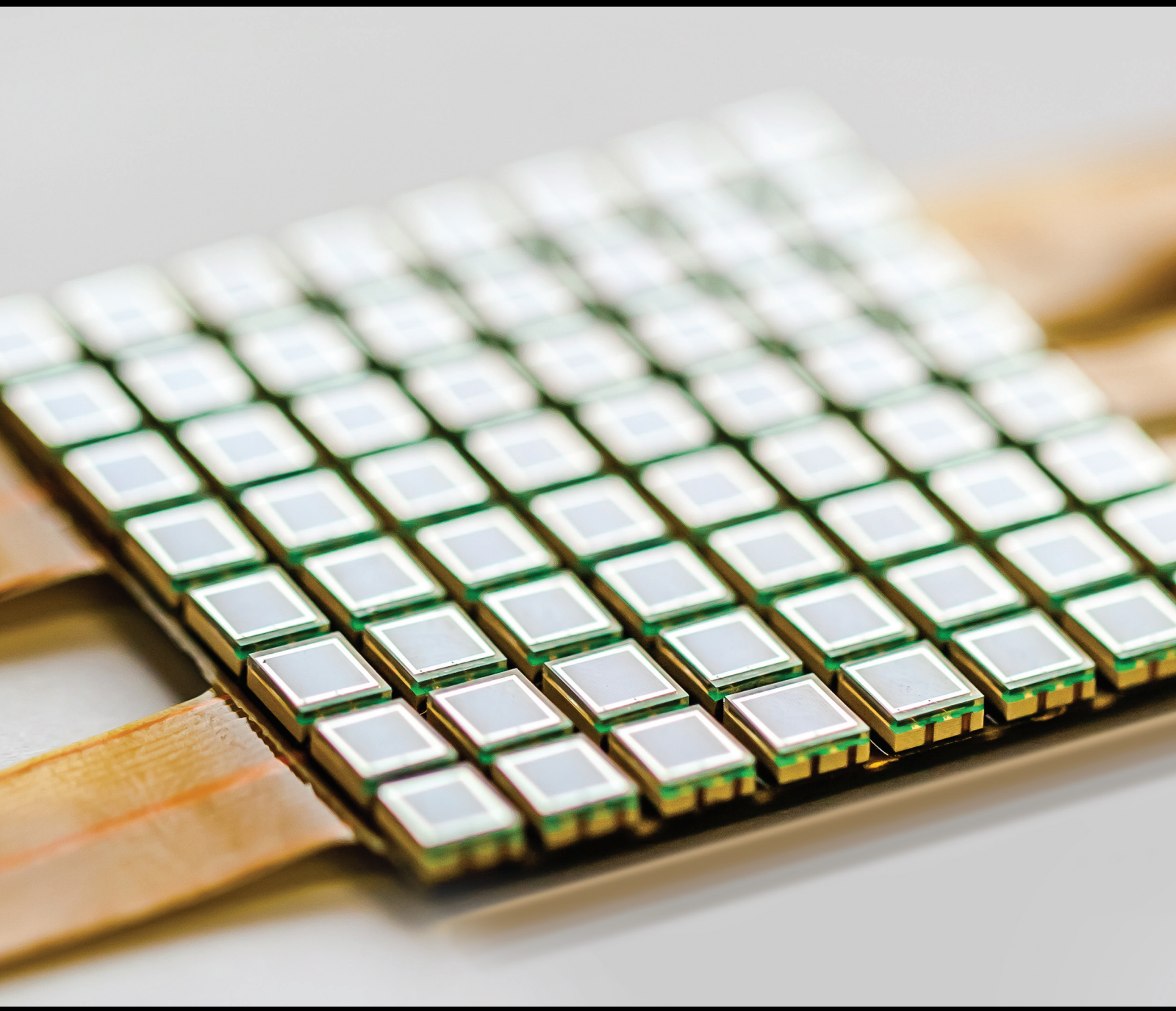
Journal of Sensors

Integration of Sensors in Control and Automation Systems

Guest Editors: Rafael Morales, Antonio Fernández-Caballero, José A. Somolinos, and Hebertt Sira-Ramírez



Integration of Sensors in Control and Automation Systems

Journal of Sensors

Integration of Sensors in Control and Automation Systems

Guest Editors: Rafael Morales, Antonio Fernández-Caballero, José A. Somolinos, and Hebertt Sira-Ramírez

Editorial Board

Contents

Hindawi
Journal of Sensors
Volume 2017, Article ID 6415876, 2 pages
http://dx.doi.org/10.1155/2017/6415876

Editorial

Integration of Sensors in Control and Automation Systems

Rafael Morales-Herrera,[1] Antonio Fernández-Caballero,[1] José A. Somolinos,[2] and Hebertt Sira-Ramírez[3]

[1]*Escuela de Ingenieros Industriales de Albacete, Universidad de Castilla-La Mancha, 02071- Albacete, Spain*
[2]*Escuela de Ingenieros Navales, Universidad Politécnica de Madrid, Madrid, Spain*
[3]*Center for Research and Advanced Studies, National Polytechnic Institute, Mexico City, Mexico*

Correspondence should be addressed to Rafael Morales-Herrera; rafael.morales@uclm.es

Received 9 January 2017; Accepted 9 January 2017; Published 12 March 2017

Control theory is an interdisciplinary branch of engineering and mathematics dealing with the behavior of dynamic systems with inputs. The objective of control theory is to calculate solutions for the proper corrective action from the controller that results in system stability and improved performance. Automation and Industrial Control Systems (ICS) encompass many applications and uses of industrial and facility control and automation systems. ICS are defined by ISA-99/IEC 62443 as "a collection of personnel, hardware, and software that can affect or influence the safe, secure, and reliable operation of an industrial process."

