cdot \left[\frac{C}{s+D}\right] e^{-sR_0}. \end{aligned} \tag{7}$$

Using the above state-space TCP/IP system model, there are several approaches to design an efficient controller with the following requirements:

(i) Queue length is as short as possible.

(ii) Sensitivity to network parameters is low.

(iii) Link utilization is high, which means we do not waste the cost of link.

(iv) Packet drop rate is low.

We, however, admit a trade-off to satisfy all of those aforementioned requirements. For example, a shorter queue length comes along with a higher packet drop rate. Our algorithm in the next section seeks for an optimal point.

3.2. Optimization Formulation of Infinite Buffer (Bufferbloat). The classic M/M/1 queuing model well represents the infinite buffer or bufferbloat phenomenon of smart grid substation routers. Optimizing the performance of even simple queue like M/M/1 is a difficult problem because of nonlinearity of objective functions and constraints and running time scales exponentially with the problem size. In [24], a convex optimization problem was proposed for M/M/1 queues and with purpose of minimizing the state probability p_k only. However, minimizing queue length is the main purpose when dealing with the bufferbloat issue. Therefore, we formulate our optimization problem to directly optimizing queue length of infinite buffer, with constraints following the aforementioned TCP state-space model.

$$\begin{aligned} \underset{w,q}{\text{minimize}} \quad & \dot{q}(t) = g(w,q) = N\frac{w(t)}{R} - C_l \\ \text{subject to} \quad & \dot{w}(t) = f(w,q,p) \\ & 0 < p(t) < 1. \end{aligned} \tag{8}$$

The optimization variables are w and q, and the constant parameters are N, C_l, and R.

4. Proposed Algorithm

In this section, we propose an algorithm, named an Enqueue-Dequeue Dropping Cascade control scheme (EDDC). We present two main design steps of our proposed algorithm from the above TCP state-space system model. The first step is to design in the continuous time domain. The second step is transform dropping probability formula into discrete domain so that we are able to implement it into the smart grid industrial router firmware (Figure 2).

4.1. EDDC in Continuous Time Domain. We design two controllers for each loop. The inner controls drop probability p_2 based on traffic information and the outer controls drop probability p_1 based on difference between measured average queue length and reference queue length.

(i) Inner Loop. An important design requirement is that the inner loop controller should behave quickly. From (7), the inner control objective is a linear first-order type:

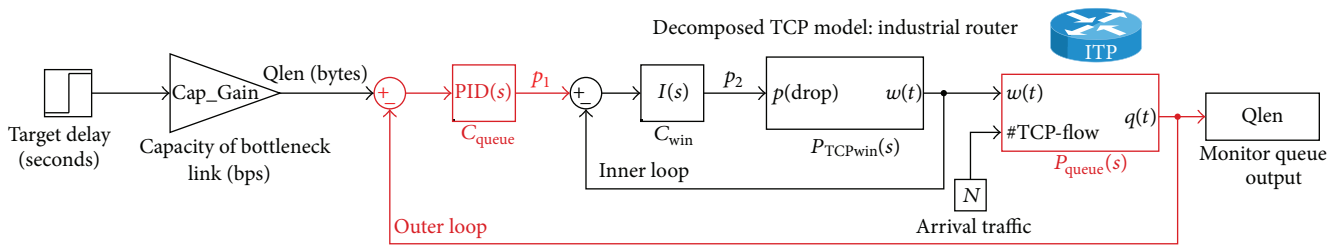


FIGURE 2: Proposed control framework for queue management in industrial router.

$P_{\text{TCPwin}}(s) = k_I/(1 + Ts)$, where $k_I = A/B$ and $T = 1/B$. Hence, we design an integral I controller for inner loop:

$$C_{\text{win}}(s) = \frac{1}{T_{C_{\text{win}}} s}. \tag{9}$$

The close-loop transfer function of the inner loop is

$$\begin{aligned} I(s) &= \frac{w(s)}{p_1(s)} = \frac{P_{\text{TCPwin}}(s) C_{\text{win}}(s)}{1 + P_{\text{TCPwin}}(s) C_{\text{win}}(s)} \\ &= \frac{k_I}{T_{C_{\text{win}}} s(1 + Ts) + k_I}. \end{aligned} \tag{10}$$

The $I(s)$ transfer function is converted into frequency domain, with ω as frequency:

$$\begin{aligned} |I(j\omega)| &= \frac{k_I}{\sqrt{\left(k_I - T_{C_{\text{win}}} T \cdot \omega^2\right)^2 + \left(\omega \cdot T_{C_{\text{win}}}\right)^2}} \Longleftrightarrow \\ I(j\omega)^2 &= \frac{k_I^2}{k_I^2 + \left(T_{C_{\text{win}}}^2 - 2k_I \cdot T_{C_{\text{win}}} T\right)\omega^2 + T_{C_{\text{win}}}^2 \cdot T^2 \omega^4}. \end{aligned} \tag{11}$$

One of quality requirements of closed-loop control system, which is represented by $I(s)$, is that the output is same to the input signal or the controller C_{win} should bring $|I(j\omega)| = 1$, $\forall\omega$, which can be called optimal gain and phase margin tuning method [16]. However, due to several reasons of real system, that requirement is rarely satisfied for all frequencies ω. An acceptable design is that $|I(j\omega)| \approx 1$, in a wide band of low frequencies ω. Hence, we propose to choose $T_{C_{\text{win}}}$ such that $T_{C_{\text{win}}}^2 - 2k_I \cdot T_{C_{\text{win}}} T = 0$ or $T_{C_{\text{win}}} = 2k_I T$. This close-form expression of $T_{C_{\text{win}}}$ is used to make decision for controller C_{win}.

(ii) Outer Loop. The close-loop transfer function of the outer loop is

$$O(s) = \frac{q(s)}{q_{\text{ref}}} = \frac{I(s) P_{\text{queue}}(s) C_{\text{queue}}(s)}{1 + I(s) P_{\text{queue}}(s) C_{\text{queue}}(s)}. \tag{12}$$

The outer control objective into zero-pole form is

$$\begin{aligned} I(s) P_{\text{queue}}(s) &= \frac{k_I}{T_{C_{\text{win}}} s(1 + Ts) + k_I} \cdot \frac{C}{s + D} \\ &= \frac{k_O}{\left(1 + T_{1_O} s\right)\left(1 + T_{2_O} s\right)\left(1 + T_{3_O} s\right)}, \end{aligned} \tag{13}$$

where $k_O = C/D$; $T_{1_O} T_{2_O} = TT_{C_{\text{win}}}/k_I$; $T_{1_O} + T_{2_O} = T_{C_{\text{win}}}/k_I$; and $T_{3_O} = 1/D$.

For the outer loop, the objective function is linear third-order type, due to inclusion of $I(s)$. Hence, we choose proportional-integral-derivative (PID) controller by using the same method at the inner loop design, or $|O(j\omega)| \approx 1$.

$$C_{\text{queue}}(s) = k_{p_O}\left(1 + \frac{1}{T_{i_O} s} + T_{d_O} s\right), \tag{14}$$

with $k_{p_O} = (T_{1_O} + T_{2_O})/2k_O T_{3_O}$, $T_{i_O} = T_{1_O} + T_{2_O}$, and $T_{d_O} = (T_{1_O} \cdot T_{2_O})/(T_{1_O} + T_{2_O})$.

4.2. EDDC in Discrete Time Domain

(i) Discrete Outer Loop Controller

$$\frac{p_1(s)}{e_1(s)} = C_{\text{queue}}(s) = k_{p_O}\left(1 + \frac{1}{T_{i_O} s} + T_{d_O} s\right), \tag{15}$$

where $p_1(s)$ is the output of the outer loop controller and $e_1(s)$ is the difference between measured queue length and desired queue length at the outer loop. We apply backward difference method to get the z-transform of the outer loop controller $C_{\text{queue}}(s)$:

$$\begin{aligned} p_1(t) &= e_1(t) k_{p_O}\left[1 + \frac{T_s}{T_{i_O}\left(1 - z^{-1}\right)} + T_{d_O}\frac{\left(1 - z^{-1}\right)}{T_s}\right]. \end{aligned} \tag{16}$$

Transforming again to time domain, we obtain

$$\begin{aligned} p_1(t) &= p_1(t-1) + k_{p_O}\left[e_1(t) - e_1(t-1)\right] \\ &\quad + \frac{k_{p_O} T_s}{T_{i_O}} e_1(t) \\ &\quad + \frac{k_{p_O} T_{d_O}}{T_s}\left[e_1(t) - 2e_1(t-1) + e_1(t-2)\right], \end{aligned} \tag{17}$$