Control systems are composed by five main elements: sensors, transducers, transmitters, controllers, and final control elements or actuators. This special issue focuses on sensors and, more concretely, on sensor integration in automation and control systems. Let us remind you that a sensor is defined as a device that converts a physical stimulus into a readable output. The role of a sensor in a control and automation system is to detect and measure some physical effect, providing this information to the control system.

The integration of sensors in control and automation systems has received a great deal of attention from a considerable number of researchers and from the industrial community in the last years. Emphasis is placed on the importance of creating improvements in control and automation systems in order to meet the challenges of developing and refining new applications. These systems have to integrate a variety of sensory information and human knowledge for the sake of efficiently carrying out tasks with or without human intervention.

In fact, the integration of sensors into intelligent devices and systems has increased the capacity to measure, analyze, and aggregate data at a localized level. Autonomous and connected sensors are able to selectively sample and measure many physical properties. Built on the increasing capabilities of fixed-access and wireless networks, smart sensor developments allow the collection of raw data, which is processed into information and conveyed via a network connection.

The concept of sensor integration is close to the sensor fusion term, which is defined as "the art of processing data from multiple sensors with an aim to replicate a physical environment or induce intelligence to control a phenomenon with increased precision and reliability." Sensor fusion or integration is evolving rapidly as the basis of robust control systems that can make sense of imperfect input despite the environment in which it operates. Data from multiple sensors are fused to increase response and accuracy, delivering control systems that until recently could only be theorized, drawing on such techniques as artificial intelligence, pattern recognition, digital signal processing, and statistical estimation. Moreover, recent advances in sensor technology and processing techniques, combined with improved hardware, make real-time fusion of data possible.

This special issue was aimed at exhibiting the latest research achievements, findings, and ideas in the integration of sensors in control and automation systems. The topics faced in this special issue were as follows: sensor systems for control and automation: sensors and sensor networks, intelligent sensors, sensor uncertainty for fault tolerant control, distributed and multimodality sensor network for control

and automation, and so on; control: adaptive control, robust control, active disturbance rejection control, complex systems, identification and estimation, nonlinear systems, intelligent systems, sensor networks, delay systems, precision motion control, control applications, and so on; automation: man-machine interactions, process automation, network-based systems, intelligent automation, planning, scheduling and coordination, and so on; robotics: modelling and identification, mobile robotics, mobile sensor networks, perception systems, visual servoing, robot sensing and data fusion, and so on; process based control: sensor development, system design, and control development; control and automation systems: fault detection and isolation, sensing and data fusion, flight control and surveillance systems, rescue and field robotics, guidance control systems, industry, military, space and underwater applications, linear and nonlinear control systems, signal and image processing, and so on; industrial informatics: embedded systems for monitoring and controlling.

Acknowledgments

We would like to thank all the authors for their excellent contributions and also the reviewers for their valuable help. We would also like to thank all members of the editorial board for approving this special issue.

Rafael Morales-Herrera
Antonio Fernández-Caballero
José A. Somolinos
Hebertt Sira-Ramírez

Hindawi
Journal of Sensors
Volume 2017, Article ID 6731204, 14 pages
http://dx.doi.org/10.1155/2017/6731204

Research Article

EM-Based High Speed Wireless Sensor Networks for Underwater Surveillance and Target Tracking

Kumudu Munasinghe,[1] Mohammed Aseeri,[2] Sultan Almorqi,[2] Md. Farhad Hossain,[3] Musbiha Binte Wali,[3] and Abbas Jamalipour[4]

[1]*Faculty of Education, Science, Technology and Mathematics, University of Canberra, Canberra, ACT 261, Australia*
[2]*National Centre for Sensors and Defense Systems Technologies, King Abdulaziz City of Science and Technology, Riyadh, Saudi Arabia*
[3]*Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh*
[4]*School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia*

Correspondence should be addressed to Kumudu Munasinghe; kumudu.munasinghe@canberra.edu.au

Received 17 June 2016; Revised 16 November 2016; Accepted 14 December 2016; Published 26 February 2017

Academic Editor: José A. Somolinos

Underwater Wireless Sensor Networks (UWSNs) are considered as tangible, low cost solution for underwater surveillance and exploration. Existing acoustic wave-based UWSN systems fail to meet the growing demand for fast data rates required in military operations, oil/gas exploration, and oceanographic data collection. Electromagnetic (EM) wave-based communication systems, on the other hand, have great potential for providing high speed data rates in such scenarios. This paper will (1) discuss the challenges faced in the utilization of EM waves for the design of tactical underwater surveillance systems and (2) evaluate several EM wave-based three-dimensional (3D) UWSN architectures differing in topologies and/or operation principles on the performance of localization and target tracking. To the best of our knowledge, this is the first of its kind in the field of underwater communications where underwater surveillance techniques for EM wave-based high speed UWSNs have been investigated. Thus, this will be a major step towards achieving future high speed UWSNs.

1. Introduction

In terms of underwater surveillance, Underwater Wireless Sensor Networks (UWSNs) are considered as a tangible, low cost solution [1, 2]. In these networks, sensor nodes are deployed at various depths in underwater and communicate with other networked floating nodes (e.g., buoys) on the surface and other communications equipment installed in maritime and airborne vehicles (e.g., ships, aircraft, and satellites) [3, 4]. Most of today's underwater surveillance systems are equipped with sonar-array based target tracking algorithms [5–8]. Sonar arrays are based on acoustic wave technology since they are capable of providing long-range communications in underwater [9, 10]. Acoustic waves however result in poor performance in shallow water environments and have extremely low data rates [11] and therefore deemed impractical for on-demand real-time target tracking applications. Moreover, acoustic transmission is affected by multipath propagation, susceptibility to environmental noise, turbidity, salinity gradients, pressure gradients, and adverse impact on marine life. Therefore, electromagnetic (EM) transmissions have been considered as a better alternative for UWSNs [12]. Despite having a relatively shorter range, EM technology is a promising technology for UWSNs as they have the ability to provide much higher data rates than those achievable with acoustic waves in harsh environments with no direct path. This new breed of UWSNs can provide real-time deep-sea oil and gas explorations, military surveillance, search and rescue operations, and environmental monitoring. A comparison on the advantages and disadvantages of acoustic and EM wave-based communications is presented in Table 1.