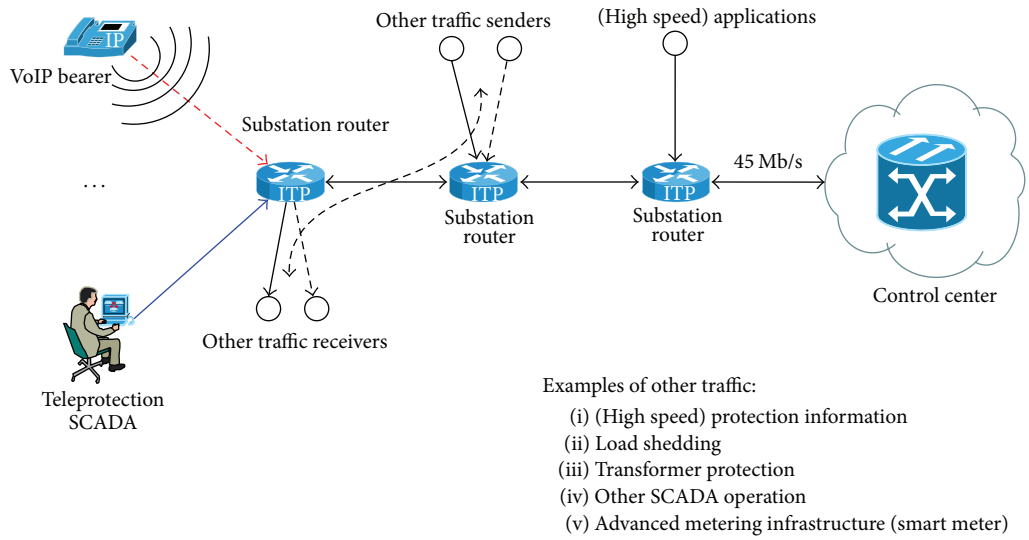


Figure 3: Cross-traffic simulation topology for smart grid applications.

or a compact form of the outer loop controller C_{queue}

$$p_1(t) = p_1(t-1) + a_O \cdot e_1(t) - b_O \cdot e_1(t-1) + c_O \cdot e_1(t-2). \quad (18)$$

(ii) Discrete Inner Loop Controller

$$\frac{p_2(s)}{e_2(s)} = C_{\text{win}} = \frac{1}{T_{C_{\text{win}}} s}, \quad (19)$$

where $p_2(s)$ is the output of the inner loop controller and $e_2(s)$ is the difference between p_1 and congestion window size $w(s)$. Doing similarly as discrete outer loop steps above, we obtain a compact form of the inner loop controller C_{win} in discrete domain as follows, with $a_I = T_s/T_{C_{\text{win}}}$:

$$p_2(t) = p_2(t-1) + a_I \cdot e_2(t-1). \quad (20)$$

Table 2: Simulation configuration.

(a)

Parameter	Value
Simulation time	600 sec
FTP packet size	(Bulk transmission) 500 bytes
VoIP (G.711)	On: 1 sec; idle: 1.35 sec
Video (MyEvalVid)	Sample foreman; packet size: 1024 bytes

(b)

AQMs	Configuration
RED	Max average queue size threshold: 20 kbytes
FB-OCQ [6]	Target queue length: 25 kbytes
PI	Target queue length: 25 kbytes
CoDel	Target delay: 10 msec
sfqCoDel	Target delay: 10 msec; 32 subflows
EDDC	Target delay: 10 msec; a_O, b_O, c_O, and a_I

5. Performance Evaluation

5.1. Simulation Setup. NS2 is a discrete event network simulator, which simulates packet-level events and facilitates the development of communication network scenarios considering the new protocols involved, either wired or wireless technologies. We use and modify the NS2 TCP evaluation tool originally developed by Wang et al. [25] for our AQM algorithm evaluation purpose. We also develop our own EDDC source code files based on the basic proportional integral (PI) algorithm code from NS2 source code tree. The simulation network topology as multiple cross bottlenecks is our choice because of its realistic smart grid IP networks (Figure 3). The simulation configuration is presented in Table 2.

5.2. Simulation Results. We evaluate our proposed algorithm performance according to three main characteristics of the bottleneck link: bandwidth (Mbps), propagation round-trip time (ms), and packet error rate ratio. Each experiment is compared between AQM algorithms (in Table 2) with respect to mean queue length, link utilization, and packet drop rate. We expect that mean queue length is as low as possible, link utilization is about 80 $\rightarrow$ 95%, and packet error rate should be low. These requirements of three criteria, however, cannot be achieved at the same time due to trade-off. Therefore, our algorithm searches for a balance point among three of them.

5.2.1. Experiment 1: Varied Link Bandwidth. In the first experiment, we vary the bottleneck link bandwidth from 1 Mbps to 1000 Mbps. This variation represents different types of links in heterogeneous smart grid networks. For example, a wired link such as Ethernet has a bandwidth of 100 Mbps while a wireless link from wireless adapter has