By and large, underwater surveillance systems used within a military context have three main characteristics. The three key characteristics are detection, identification, and tracking submerged targets (localization). Target detection is how the network identifies a potential target within its

Table 1: Advantages, disadvantages, and challenges of underwater networking using acoustic and EM waves [12–14].

Particulars	Acoustic wave	EM wave
Advantages	(i) Significantly lower signal attenuation (ii) Longer transmission area in the range of km (iii) Can function in the absence of line-of-sight (LOS) path between transmitting and receiving nodes	(i) Large bandwidth (ii) High data rates in the range of few Mbps (iii) Faster response due to higher propagation speed and significantly lower delay (iv) LOS for communication is not essential (v) No need of clear water (vi) No noticeable impact of underwater environment, such as temperature, turbidity, salinity, bubbles, and pressure gradients and thus improving robustness in unpredictable underwater environment (vii) Not affected by sediments and aeration (viii) Immune to other noise except electromagnetic interference (EMI) (ix) Lower Doppler shift (x) More reliable communication (xi) Can cross water-to-air or water-to-earth boundaries easily (xii) No impact on marine life (xiii) Lower cost of nodes (xiv) Good performance in shallow water (xv) Higher attenuation is beneficial in an environment of multiuser interference
Disadvantages	(i) Significantly slower response as propagation speed is much lower (1500 m/s) than that of EM wave (ii) Significantly lower data rate (up to 20 kbps) as bandwidth is low (iii) Surface repeater is required as strong reflections and attenuation occurs in crossing water/air boundary (iv) Poor performance in shallow water (v) Less reliable and robust communication as easily affected by turbidity, ambient noise, temperature, salinity, and pressure gradients (vi) Adverse impact on the marine life and ecosystem (vii) Higher cost of network nodes	(i) Easy to be affected by EMI (ii) Higher attenuation, which increases with the salinity of water (iii) Limited communication range in high data rate applications (e.g., 50 m for 150 kbps and less than 10 m for Mbps range) (iv) Dense deployment of nodes is required for higher frequency range
Challenges	(i) Higher and variable latency (ii) Difficult time synchronization due to variable delay (iii) Higher bit error rate (iv) Multipath propagation and fading (v) Easy signal corruption due to ambient noise (vi) Mobility of nodes	(i) Timing synchronization is difficult as the symbol duration is smaller for higher data rates (ii) Multipath propagation and fading (iii) Mobility of nodes

vicinity. Target identification is the classification of the aforementioned target (i.e., submarines, divers, naval mines, sea animals, etc.). Once a potential target is identified, localization techniques are required for tracking of its movements. UWSN localization is a very challenging problem due to the unavailability of the Global Positioning System (GPS) underwater. Energy efficiency and access delay are two major constraints of any UWSN [13–15]. On the other hand, high throughput with low packet collisions is desirable for modern applications [12]. Fair resource allocation among multiple

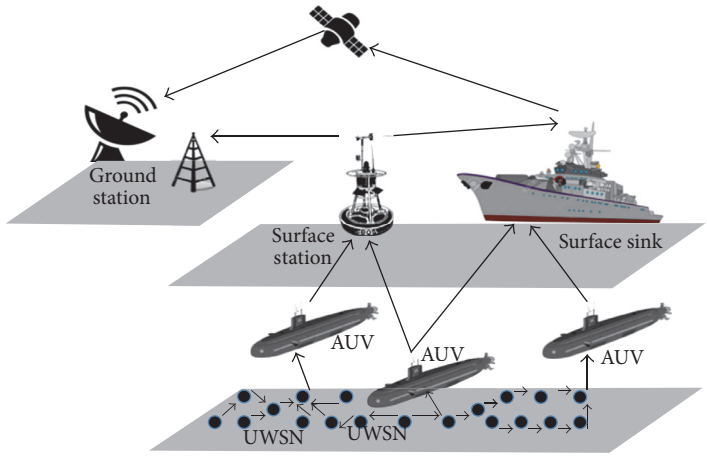


Figure 1: FORCEnet project from the US Navy's unmanned undersea vehicles (UUV) program office.

nodes is another critical factor for effective operation of UWSNs [10]. In addition, the sensor node topology (node distribution) and network architecture play a key role in the design process of these surveillance techniques [13]. For example, node distribution relates to how the sensor nodes are distributed with respect to the characteristics of a particular target, which also needs attention during the UWSN design process.

Our motivation is driven by the fact that existing surveillance systems designed for terrestrial WSNs are inappropriate for EM-based UWSNs due to the fundamental differences between characteristics of the two mediums. Furthermore, due to the marked differences between the propagation characteristics of acoustic and EM waves, most existing underwater surveillance systems that are based on acoustic technology will not be applicable for EM wave-based communications. It is therefore extremely rare to find existing works on underwater surveillance systems for EM wave-based UWSNs. Thus, it is of utmost significance to divert considerable research effort in developing underwater surveillance systems suitable for EM wave-based high speed UWNSs, which is the foremost objective of this paper. This paper investigates (1) the opportunities and challenges of using EM wave-based UWSNs supporting high speed data transmissions with a particular focus on the design of suitable tactical underwater surveillance systems and (2) evaluates several EM wave-based three-dimensional (3D) UWSN architectures on the performance of localization and target tracking. To the best of our knowledge, development of appropriate underwater surveillance techniques for EM wave-based high speed UWSNs will be the first of its kind in the field of underwater communications.