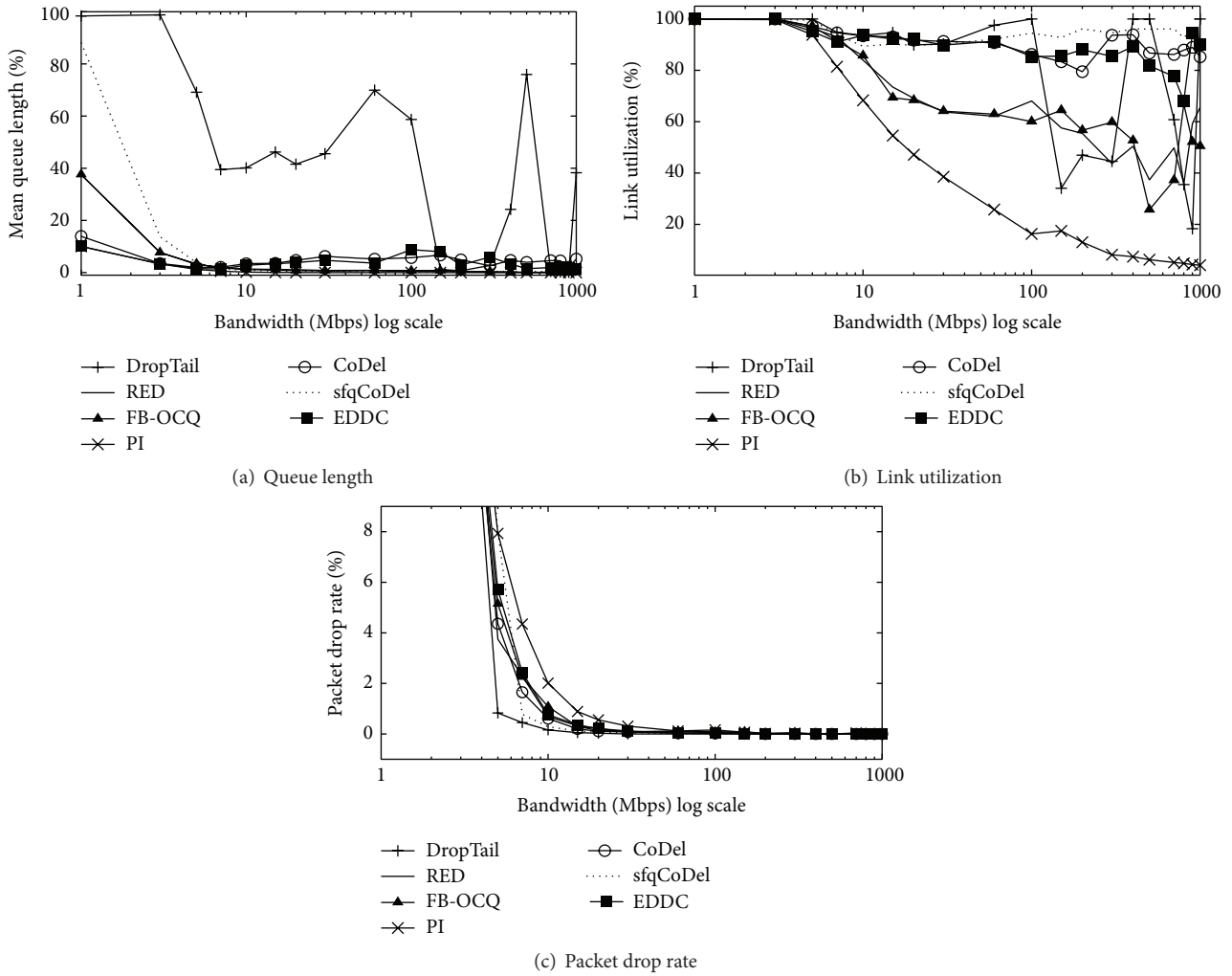


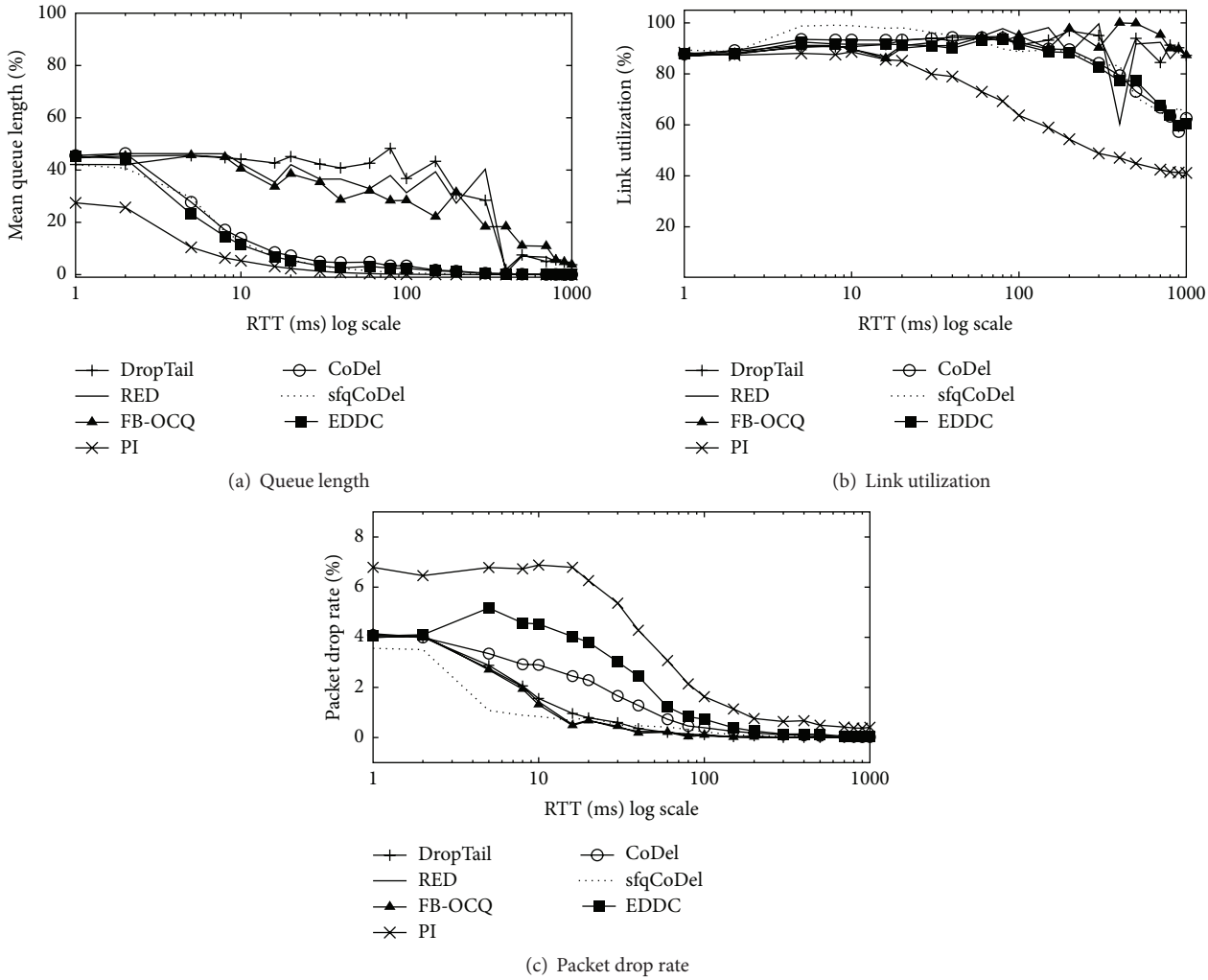
Figure 4: Experiment 1: varied link bandwidth.

varied bandwidth because of wireless signal attenuation and fading.

Firstly, Figure 4 presents our comparison results when we change the value of link bandwidth from 1 Mbps to 1000 Mbps. Our algorithm achieves very low and stable mean queue length (≈10%); even bandwidth is varied in a wide range. Meanwhile, the link utilization of EDDC is about 95%, which is higher than the default DropTail or RED and nearly approaches the modern sfqCoDel algorithm (96%). The packet drop rate performance of EDDC, however, is almost the same as the others. It makes sense that when we use a low-bandwidth link, the packet drop rate is high due to full buffer phenomenon. We conclude that our EDDC algorithm promisingly maintain the queue length of substation router, while link utilization or the busy level of link is acceptable under different kinds of links.

5.2.2. Experiment 2: Varied Link Round-Trip Time. Figure 5 shows our comparison of AQM algorithms performance in case of varied link propagation round-trip time (RTT). This criterion represents the latency of packets when using a loss-based TCP congestion control scheme, for example, TCP Reno. As mentioned before, one consequence of a large buffer (bufferbloat) is that TCP operation might be damaged because of untimely congestion information feedback of acknowledgment (ACK) packets. The higher the propagation RTT is, the higher chance of incorrect ACK feedback to the sources is. We observe the same results of EDDC performance for the mean queue length and link utilization. In Figure 5(c), the packet drop rate results of EDDC, however, are higher than the other AQM algorithms in the RTT range from 1 ms to 100 ms.

5.2.3. Experiment 3: Varied Link Packet Error Rate. Lastly, we change the packet error rate (PER) of the congested link. A wired link, for example, Ethernet cable, usually has a PER value of zero. A wireless link, however, has the wide range of PER values because of wireless signal characteristics such as



(a) Queue length

(b) Link utilization

(c) Packet drop rate

Figure 5: Experiment 2: varied link propagation round-trip time.

attenuation or fast fading. In order to simulate them, we vary PER from 0 to 0.9; for instance, a link with PER = 0.9 means that if we transmit 10 packets, only 1 packet is successfully delivered to destination.

In Figure 6(a), we see that EDDC keeps the router queue length at 2% which is lower and better than DropTail, RED, FB-OCQ, and CoDel algorithms. The modern sfqCoDel using the stochastic fair method seems to manage queue length to be too short (nearly 0%) which might increase the number of packet drop. It is, in fact, confirmed from Figure 6(c) that sfqCoDel has the highest packet drop rate (≥2%), while our algorithm EDDC only drops packets smartly at 1% ratio.

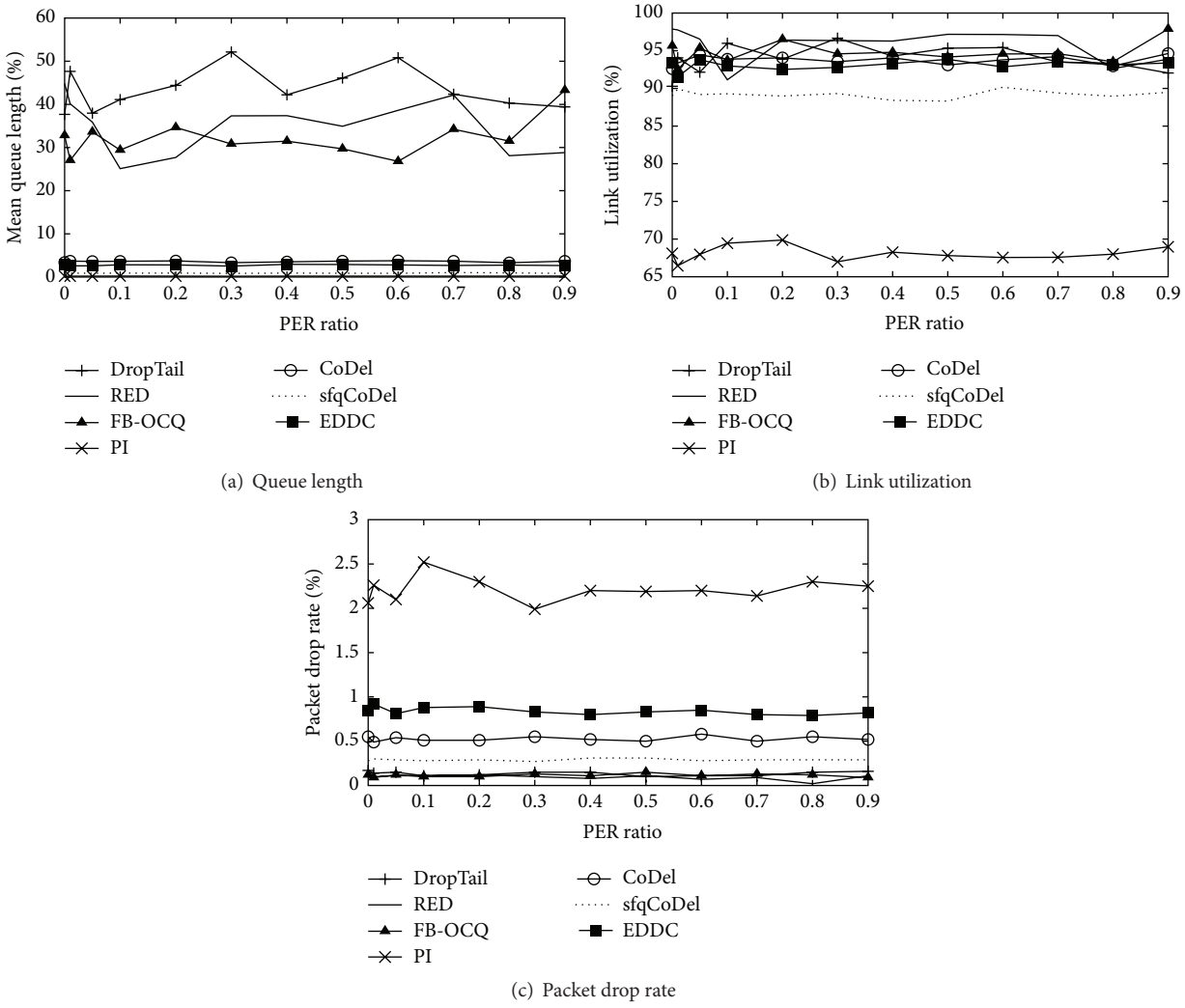
The above experiments support us to conclude some important advantages of our queue management algorithm EDDC. By smartly dropping packets with EDDC scheme, the mean queue length is kept at a stable and low level (2–5%) comapred with the other popular algorithms. Moreover, link utilization value is reasonable (80–95%) while the packet drop rate of EDDC is lower than the newest and modern algorithm (sfqCoDel).

6. Conclusions

An algorithm for better queue management of heterogeneous smart grid traffic was proposed. The algorithm, based on cascade control theory, smartly dropped packets waiting in the queue based on information of inflow smart grid traffic and current state of input queue. The algorithm was shown to effectively manage the queue length or the number of packets that are waiting in queue at a stable and low level (2–5%) while the link utilization does not reduce too much.

Competing Interests

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



(a) Queue length

(b) Link utilization

(c) Packet drop rate

Figure 6: Experiment 3: varied link packet error rate.

Acknowledgments

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the Grand Information Technology Research Center Support Program (IITP-2016-R71181610050001002) supervised by the IITP (Institute for Information & Communications Technology Promotion). This work was also supported by a 2-year research grant of Pusan National University.

References

[1] F. Baker and D. Meyer, "Internet protocols for the smart grid," Tech. Rep., 2011.

[2] P. Kansal and A. Bose, "Bandwidth and latency requirements for smart transmission grid applications," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1344–1352, 2012.

[3] J. G. Deshpande, E. Kim, and M. Thottan, "Differentiated services QoS in smart grid communication networks," *Bell Labs Technical Journal*, vol. 16, no. 3, pp. 61–81, 2011.

[4] O.-H. Borlaug, "Real-time applications in smart grids using internet communications:: latency issues and mitigations," 2014.

[5] The Virtual Access GW2027 Industrial Router, Virtual Access, 4, 2016.

[6] A. Radwan, H.-L. To, and W.-J. Hwang, "Optimal control for bufferbloat queue management using indirect method with parametric optimization," *Scientific Programming*, vol. 2016, Article ID 4180817, 10 pages, 2016.

[7] J. Gettys and K. Nichols, "Bufferbloat: dark buffers in the internet," *Queue*, vol. 9, no. 11, p. 40, 2011.

[8] D. Wischik and N. McKeown, "Part I: buffer sizes for core routers," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 3, pp. 75–78, 2005.

[9] D. Taht and J. Gettys, "Best practices for benchmarking codel and fq codel," Wiki Page on bufferbloat.net Web Site, 2012.

[10] V. G. Cerf, "Bufferbloat and other internet challenges," *IEEE Internet Computing*, vol. 18, no. 5, p. 80, 2014.

[11] K. Nichols and V. Jacobson, "Controlling queue delay," *Communications of the ACM*, vol. 55, no. 7, pp. 42–50, 2012.

[12] V. P. Rao, M. P. Tahiliani, and U. K. K. Shenoy, "Analysis of sfqCoDel for active queue management," in *Proceedings of the 5th International Conference on the Applications of Digital Information and Web Technologies (ICADIWT '14)*, pp. 262–267, Bengaluru, India, February 2014.

[13] R. Pan, P. Natarajan, C. Piglione et al., "PIE: a lightweight control scheme to address the bufferbloat problem," in *Proceedings of the IEEE 14th International Conference on High Performance Switching and Routing (HPSR '13)*, pp. 148–155, IEEE, Taipei, Taiwan, July 2013.

[14] C. V. Hollot, V. Misra, D. Towsley, and W.-B. Gong, "A control theoretic analysis of RED," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 3, pp. 1510–1519, April 2001.

[15] G. Raina, D. Towsley, and D. Wischik, "Part ii: control theory for buffer sizing," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 3, pp. 79–82, 2005.

[16] W. K. Ho, K. W. Lim, and W. Xu, "Optimal gain and phase margin tuning for PID controllers," *Automatica*, vol. 34, no. 8, pp. 1009–1014, 1998.

[17] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, 1993.

[18] M. May, J. Bolot, C. Diot, and B. Lyles, "Reasons not to deploy red," in *Proceedings of the IEEE 7th International Workshop on Quality of Service (IWQoS '99)*, pp. 260–262, London, UK, June 1999.

[19] T. Høiland-Jørgensen, "Technical description of fq codel," 2014.

[20] C.-K. Chen, H.-H. Kuo, J.-J. Yan, and T.-L. Liao, "GA-based PID active queue management control design for a class of TCP communication networks," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1903–1913, 2009.

[21] K. B. Kim, "Design of feedback controls supporting TCP based on the state-space approach," *IEEE Transactions on Automatic Control*, vol. 51, no. 7, pp. 1086–1099, 2006.

[22] D.-M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1–14, 1989.

[23] G. Patil and G. Raina, "Some guidelines for queue management in high-speed networks," in *Proceedings of the 3rd Asian Himalayas International Conference on Internet (AH-ICI '12)*, pp. 1–6, IEEE, Kathmandu, Nepal, November 2012.

[24] M. Chiang, A. Sutivong, and S. Boyd, "Efficient nonlinear optimizations of queuing systems," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '02)*, vol. 3, pp. 2425–2429, IEEE, Taipei, Taiwan, November 2002.

[25] G. Wang, Y. Xia, and D. Harrison, "An NS2 TCP evaluation tool," Tech. Rep., Internet Engineering Task Force, 2007.

Hindawi Publishing Corporation
Scientific Programming
Volume 2016, Article ID 5369780, 8 pages
http://dx.doi.org/10.1155/2016/5369780

Research Article

A Novel Metric Online Monocular SLAM Approach for Indoor Applications

Yongfei Li, Shicheng Wang, Dongfang Yang, and Dawei Sun

High-Tech Institute of Xi'an, Xi'an, Shaanxi 710025, China

Correspondence should be addressed to Yongfei Li; lyfei314@163.com

Received 9 May 2016; Revised 29 July 2016; Accepted 7 August 2016

Academic Editor: Wenbing Zhao

Monocular SLAM has attracted more attention recently due to its flexibility and being economic. In this paper, a novel metric online direct monocular SLAM approach is proposed, which can obtain the metric reconstruction of the scene. In the proposed approach, a chessboard is utilized to provide initial depth map and scale correction information during the SLAM process. The involved chessboard provides the absolute scale of scene, and it is seen as a bridge between the camera visual coordinate and the world coordinate. The scene is reconstructed as a series of key frames with their poses and correlative semidense depth maps, using a highly accurate pose estimation achieved by direct grid point-based alignment. The estimated pose is coupled with depth map estimation calculated by filtering over a large number of pixelwise small-baseline stereo comparisons. In addition, this paper formulates the scale-drift model among key frames and the calibration chessboard is used to correct the accumulated pose error. At the end of this paper, several indoor experiments are conducted. The results suggest that the proposed approach is able to achieve higher reconstruction accuracy when compared with the traditional LSD-SLAM approach. And the approach can also run in real time on a commonly used computer.

1. Introduction

In the robotics community, simultaneous localization and mapping (SLAM) refers to creating the surrounding map and determining self-position, which is necessary for a robot to autonomously navigate in an unknown environment [1]. The map of the unknown environment must be built incrementally. This means the class of methods must focus on computational algorithms which integrate the new information incrementally [2]. Because camera sensor is cheap, is light, and has low power requirement when compared with other sensors like depth camera, visual SLAM has become a popular research topic.

Traditional visual navigation usually uses a stereo visual system, which can directly provide the 3-dimensional information of circumstance, and the position of cameras can be easily estimated by utilizing the visual difference coming from two or more cameras. Whereas the accuracy of stereo visual navigation is limited by the length of base line, this problem is crucial especially in applications where the base line is seriously limited, such as remote sensing and micro UAVs. Therefore, the monocular visual navigation tends to be more general and commonly used.

Generally speaking, there exist two classes of monocular SLAM: feature-based methods and direct methods. In feature-based approaches, including filtering-based [3, 4] and keyframe-based, the geometric information is estimated from image sequences. The whole process is usually split into two sequential steps: extracting feature observations from the image and calculating the scene geometry and camera pose as a function of these feature observations only by using multiview stereo matches.

This uncoupling predigests the overall problem at the cost of information lose, such as information presenting curved edges. This class of image information often makes up a large part of the image especially in man-made environment and is important for a robot to fulfill tasks like obstacle avoidance.

Direct visual odometry (VO) methods overcome this limitation by calculating camera pose directly on the image intensities, in which all information contained in the image is

used. Moreover, more geometry information of the environment can be used in the direct methods, and that is helpful for obtaining higher accuracy and robustness, especially in simplex environments where few key points are available. The geometry information about the scene is valuable for robotics in many applications such as augmented reality. It is well known that direct image alignment is well established for stereo sensors or RGB-D [5, 6]. However, in monocular visual applications, the existed scale ambiguity problem is hard to solve by direct approach. Until recently [7–9], the precise and intact dense depth maps are computed with a variation formulation. However, its computational complexity is so large that a powerful GPU is required to guarantee the online availability. In [10], a semidense depth filtering formulation is proposed, which significantly reduces computational complexity. Thus, it is able to run online on a CPU and even on a smartphone [11]. With the assistance of key-points methods, direct tracking method even can achieve higher frame-rates on embedded platforms [12].

In both feature-based methods and direct methods, monocular SLAM can only get the reconstruction of scene up to a scale. There still exist more challenges of scale drift, which is seen as the major reason for accumulated error [13]. Due to the scale-drift phenomenon in monocular SLAM, the 3D similarity transform is adopted to represent camera poses instead of rigid body transformations [12]. And also loop closures detection is used as constraints to correct the scale drift. Using landmarks is also a method to reduce the scale drift, but it only can be applied in a cooperative scene [14]. As to the scale-ambivalent for the monocular SLAM method, it must be handled with either scale as a dubious factor or some additional information, such as point coordinates on a calibration object in the world coordinate and must be introduced to calculate the scale [15], and exploiting nonhomonymic motion constraints also work [16]. In our work, we solve the two problems by using a chessboard as a calibration reference. We need to be reminded here that the chessboard is widely and easily used. Therefore, in our approach, the chessboard is supposed to be involved in the visual field at the beginning of the SLAM process.