The remainder of this paper is organized as follows. Next we present a literature review that discussed the challenges encountered in the design of underwater surveillance system with particular emphasis to UWSNs. Followed by this we present the system design considerations that need attention for such a design. Next we introduce the EM wave-based three-dimensional UWSN architectures for localization and target tracking followed by a discussion and conclusion.

2. Underwater Surveillance Systems and Challenges

Approximately 70% of the surface of the earth is covered by water. Further 97% of the aforementioned is seawater [16]. Due to the lack of efficient underwater information collecting networks, this vast area of underwater world, which is in abundance of extremely rich natural resources, has hardly been explored. Furthermore, military and political tensions between nations also call for efficient underwater surveillance systems for maritime boundary protection [16]. Issues to be considered with surveillance systems are target detection, localization, classification, and tracking. Traditionally, sonar-array based systems were used for underwater target tracking. As a result, a number of sonar-array systems have been designed for this purpose [9, 16–18]. When sonar-array equipment is submerged and dragged by some sort of vessels (e.g., ship and submarine), they become unsuitable for on-demand tracking missions [19, 20]. Further, if the platform, which it tows, breaks down, the entire system fails [21, 22].

2.1. WSNs for Underwater Surveillance. In order to avoid the issues of sonar-array based target tracking mechanisms, UWSNs have been proposed as an alternative solution. In terms of underwater target tracking, UWSNs seem to offer a promising approach. Low cost, rapid deployment, self-organization, and fault tolerance are the main advantages of UWSNs [23]. As a result, there has been growing interest in research and development in using UWSN as a tangible, low cost solution [13]. UWSNs typically include a large number of intercommunicating underwater devices such as sensors, buoys, gateways, sinks, anchors, and autonomous underwater vehicles (AUVs), which are coordinated for carrying out certain tasks in a collaborative manner. Such networks are often integrated with water surface sinks and stations, submarines, satellite networks, aviation systems, and onshore base stations (sinks) enabling extended functionalities [15, 16]. FORCEnet project of US Navy, as illustrated in Figure 1, is one example of such systems, where an artist's conception of their system can be found in [2].

2.2. Acoustic UWSN Based Surveillance Techniques. Some effort has been made for acoustic wave-based UWSN based underwater target tracking. For example, location estimation is discussed in [24]. To detect underwater target size, a maximum likelihood estimation algorithm is proposed in [25] but, however, lacks a tracking mechanism. In [26], two tracking schemes based on the distributed particle filter have been proposed for cluster based UWSNs.

The biggest downside of these tracking schemes is that they only considered two dimensions. This imposes severe limitations for applications. In response to the above, a 3D target tracking scheme with the combination of interacting multiple models is designed to solve the nonlinear and manoeuvring problems [27]. However, this scheme does not consider the energy consumption problem, which is also a limitation for practical applications. In response, Isik and Akan designed a 3D target tracking solution [28]. The arrival time of these echo messages coming from the target is used for determining the distance from the sensor to the target. Then trilateration is utilized to calculate target's position. Node's position and the above velocity will then be used for tracking the underwater subject. In [29], the waking-up sleep mechanism is utilized to save energy consumption. Despite the aforementioned advances, all of these are based on acoustic UWSNs. Current underwater acoustic communications technologies suffer from serious challenges such as susceptibility to environmental noise and require expensive signal processing to deal with the multipath acoustic channel [30].

2.3. Motivations for EM-Based UWSN Surveillance Techniques. As discussed in Table 1, due to the behaviour of acoustic waves under water, it can be justified that acoustic UWSNs are inappropriate for most modern day underwater applications. Further, an extensive survey conducted by the authors has indicated that there is no complete published works on underwater surveillance systems specifically designed for EM-based UWSNs. Due to aforementioned drawbacks of current underwater acoustic communication technologies, very recent works have proposed EM-based UWSNs as a cost-effective and reliable way forward [12].

To further justify, the advantages of EM waves over acoustic waves for UWSNs could be stated as follows. Firstly, the relatively higher channel bandwidth and data rates (up to 100 Mbps) of EM UWSNs are a clear benefit over relatively lower bandwidth and data rates of acoustic UWSNs (up to 20 kbps). Secondly, the relatively higher propagation speeds would give the EM UWSNs the capabilities such as fast detection, instantaneous tracking, and quick countermeasuring. Thirdly, unlike acoustic UWSNs, EM UWSNs are unaffected by temperature, salinity, turbidity, pressure gradients, and wind speed of the sea. Further, EM UWSNs are highly susceptible to various sources of acoustic noise (e.g., marine life at the seabed and wind speed). In addition, EM UWSNs outperform acoustic UWSNs with its capabilities of nonline of sight operation (e.g., unaffected by aeration and sediments at the seabed). Moreover, EM wave suffers less attenuation in shallow water enabling longer range UWSNs, whereas the seeming drawback of higher attenuation of EM wave in deep water can be exploited in a beneficial way for multiuser parallel data transmission enabling localized communications in UWSNs. Further, the relatively lower cost of RF nodes will further add to the aforementioned reliability making EM UWSNs a clear winner. Lastly, EM UWSNs have no known impact on the marine life and ecosystem.

Nevertheless, the biggest challenge of using EM-based radios underwater is its limited communication range due to high attenuation in water. Therefore, a relatively short communication range and a relatively large number of nodes are required to provide connectivity across large areas. Consequently, optimal node placement algorithms have to be developed. Nevertheless, since the cost of EM-based radios is significantly lower than that of acoustic-based radios, the cost of high density EM-based WSNS should not be an issue. On the other hand, because of the fundamental differences between propagation characteristics of EM wave on terrestrial and underwater channels as well as the topological differences of wireless sensor networks, it is apparent that the surveillance protocols designed for terrestrial WSNs will not operate effectively in underwater environment. Similarly, the absolute differences in generation, propagation, and detection of acoustic and EM waves clearly rule out the appropriateness of the existing acoustic surveillance techniques from being used in EM UWSNs. Some of the challenges, other than those inherent properties, which cannot be compensated by developing sophisticated schemes, of underwater communications for both acoustic and EM waves are also summarized in Table 1.