Contributions of This Paper. We present a metric online direct semidense monocular SLAM method, which calculates the real-time camera pose and builds consistent maps of the environment in metric scale, even in a large-scale environment. The method estimates the depth maps using a filtering-based algorithm, coupled with direct image alignment, and a chessboard is used to provide an initial depth estimation, in which the scale of the world is introduced into the mapping. The method presents steady tracking of the motion of the camera and accurate mapping of the environment and can run in real time on a CPU. The main contributions of this paper are (1) a method to introduce the metric scale into the monocular SLAM system and (2) a method to correct scale drift with a calibration object where a rule is proposed to determine when the calibration object is within the horizon of the camera and required to be detected for the purpose of reducing computation cost.

2. Preliminaries

In this section, some relevant mathematical definitions and notations used in this paper will be introduced. In particular, we introduce the most widely used camera model, pinhole camera model (Section 2.1), and represent the 3D poses as elements of Lie-Algebras (Section 2.2). In addition, we briefly introduce the solution for a weighted least-squares minimization on Lie-manifolds (Section 2.3).

Notations. Matrices are donated by bold, capital letters (R) and vectors by bold, lower case letters (ξ):

$\mathbf{R}_{ab}$: direction cosine matrix between a- and b-coordinates;

$\mathbf{t}_{ab}$: relative translation-vector between a- and b-coordinates;

$\mathbf{K}_i$: camera matrix for ith camera;

$\mathbf{X}_C, \mathbf{X}_W$: 3D point coordinate in camera coordinate or world coordinate;

Ω: set of normalized pixel coordinates;

$I : \Omega \to R$: images;

$D : \Omega \to R^+$: inverse depth map;

$V : \Omega \to R^+$: inverse depth variance map.

2.1. Camera Model and Coordinate Definition. The most widely used camera model is the pinhole camera model which is shown in Figure 1.

As described in Figure 1, the camera perspective model can be expressed as

$$\begin{aligned}\mathbf{Z}_c \cdot \mathbf{r}_{O_I P_I} &= \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{r}_{O_C P_C} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} a_x & 0 & u_0 & 0 \\ 0 & a_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}_{WC} & \mathbf{t}_{WC} \\ 0_{(1\times 3)} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{r}_{O_W P_W} \\ 1 \end{bmatrix},\end{aligned} \tag{1}$$

where (u_0, v_0) is the camera principal point, f is the metric focal length, and dx and dy denote the physical width and height of one pixel. $\mathbf{R}_{WC}$, $\mathbf{t}_{WC}$ are external coefficients between C and W.

As described in aforementioned introduction, a chessboard is used to initialize the depth map. During the initialization process, the relationship between the camera coordinate and the world coordinate is estimated. Herein, the world coordinate is defined as follows: the original point is the left-top point of the chessboard, the X-coordinate is the top line of the chessboard and points to the right, and Y-coordinate is the left line of the chessboard and points to the bottom, as described in Figure 2.

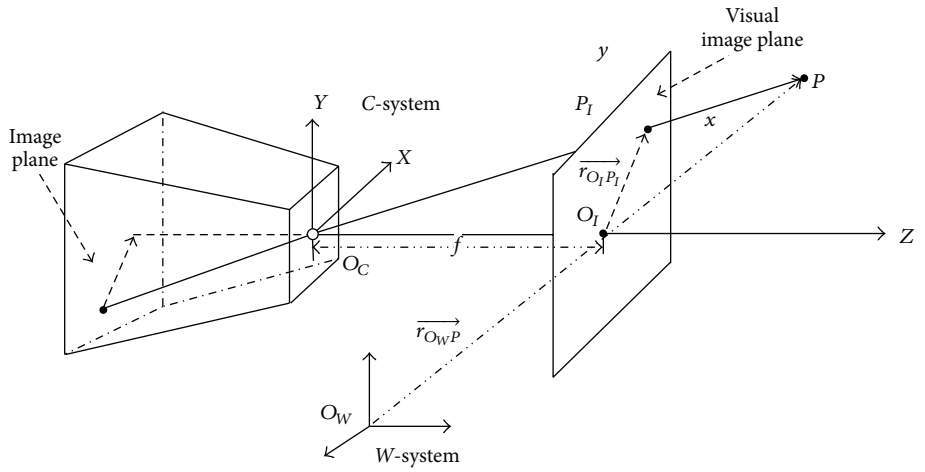


Figure 1: Camera perspective model.

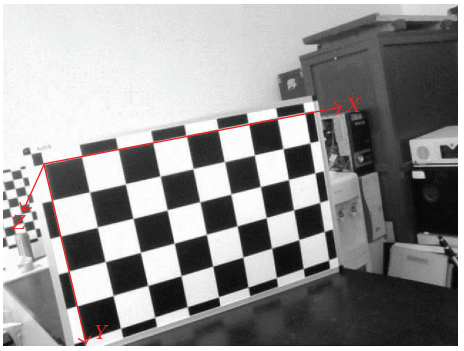


Figure 2: World coordinate definition.

2.2. 3D Pose Representing as Elements of Lie-Algebras. In this section,the representation of 3D pose transformation is described in the same way as in previous researches, such as [9, 10]. In order to guarantee the fluency of the presentation, we still utilize the most commonly used description of this problem. Usually, 3D rigid body transform $\mathbf{G} \in \mathrm{SE}(3)$ denotes translation and rotation in 3D that is defined by

$$\mathbf{G} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}, \quad \text{with } \mathbf{R} \in \mathrm{SO}(3),\ \mathbf{t} \in R^3. \tag{2}$$

And 3D similarity transform $\mathbf{S} \in \mathrm{Sim}(3)$ denotes scaling, translation, and rotation that is defined by

$$\mathbf{G} = \begin{pmatrix} s\mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix}, \quad \text{with } \mathbf{R} \in \mathrm{SO}(3),\ \mathbf{t} \in R^3,\ s \in R^+. \tag{3}$$

But a nonredundant expression for the camera pose is needed during optimization, which cannot be given by definition above, so the corresponding element $\boldsymbol{\xi} \in \mathrm{SE}(3)$ of the associated Lie-algebra is used to represent 3D rigid body transform and $\boldsymbol{\xi} \in \mathrm{Sim}(3)$ for 3D similarity transform. Elements are transformed into SE(3) by the exponential map $\mathbf{G} = \exp_{\mathrm{SE}(3)}(\boldsymbol{\xi})$ for rigid body transform and into Sim(3) by map $\mathbf{G} = \exp_{\mathrm{Sim}(3)}(\xi)$ for similarity transform, and their inverse is denoted by $\boldsymbol{\xi} = \log_{\mathrm{SE}(3)}(\mathbf{G})$ and $\boldsymbol{\xi} = \log_{\mathrm{Sim}(3)}(\mathbf{G})$. So the transformation moving a point from frame i to frame j is written as $\boldsymbol{\xi}_{ij}$. As in rigid body transform description, the pose concatenation operator $\circ : \mathrm{SE}(3) \times \mathrm{SE}(3) \to \mathrm{SE}(3)$ is defined as

$$\begin{aligned} \boldsymbol{\xi}_{ki} &:= \boldsymbol{\xi}_{kj} \circ \boldsymbol{\xi}_{ji} \\ &:= \log_{\mathrm{SE}(3)} \left(\exp_{\mathrm{SE}(3)} \left(\boldsymbol{\xi}_{kj} \right) \cdot \exp_{\mathrm{SE}(3)} \left(\boldsymbol{\xi}_{ji} \right) \right) \end{aligned} \tag{4}$$

which can be defined analogously for similarity transform. Please see [10] for more details.

2.3. Solution for Weighted Gauss-Newton Optimization. The Gauss-Newton algorithm is effective with nonlinear least-squares problems, with the advantage of a small computation cost [17]. The problem is usually described as follows: given m functions $\mathbf{r} = (r_1, \ldots, r_m)$ of n variables $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$, with $m > n$, the Gauss-Newton algorithm iteratively finds the minimum of the sum of squares:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} r_i^2(\boldsymbol{\beta}). \tag{5}$$

Beginning with an initial guess $\boldsymbol{\beta}^{(0)}$, the method proceeds with the iterations [17]:

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\mathbf{J}_{\mathbf{r}}^T \mathbf{J}_{\mathbf{r}} \right)^{-1} \mathbf{J}_{\mathbf{r}}^T \mathbf{r} \left(\boldsymbol{\beta}^{(s)} \right), \tag{6}$$

where if $\mathbf{r}$ and $\boldsymbol{\beta}$ are column vectors, the entries of the Jacobian matrix are [17]

$$\left(\mathbf{J}_{\mathbf{r}} \right)_{ij} = \frac{\partial r_i \left(\boldsymbol{\beta}^{(s)} \right)}{\partial \beta_j}. \tag{7}$$

And an iteratively reweighted least-squares problem is proposed to be robust to outliers arising, for example, from occlusions or reflections, which can be expressed as [17]

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \omega_i(\boldsymbol{\xi}) r_i^2(\boldsymbol{\beta}) \tag{8}$$

and can be proceeded by the iterations [17]

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \left(\mathbf{J}_{\mathbf{r}}^{T}\mathbf{W}\mathbf{J}_{\mathbf{r}}\right)^{-1}\mathbf{J}_{\mathbf{r}}^{T}\mathbf{W}\mathbf{r}\left(\boldsymbol{\beta}^{(s)}\right), \tag{9}$$

where $\mathbf{W} = \mathbf{W}(\boldsymbol{\beta}^{(s)})$ is a weight matrix computed in each iteration and used to downweight large residuals.