3. System Design Considerations

According to our comprehensive literature survey, we have found out considerable research challenges such as EM wave propagation behaviour in underwater environments, underwater channel models, physical and chemical properties of the underwater environment, water dynamics, geological distribution of seabed, and other factors influencing the UWSN performance [12–14]. Knowledge on these underwater propagation characteristics of EM waves can be directly used in the development process. Therefore, the new challenges identified in this paper are novel surveillance techniques for underwater EM-based UWSNs and the formulation of analytical models as well as development of simulation platforms for evaluating the performance of the proposed techniques. Other challenges identified are theoretical analysis to investigate the performance of the proposed surveillance technique and algorithms. Simulation tools including MATLAB, NS2, and OPNET can be used to verify the theoretical analysis results. Aqua-Sim, an NS2 based network simulator specifically designed for UWSNs by the University of Connecticut, can also be used for performance evaluation [31].

3.1. Node Topology Design Considerations. Node topology has a great effect on the performance of any acoustic UWSNs [32, 33]. Thus, it can be inferred that underwater node/target localization as well as target tracking performance with EM communication will also be highly dependent on the node topology. Firstly, the design objectives of UWSN node topologies for mobile UWSNs would greatly differ between

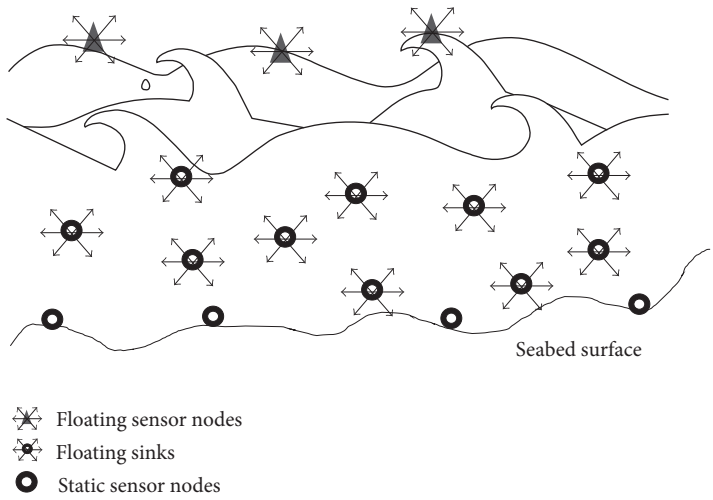


Figure 2: Cyclic and irregular mobility patterns of nodes in an UWSN.

non-time-critical long-term (say, aquatic species monitoring) and time-critical short-term (say, submarine detection) applications. Secondly, if the deployed field is not under the designer's control or random deployment of the sensors is more feasible, stochastic deployment may be preferred [34]. In the latter case, an optimization problem on sensor deployment needs to be formulated to provide sufficient grid coverage of the sensor field such that maximum coverage may be achieved.

Due to the ocean current, UWSN nodes may move at speeds up to six kilometres an hour in a typical underwater scenario [3]. Hence, unlike most terrestrial sensor networks, where sensor nodes are mostly static, most sensor nodes placed underwater have slow to medium mobility. Consequently, any surveillance techniques designed by ignoring the mobility of sensor nodes may perform suboptimally when directly integrated into mobile UWSNs. It is important to note that the mobility models must be developed. This is due to the fact that the node mobility pattern in an UWSN is completely different from those usually considered in the above ground wireless sensor networks literature. The new mobility models have to be 3D in nature because of the cyclic or irregular patterns in forward and backward ways of ocean waves, as illustrated in Figure 2, something that is not the case in terrestrial networks.

Finally, the optimal architecture has to be of distributed types that takes the node depth into consideration with respect to the surface buoys. This way, the 3D coverage of the UWSN can be guaranteed. Therefore, the solutions should use adaptive node topology capable of adjusting the depth of sensor nodes in the event they drift by currents, winds, and so on.

3.2. Target Detection Considerations. Target detection deals with how the network detects or recognizes the presence of a mobile target in the proximity. As per the literature, a number of efforts have been made for underwater target detection, where most mechanisms have been based on the classical Doppler equation [35]. The primary motive behind using Doppler equation based techniques is due to the fact that most UWSNs were based on acoustic wave-based sensors. Further, the accuracy of the Doppler equation can only be guaranteed when the target moves to or from the listener, which means that there is a degree of inaccuracy involved in this method. Also, the node placement and the target characteristics must be known, which cannot be done for an underwater stochastically deployed environment. Therefore, due to the previously mentioned deficiencies of acoustic wave-based systems there are a number of prospects in using EM-based UWSNs.

In the case of stochastically deployed EM-based UWSNs, target detection can be achieved probabilistically. By the evaluation of the detection probability of at least k sensors, the success of the target detection can be measured [34]. However, in more realistic scenarios, a sensor needs to collect multiple samples of the target before it can perform reliable detection [36, 37]. Hence, a sensor s must sample the target X for at least t units of time, before s can reliably determine the presence of X. The most important point to be considered is that, since [34–37] are aboveground scenarios, underwater propagation characteristics of EM waves have to be taken into consideration in the development process.

3.3. Target Classification Considerations. Classification deals with how the network classifies the target type. Under this method, a classification-based data mining scheme can be implemented, where the system can classify the submerged targets as submarines, mines, divers, and sea animals. This is achieved with a combination of sensors based on radiation, mechanical, magnetic, thermal, and chemical signaling. In this particular case, the method that can be used for classification of targets is decision trees. The reason behind choosing decision trees is due to its fast executing, scalability, and ease of interpreting characteristics [38]. In order to use collaborative computing, classification-based data mining combined with the previous decision tree to detect and classify a target can be used. Each target type can be studied for defining the possible set of data values that a sensor could be measuring in the vicinity of a target. Next the challenge of

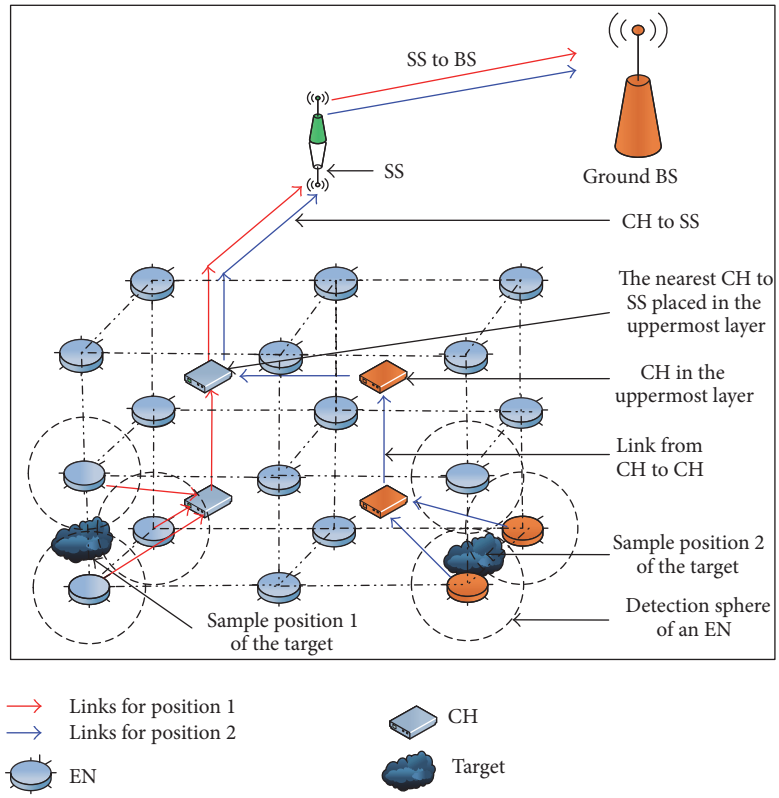


Figure 3: A general view of the proposed underwater target detection and tracking system.

preparation of the aforementioned data set for a classification mining based detection algorithm is involved.

3.4. Target Tracking Considerations. Once a target has been detected and classified, localization deals with sustained target tracking. As mentioned under *Node Topology Design Considerations*, a distributed tracking architecture is the preferred choice for UWSNs. In a potential distributed tacking architecture, the processing node must be chosen near the target and a subset of sensors in its vicinity are chosen as sensor data collection nodes. The moving target tracking problem can be handled as a multisensor data fusion problem. Under this, measurements from various sensors are combined and pull all information together as one coherent structure. Interacting multiple model (IMM) filters in this collaborative manoeuvring target tracking problem can also be used. Based on the received data, the IMM filter is capable of achieving its movement detection functionality by updating the mode probabilities making it advantageous over simpler estimators like Kalman filter [39]. Next the distributed IMM filter is used for estimating states of a target on a given sensor platform (assuming the target moves along a variety of trajectories). The distributed IMM filter then can combine different models as per the target's motion characteristics such that it will adapt to any change of trajectory and then the probability for each model can be calculated. Finally the complete target tracking model can be combined with an adaptive sensor selection scheme and an appropriate sleep/wake model for ensuring improved estimation and energy-efficient performance.

4. EM Wave-Based Three-Dimensional UWSN Architectures for Localization and Target Tracking

4.1. Basic Network Layout. We consider a 3D target surveillance area of dimension $D_X \times D_Y \times D_Z$, where D_X, D_Y, and D_Z are the length of the surveillance area along the X-, Y-, and Z-axis, respectively. For the convenience, we assume a plane seabed. The X-Y plane of the area is assumed parallel to the seabed and the Z-axis is along the depth of the sea. We propose grid-based network topologies by dividing the entire network into $N_X = D_X/\Delta_X$, $N_Y = D_Y/\Delta_Y$, and $N_Z = D_Z/\Delta_Z$ segments along the three axes respectively, where Δ_X, Δ_Y, and Δ_Z are the corresponding segment lengths. The proposed UWSN topologies consist of two basic elements: the elementary nodes (ENs) which are essentially sensor nodes and the cluster heads (CHs). The general view of the overall target detection and tracking scheme as illustrated in Figure 3 consists of a UWSN, a surface sink (SS), and a ground base station (BS). When a target enters the surveillance area, some ENs sense the presence of the target and broadcast their location information to the surrounding CHs using EM wave. Upon receiving EM signal from ENs, nearby CHs collect different information and then forward the selected information to the respective CHs of immediate upper layer and so on. The information is subsequently routed to the CH nearest to the SS and finally to the SS. The SS then transmits the gathered information to the ground BS for processing and estimating the location and travelling path of the target. All the links