3. Metric Online Monocular SLAM

This main contribution of this paper is that it provides a metric online monocular SLAM approach, by using a commonly used chessboard reference. The process can be divided into 5 parts. The chessboard provides the initial depth estimation of the scene, and the initial guess can be seen as the scaled source of the scene reconstruction. The initial depth estimation results are able to correct the scale drift through the key frames transfer, and thus we can obtain a global metric reconstruction. The whole process of this approach is shown in Figure 3.

In this section, we introduce our work from 4 parts: the initial depth estimation in Section 3.1, the estimation of camera pose using alignment in Section 3.2, method to correct the accumulated pose error with the aid of a chessboard in Section 3.3, and the depth map estimation and optimization in Section 3.4.

3.1. Initial Metric Depth Estimation. In the initial process, unlike in the traditional LSD-SLAM approach, we use a standard, key point-based method to obtain the initial depth map with the aid of a calibration object, which is a commonly used chessboard in this paper. We need to be reminded here that any other reference object is also good to fulfill this initial depth estimation process.

During the calibration process, the chessboard corners detection should be executed at the beginning. With the known 2D coordinates of chessboard corners and corresponding 3D coordinates, the relative pose of camera to the world coordinate can be got, which is known as the PNP problem [18].

3.1.1. Image Points Matching. To run online, only the depth of pixels with sufficiently large intensity gradient, which means that the pixel is a corner or on the edges, is estimated. We search the corresponding points of those pixels on the epipolar lines in the second image using a window-based matching approach with a window size of 3 pixels. Also, parallax constraint and sequence constraint are used to reduce the mismatching.

3.1.2. Initial Depth Map Estimation. With the known intrinsic camera parameters, extrinsic parameters, and corresponding image point pairs, the initial depth map can be estimated:

$$\begin{bmatrix}\mathbf{R}_{c_1 w}^{-1}\mathbf{K}^{-1}\mathbf{x}_1 & -\mathbf{R}_{c_2 w}^{-1}\mathbf{K}^{-1}\mathbf{x}_2\end{bmatrix}\begin{bmatrix}k_1\\k_2\end{bmatrix} = \mathbf{R}_{c_1 w}^{-1}\mathbf{t}_{c_1}^{w} - \mathbf{R}_{c_2 w}^{-1}\mathbf{t}_{c_2}^{w}, \tag{10}$$

where k_1, k_2 is inverse depth of pixel in the 1st image and 2nd image and $\mathbf{x}_1, \mathbf{x}_2$ is the homogeneous image pixel coordinate.

3.2. Pose Tracking. As a recursive process in the visual localization, the pose tracking is the main task. Now the commonly used pose tracking method is the image alignment, in which image sequences sampled in different time steps are consequently utilized, to provide the location verification of the moving camera. As the same in traditional monocular visual odometry researches, such as [7], we use the image alignment method by utilizing the direct features. The 3D pose $\boldsymbol{\xi}_{ji} \in \mathrm{SE}(3)$ of a new frame related to its keyframe is calculated using the direct SE(3) image alignment, and pose of keyframe is tracked using the direct Sim(3) image alignment. We need to be reminded here that, as described in the commonly used approaches, we use the same image alignment based pose tracking approach; please see [19] for more details. In order to make the description more fluent, we use the same nomination during the problem formulation.

3.2.1. Direct SE(3) *Image Alignment.* The pose estimation of a new frame is treated as a problem to minimize the variance-normalized photometric error:

$$E_p\left(\boldsymbol{\xi}_{ji}\right) = \sum_{\mathbf{p}\in\Omega_{D_i}}\left\|\frac{r_p^2\left(\mathbf{p},\boldsymbol{\xi}_{ji}\right)}{\sigma_{r_p(\mathbf{p},\boldsymbol{\xi}_{ji})}^2}\right\|_{\delta}$$
$$\text{with } r_p^2\left(\mathbf{p},\boldsymbol{\xi}_{ji}\right) \coloneqq I_i\left(\mathbf{p}\right) - I_j\left(\omega\left(\mathbf{p}, D_i\left(\mathbf{p}\right), \boldsymbol{\xi}_{ji}\right)\right),\ \sigma_{r_p(\mathbf{p},\boldsymbol{\xi}_{ji})}^2 \coloneqq 2\sigma_I^2 + \left(\frac{\partial r_p^2\left(\mathbf{p},\boldsymbol{\xi}_{ji}\right)}{\partial D_i\left(\mathbf{p}\right)}\right)^2 V_i\left(\mathbf{p}\right), \tag{11}$$

where $\|\cdot\|_{\delta}$ is the Huber norm

$$\left\|r^2\right\|_{\delta} \coloneqq \begin{cases}\dfrac{r^2}{2\delta} & \text{if } |r| \le \delta\\ |r| - \dfrac{\delta}{2} & \text{otherwise.}\end{cases} \tag{12}$$

The Huber norm is applied to normalize the estimation residual. The residual's variance $\sigma_{r_p(\mathbf{p},\boldsymbol{\xi}_{ji})}^2$ is computed using covariance propagation:

$$\sum_f \approx \mathbf{J}_f \sum_{\mathbf{x}} \mathbf{J}_f^T. \tag{13}$$

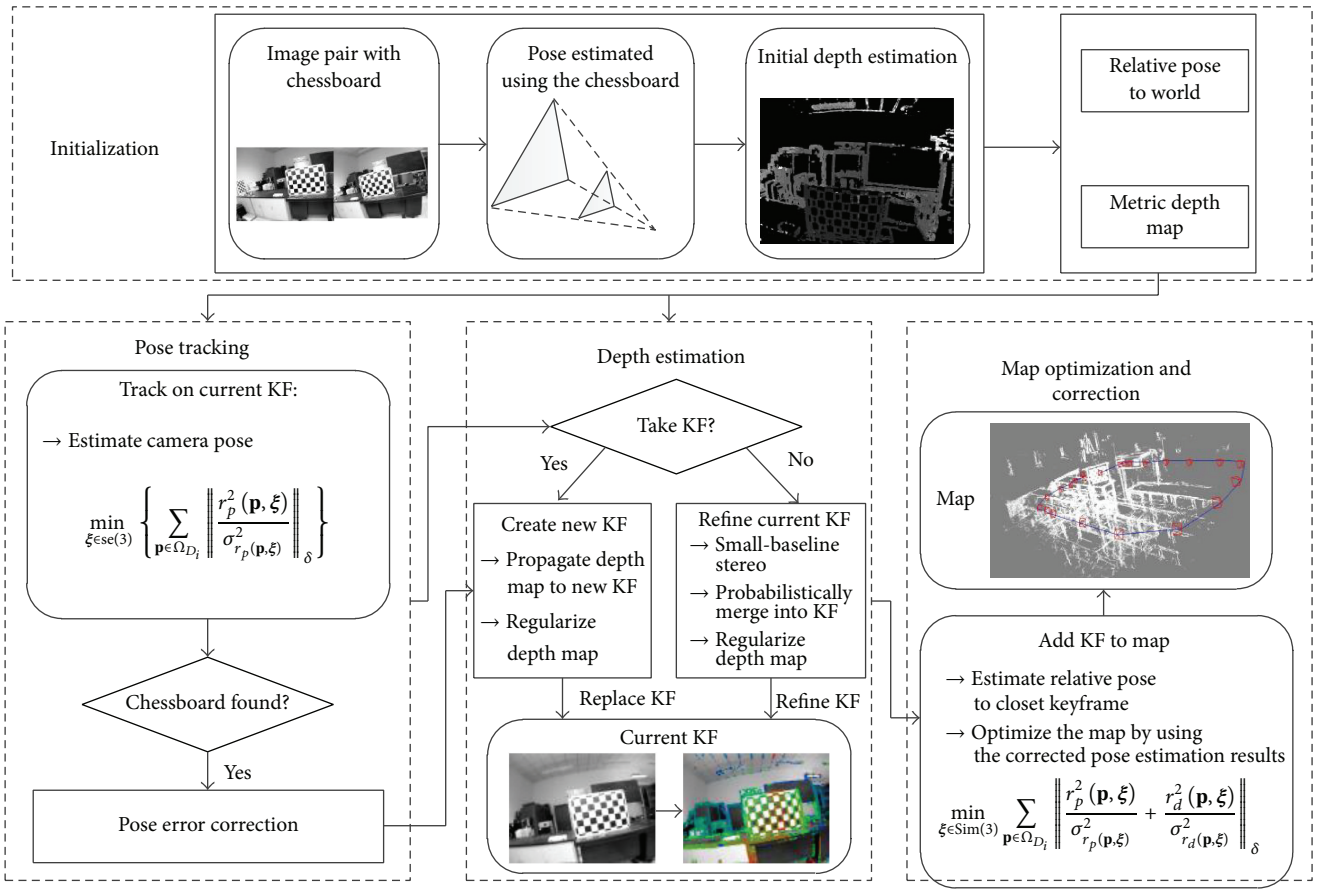


Figure 3: Overview over the whole SLAM system.

V_i is the inverse depth variance, and σ_I^2 is the image dense noise which is assumed to follow Gaussian distribution. The problem is solved using iteratively reweighted Gauss-Newton optimization described in Section 2.3.