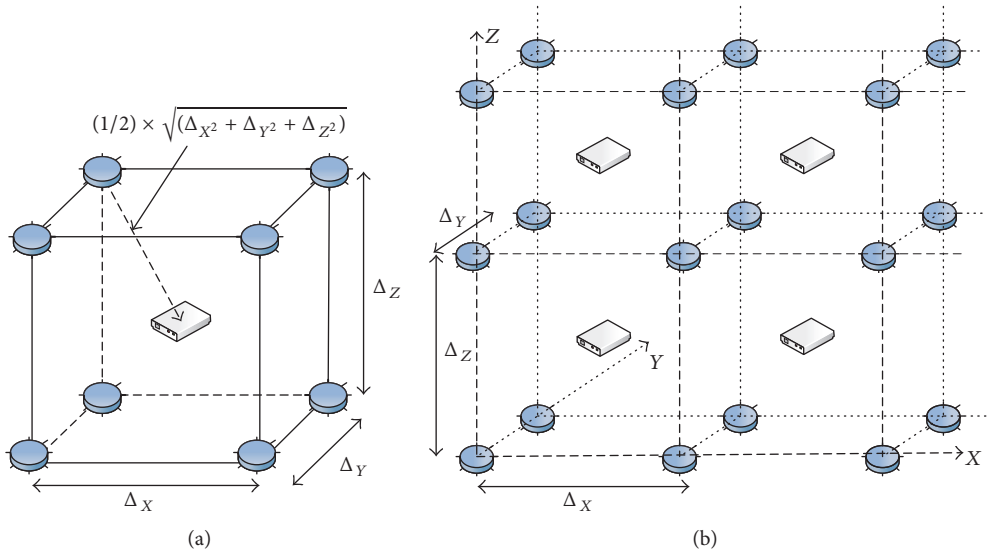


Figure 4: Proposed network topology for architecture A_1 and A_2: (a) basic building block and (b) UWSN.

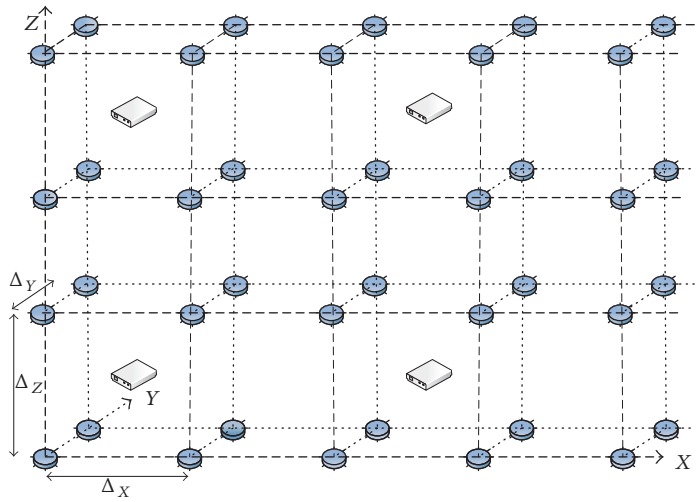


Figure 5: Proposed UWSN with architecture A_3 and A_5.

including EN-CH, CH-CH, CH-SS, and SS-BS are proposed to be EM wave-based.

4.2. Architectures. Under the proposed architectures, ENs and CHs are arranged in three different formations in 3D space evolving to three distinct basic building blocks of the UWSN. The basic building blocks of the topologies are repeated in 3D space to form the entire UWSNs covering the surveillance area of dimension $D_X \times D_Y \times D_Z$. The resulting three UWSNs using three distinct topologies are shown in Figures 4–6. The basic building block is shown only in Figure 4, which can similarly be drawn for the networks in Figures 5 and 6. Now, based on these three distinct node topologies and different location estimation algorithms, we propose five different architectures denoted as $A_1, A_2, \ldots, A_5$ as presented below.

4.2.1. Architectures A_1 and A_2. The topology of the architectures A_1 and A_2 are same. However, the location estimation principles have significant difference. As shown in Figure 4(a), ENs are placed at all the eight vertices of a rectangular cuboid of dimension $\Delta_X \times \Delta_Y \times \Delta_Z$. Thus, the entire UWSN is formed by placing ENs every Δ_X, Δ_Y, and Δ_Z distance along X-, Y-, and Z-axis, respectively. On the other hand, one CH is placed at the centre of each cuboid. The UWSN thus has eight ENs surrounding each CH and vice versa. Thus the total number of CHs, N_{CH} and ENs, N_{EN} can be written as $N_{CH} = N_X N_Y N_Z$ and $N_{EN} = (N_X + 1)(N_Y + 1)(N_Z + 1)$.

Now, for architecture A_1, we assume that each CH knows its own 3D coordinate. When an EN senses the presence of a target, it alerts the surrounding CHs by transmitting signal using EM wave. When a CH receives this EM signal with a power above a certain threshold P_{th}, it transmits its 3D coordinate to the immediate upper CH using EM wave. Through

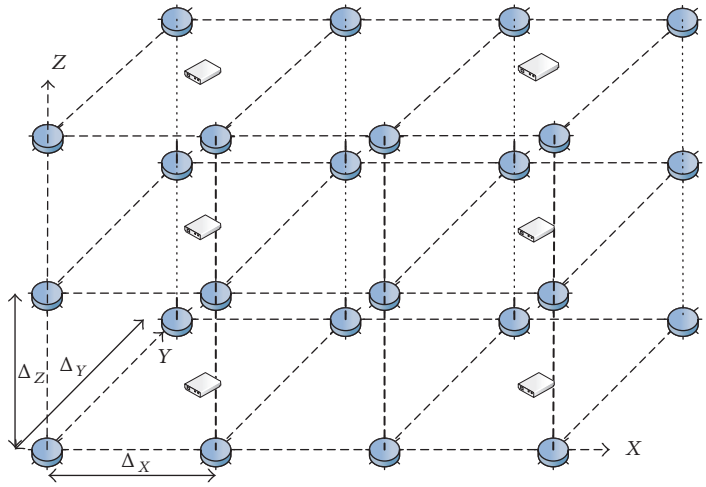


Figure 6: Proposed UWSN with architecture A_4.

this flow of location data from one layer of CHs to those of the upper layers, the coordinates of the CHs nearest to the sensing ENs are known to the BS. Then the ground BS estimates the 3D coordinate $(x_{est}, y_{est}, z_{est})$ of the target as follows:

$$x_{est} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$
$$y_{est} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad (1)$$
$$z_{est} = \frac{1}{N}\sum_{i=1}^{N} z_i,$$

where (x_i, y_i, z_i) is the 3D coordinate of the ith CH and N is the number of CHs from which the ground BS has received information.