3.2.2. Direct Sim(3) *Image Alignment.* To solve the scale-drift problem, direct Sim(3) image alignment is used to estimate edges between keyframes. After the depth map of a keyframe is refined, it is scaled to make its mean inverse depth to be one. Then, the direct Sim(3) image alignment is performed to elegantly incorporate the scaling difference between keyframes. Similarly to direct SE(3) image alignment, the pose between two keyframes, represented as 3D similarity transform $\mathbf{S} \in \mathrm{Sim}(3)$, is estimated as a problem to minimize the variance-normalized photometric error:

$$E\left(\boldsymbol{\xi}_{ji}\right) = \sum_{\mathbf{p}\in\Omega_{D_i}} \left\| \frac{r_p^2\left(\mathbf{p}, \boldsymbol{\xi}_{ji}\right)}{\sigma^2_{r_p(\mathbf{p},\boldsymbol{\xi}_{ji})}} + \frac{r_d^2\left(\mathbf{p}, \boldsymbol{\xi}_{ji}\right)}{\sigma^2_{r_d(\mathbf{p},\boldsymbol{\xi}_{ji})}} \right\|_\delta. \tag{14}$$

Here, a depth residual r_d is incorporated, which penalizes deviations in inverse depth between keyframes, allowing to directly estimate the scaled transformation between them. And the depth residual r_d and its variance $\sigma^2_{r_d(\mathbf{p},\boldsymbol{\xi}_{ji})}$ are computed as [10]

$$\begin{aligned} r_d\left(\mathbf{p}, \boldsymbol{\xi}_{ji}\right) &\coloneqq \left[\mathbf{p}'\right]_3 - D_j\left(\left[\mathbf{p}'\right]_{1,2}\right), \\ \sigma^2_{r_d(\mathbf{p},\boldsymbol{\xi}_{ji})} &\coloneqq \left(V_j\left(\left[\mathbf{p}'\right]_{1,2}\right)\right)\left(\frac{\partial r_d\left(\mathbf{p}, \boldsymbol{\xi}_{ji}\right)}{\partial D_j\left[\mathbf{p}'\right]_{1,2}}\right)^2 \\ &\quad + V_i\left(\mathbf{p}\right)\left(\frac{\partial r_d\left(\mathbf{p}, \boldsymbol{\xi}_{ji}\right)}{\partial D_i\left(\mathbf{p}\right)}\right)^2 \end{aligned} \tag{15}$$

and $\mathbf{p}' \coloneqq \omega_s(\mathbf{p}, D_i(\mathbf{p}), \boldsymbol{\xi}_{ji})$ denotes the corresponding point.

The problem can also be solved with iteratively reweighted Gauss-Newton optimization, which is the most commonly adopted approach in nonlinear optimization problem, as described in Section 2.3.

3.3. Pose Error Correction. In this section, we will show the outer assistance of a chessboard work in the pose error correction and in the depth maps accumulation errors correction. This is the main contribution of our work. Our SLAM approach is designed for the indoor robots, which means that it is possible for the robots to see the calibration object more

than once while moving. So it is practical to correct the pose estimation error accumulated with the calibration object.

3.3.1. Chessboard Corners Detection. It is unnecessary to detect the chessboard in every frame, which is of great computational cost. We give a principle to judge whether to detect the chessboard with the aid of pose of current frame.

When the four corners of the chessboard can be observed, the whole chessboard is inside the horizon, which can be used to decide when to detect the chessboard. In the world coordinate, the four corners of the chessboard are $\mathbf{X_1} = (0,0,0)$, $\mathbf{X}_2 = (w,0,0)$, $\mathbf{X}_3 = (0,h,0)$, and $\mathbf{X}_4 = (w,h,0)$, where w and h are the width and height of the chessboard. The pose of current frame to world coordinate can be calculated:

$$\begin{aligned}\mathbf{R}_{c_{\mathrm{cf}}}^{w} &= s\mathbf{R}_{c_{\mathrm{kf}}}^{w}\mathbf{R}_{c_{\mathrm{cf}}}^{c_{\mathrm{kf}}},\\ \mathbf{t}_{c_{\mathrm{cf}}}^{w} &= s\mathbf{R}_{c_{\mathrm{kf}}}^{w}\mathbf{t}_{c_{\mathrm{cf}}}^{c_{\mathrm{kf}}} + \mathbf{t}_{c_{\mathrm{kf}}}^{w},\end{aligned} \tag{16}$$

where $s\mathbf{R}_{c_{\mathrm{kf}}}^{w}$ and $\mathbf{t}_{c_{\mathrm{kf}}}^{w}$ are the pose of keyframe and $\mathbf{R}_{c_{\mathrm{cf}}}^{c_{\mathrm{kf}}}$ and $\mathbf{t}_{c_{\mathrm{cf}}}^{c_{\mathrm{kf}}}$ are the relative pose of current frame to current keyframe.

And homogeneous image pixel coordinate can be expressed as

$$\mathbf{x}_i = \mathbf{K}\left(\mathbf{R}_{w}^{c_{\mathrm{cf}}}\mathbf{X}_i + \mathbf{t}_{w}^{c_{\mathrm{kf}}}\right), \quad i = 1,2,3,4. \tag{17}$$

A chessboard detection is performed only under the condition that $\mathbf{x}_i$ is within the image:

$$\begin{aligned}0 \le \frac{\mathbf{x}_{i(1,1)}}{\mathbf{x}_{i(3,1)}} \le L_w,\\ 0 \le \frac{\mathbf{x}_{i(2,1)}}{\mathbf{x}_{i(3,1)}} \le L_h,\\ i = 1,2,3,4.\end{aligned} \tag{18}$$

Usually, successful chessboard detection is hard to achieve when the camera is too far away from the chessboard, even when the whole chessboard can be observed, so we add a limitation to avoid the case:

$$\left\|\mathbf{t}_{c_{\mathrm{cf}}}^{w}\right\| \le 3. \tag{19}$$

3.3.2. Pose Correction. When chessboard detection is performed successfully on current frame while failing on previous frame, a new keyframe will be created with a corrected pose which is estimated with the correspondence between the image coordinate and the world coordinate of those chessboard corners. Then, relative pose between new created keyframe and previous keyframe can be calculated as

$$\begin{aligned}\mathbf{R}_{c_{\mathrm{cf}}}^{c_{\mathrm{kf}}} &= \frac{1}{s}\mathbf{R}_{c_{\mathrm{kf}}}^{w\ -1}\mathbf{R}_{c_{\mathrm{cf}}}^{w},\\ \mathbf{t}_{c_{\mathrm{cf}}}^{c_{\mathrm{kf}}} &= \frac{1}{s}\mathbf{R}_{c_{\mathrm{kf}}}^{w}\left(\mathbf{t}_{c_{\mathrm{kf}}}^{w} - \mathbf{t}_{c_{\mathrm{cf}}}^{w}\right)\end{aligned} \tag{20}$$

with which the depth map of new keyframe can be initialized by projecting points from the previous frame.

3.4. Depth Map Estimation and Optimization. The depth map estimation problem is the most commonly referred problem in monocular SLAM, and in this paper we still use the most common method to execute the depth map estimation using the method proposed in [19]. When a new frame is obtained, we first measure whether the camera has moved far away enough from its keyframe that a new keyframe should be created using it. To do this, a weighted combination of relative distance and angle to the current keyframe is threshold, which is the same as in [19].

After the pose estimation process, the pose graph optimization is necessary to continuously optimize the map which consists of a set of keyframes and their camera poses. The error function is defined in the following equation, the same in [20],

$$\begin{aligned}&E\left(\boldsymbol{\xi}_{W1}\cdots\boldsymbol{\xi}_{Wn}\right)\\ &\quad\coloneqq \sum_{(\xi_{ji},\Phi_{ji})\in\varepsilon}\left(\boldsymbol{\xi}_{ji}\cdot\boldsymbol{\xi}_{Wi}^{-1}\cdot\boldsymbol{\xi}_{Wj}\right)^{T}\Phi_{ji}^{-1}\left(\boldsymbol{\xi}_{ji}\cdot\boldsymbol{\xi}_{Wi}^{-1}\cdot\boldsymbol{\xi}_{Wj}\right).\end{aligned} \tag{21}$$

4. Experiments

In the experiments process, the SLAM approach is executed in an indoor environment, where the visual scene is occupied by artificial equipment. In the experiment, the camera is selected as a commonly used industrial camera, the frame-frequency of which is higher than 30 frames/sec.

Before the experiment, the camera is placed in front of a chessboard for the initial alignment, from which the initial positions and gestures between camera and word coordinate are calculated. During the experiment process, the hand-held camera is moved around the house arbitrarily and returned to the start position at the end of the experiment. Based on the same datasets sampled in the moving process, our SLAM approach is run twice to provide the comparisons of these two different methods. The first experiment utilizes the most common monocular SLAM approach, as described in [9]. The second experiment is calculated with the assistance of the chessboard, especially in the correction of error accumulation. Both the reconstruction results and the ego-motion estimation results are depicted in Figure 4. Herein the SLAM results in only two keyframes are provided for simplicity, and the comparison in the whole moving process is the same as in the selected keyframes.

As depicted in the experimental results, we can easily find that there exits accumulated error in pose estimation when the traditional LSD-SLAM method is used, which is shown in Figure 4. As a result, the corresponding reconstruction results are also degenerated obviously. Thus, we can intuitively find that there exist ghost images of some reconstructed objects. While the proposed method with the outer reference assistance is used, the ego-motion estimation results can be improved, from which the reconstruction results can also be corrected with the assistance of chessboard calibration. Thus, we can achieve higher accuracy mapping of the experiment scene, and the reconstruction map of the scene in the second method is able to obtain solely results, without any ghost reconstruction in Figures 5 and 6.

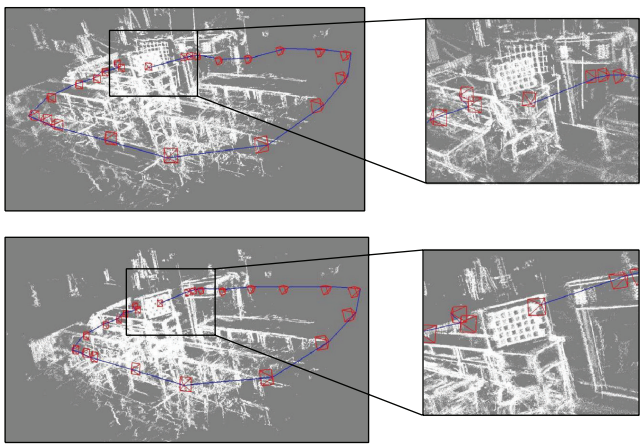


Figure 4: Reconstruction of scene.

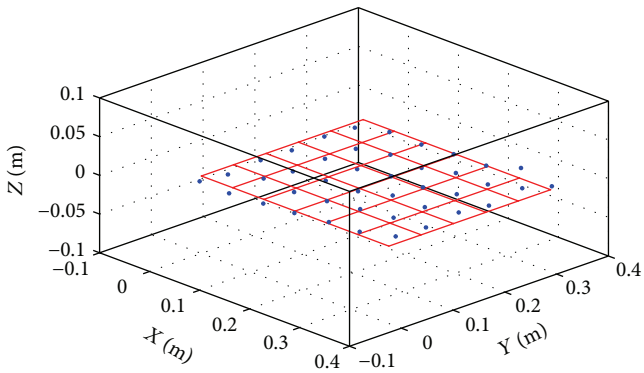


Figure 5: A comparison between the reconstructed points and the real ones. The red points are real chessboard corners and the blue ones are those reconstructed.

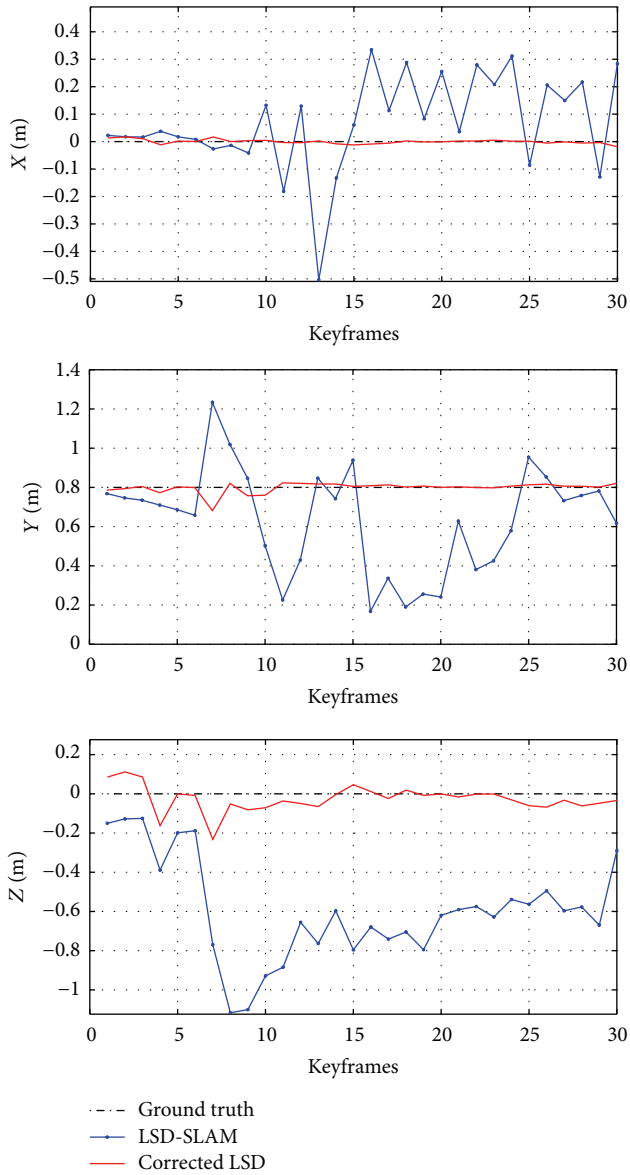


Figure 6: Reconstructed chessboard corners in different keyframes.

The aforementioned results can only provide a visualized comparison. In order to assess the proposed method quantitatively, we also provide a numerical analysis of the reconstruction results. Because the around scene during the experimental process is randomly selected, without any other information about the accurate scale and size, thus we choose the reconstructed result of the known chessboard to verify the accuracy of our method. As depicted in Figure 5, the comparison between the reconstructed chessboard corners and the real ones is shown.

As depicted in Figure 5, because our SLAM system can achieve a metric reconstruction of scene, we can find that the reconstruction precision is well improved remarkably with the assistance of the chessboard measurements.

In addition, all the coordinates of a randomly selected chessboard corner in different keyframes are also counted, and it is used to show the variation of accuracy about the reconstructed results.

From the results shown in Figure 6, we find that, at the beginning of the experiments, both the original LSD-SLAM method and the proposed corrected LSD-SLAM method can achieve accurate reconstruction of the chessboard corner. But along with the increasing measurements, the reconstruction error of the original LSD-SLAM method increases rapidly, while that of those proposed method is kept in a limit range, which indicates that our method can correct the accumulative error.

5. Conclusion

In this article, a metric direct monocular SLAM system is introduced, which can run in real time on a CPU and can obtain metric reconstruction of the scene. Based on the assistance of a chessboard, the initial depth map is estimated; meanwhile, the similarity transform between known world coordinate and the map coordinate is calculated, which can be used to convert the map to the known world coordinate. The system is tested in a complex indoor environment, and its accuracy is verified with a comparison between the estimated

chessboard corner coordinates and the real ones. The indoor experiments prove the effectiveness of the proposed metric monocular SLAM approach.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Natural Science Fund of China 61403398.

References

[1] M.-H. Le, A. Vavilin, and K.-H. Jo, "3D scene reconstruction enhancement method based on automatic context analysis and convex optimization," *Neurocomputing*, vol. 137, pp. 71–78, 2014.

[2] H. Yu, R. Sharma, R. W. Beard, and C. N. Taylor, "Observability-based local path planning and obstacle avoidance using bearing-only measurements," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1392–1405, 2013.

[3] D. Yang, Z. Liu, F. Sun, J. Zhang, H. Liu, and S. Wang, "Recursive depth parameterization of ego-motion estimating: observability analysis and performance evaluation," *Information Sciences*, vol. 287, pp. 38–49, 2014.

[4] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.

[5] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proceedings of the 26th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '13)*, pp. 2100–2106, Tokyo, Japan, November 2013.

[6] M. Dhome, M. Richetin, J.-T. Lapreste, and G. Rives, "Determination of the attitude of 3-D objects from a single perspective view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 12, pp. 1265–1278, 1989.

[7] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Pattern Recognition*, M. Goesele, S. Roth, A. Kuijper, B. Schiele, and K. Schindler, Eds., vol. 6376 of *Lecture Notes in Computer Science*, pp. 11–20, Springer, Berlin, Germany, 2010.

[8] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[9] D. Yang, F. Sun, S. Wang, and J. Zhang, "Simultaneous estimation of ego-motion and vehicle distance by using a monocular camera," *Science China Information Sciences*, vol. 57, no. 5, pp. 1–10, 2014.

[10] J. Engel, J. Sturm, and D. Cremers, "Scale-aware navigation of a low-cost quadrocopter with a monocular camera," *Robotics and Autonomous Systems*, vol. 62, no. 11, pp. 1646–1656, 2014.

[11] T. Schops, J. Enge, and D. Cremers, "Semi-dense visual odometry for AR on a smartphone," in *Proceedings of the 13th IEEE International Symposium on Mixed and Augmented Reality (ISMAR '14)*, pp. 145–150, IEEE, Munich, Germany, September 2014.

[12] E. Rodolà, A. Albarelli, D. Cremers, and A. Torsello, "A simple and effective relevance-based point sampling for 3D shapes," *Pattern Recognition Letters*, vol. 59, pp. 41–47, 2015.

[13] H. Strasdat, J. Montiel, and A. Davison, "Scale drift-aware large scale monocular SLAM," *Robotics: Science and Systems*, vol. 6, no. 6, pp. 56–62, 2010.

[14] E. Hernández, J. M. Ibarra, J. Neira, and R. Cisneros, "Visual SLAM with oriented landmarks and partial odometry," in *Proceedings of the 21st International Conference on Electrical Communications and Computers*, pp. 39–45, San Andres Cholula, Mexico, September 2011.

[15] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proceedings of the International Conference on Computer Vision (ICCV '03)*, pp. 1403–1410, Nice, France, October 2003.

[16] N. Ohnishi and A. Imiya, "Appearance-based navigation and homing for autonomous mobile robot," *Image and Vision Computing*, vol. 31, no. 6-7, pp. 511–532, 2013.

[17] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, Pa, USA, 1996.

[18] Y. Wu and Z. Hu, "PnP problem revisited," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 1, pp. 131–141, 2006.

[19] J. Engel, T. Schops, and D. Cremers, "LSD-SLAM: large-scale direct monocular SLAM," in *Proceedings of the European Conference on Computer Vision (ECCV '14)*, pp. 834–849, Zurich, Switzerland, September 2014.

[20] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: a general framework for graph optimization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '11)*, pp. 3607–3613, Shanghai, China, May 2011.