On the other hand, architecture A_2 utilizes both the location information of the CHs sending the target presence information to the ground BS as well as the EM signal power strength received from the surrounding ENs. The first step of estimating the target location is similar to architecture A_1. Thus, an initial estimate of the 3D target location denoted by $(x_{int}, y_{int}, z_{int})$ is evaluated using (1). Then, by using the information of the received power at CHs from the surrounding ENs, the initially estimated location $(x_{int}, y_{int}, z_{int})$ is fine-tuned to achieve a new coordinate, which can be given by

$$x_{est} = \frac{1}{N}\sum_{i=1}^{N} \frac{(P_{max} - P_i)\, x_{int} - (P_{int} - P_i)\, x_i}{P_{max} - P_{int}},$$
$$y_{est} = \frac{1}{N}\sum_{i=1}^{N} \frac{(P_{max} - P_i)\, y_{int} - (P_{int} - P_i)\, y_i}{P_{max} - P_{int}}, \quad (2)$$
$$z_{est} = \frac{1}{N}\sum_{i=1}^{N} \frac{(P_{max} - P_i)\, z_{int} - (P_{int} - P_i)\, z_i}{P_{max} - P_{int}},$$

where P_i is the actual EM power received at ith CH; P_{max} and P_{int} are the received power at ith CH if ENs were located at the location of CH and $(x_{int}, y_{int}, z_{int})$, respectively. All the values of power used in (2) are given in dBm. Equation (2) is developed based on the fact that if we ignore fading, then the received power expressed in dBm decreases linearly with distance between the transmitter and the receiver.

4.2.2. Architectures A_3 and A_5. The topologies of architectures A_3 and A_5 (Figure 5) differ from those of architectures A_1 and A_2 (Figure 4) in the number and the locations of CHs. As illustrated for architectures A_3 and A_5, CHs are placed at the centre of every other rectangular cuboid of dimension $\Delta_X \times \Delta_Y \times \Delta_Z$ along all the three axes. Thus CHs are placed every $2\Delta_X$, $2\Delta_Y$, and $2\Delta_Z$ distances along X-, Y-, and Z-axis, respectively. Hence, $N_{CH} = \lceil N_X/2 \rceil \lceil N_Y/2 \rceil \lceil N_Z/2 \rceil$ and $N_{EN} = (N_X + 1)(N_Y + 1)(N_Z + 1)$. Here $\lceil x \rceil$ implies the ceiling operation and is equal to the smallest integer equal to or greater than x.

Now, similar to architecture A_1, architecture A_3 estimates the 3D location of a target from the location of the active CHs by using (1). On the other hand, although the topology of architecture A_5 is the same as that of A_3, A_5 integrates additional feature into the CHs. It is assumed that each CH is equipped with eight directional receivers directed to the eight ENs surrounding it. So when a CH receives signal from surrounding ENs, it sorts out the strongest EM signal and the location of the corresponding EN from which the signal is received. Unlike A_1–A_3, instead of its own location, a CH then transmits the location of this EN of the strongest EM signal to the CH of its upper layer and so on until it is received by the BS. The BS then estimates the location of the target using (1) replacing x_i, y_i, and z_i by the coordinate of ith EN contributing the strongest EM signal.

4.2.3. Architecture A_4. In the topology of architecture A_4 as shown in Figure 6, though the placement of ENs is the same as all other architectures, CHs are positioned in a very different way. Instead of placing CHs at the centre of cuboids, they are deployed on the Z-planes at the centre of four coplanar ENs. Thus, ENs are placed at regular intervals of $2\Delta_X$, $2\Delta_Y$, and

Table 2: A summary of the key features of the proposed architectures.

Architectures	Spacing between two ENs	Spacing between two CHs	Total number of ENs and CHs	Directional receiving antennas in CHs	Information used for location estimation
A_1	$\Delta_X, \Delta_Y, \Delta_Z$	$\Delta_X, \Delta_Y, \Delta_Z$	$N_{CH} = N_X N_Y N_Z$ $N_{EN} = (N_X+1)(N_Y+1)(N_Z+1)$	No	Location of CHs
A_2	$\Delta_X, \Delta_Y, \Delta_Z$	$\Delta_X, \Delta_Y, \Delta_Z$	$N_{CH} = N_X N_Y N_Z$ $N_{EN} = (N_X+1)(N_Y+1)(N_Z+1)$	No	Location of CHs and EM power received at CHs
A_3	$\Delta_X, \Delta_Y, \Delta_Z$	$2\Delta_X, 2\Delta_Y, 2\Delta_Z$	$N_{CH} = \lceil \frac{N_X}{2} \rceil \lceil \frac{N_Y}{2} \rceil \lceil \frac{N_Z}{2} \rceil$ $N_{EN} = (N_X+1)(N_Y+1)(N_Z+1)$	No	Location of CHs
A_4	$\Delta_X, \Delta_Y, \Delta_Z$	$2\Delta_X, 2\Delta_Y, \Delta_Z$	$N_{CH} = \lceil \frac{N_X}{2} \rceil \lceil \frac{N_Y}{2} \rceil (N_Z+1)$ $N_{EN} = (N_X+1)(N_Y+1)(N_Z+1)$	Four directional antennas at each CH	Location of ENs contributing the strongest EM signal
A_5	$\Delta_X, \Delta_Y, \Delta_Z$	$2\Delta_X, 2\Delta_Y, 2\Delta_Z$	$N_{CH} = \lceil \frac{N_X}{2} \rceil \lceil \frac{N_Y}{2} \rceil \lceil \frac{N_Z}{2} \rceil$ $N_{EN} = (N_X+1)(N_Y+1)(N_Z+1)$	Eight directional antennas at each CH	Location of ENs contributing the strongest EM signal

Δ_Z distance along X-, Y-, and Z-axis, respectively, as shown in the figure. Hence, $N_{CH} = \lceil N_X/2 \rceil \lceil N_Y/2 \rceil (N_Z+1)$ and N_{EN} is same as that of the previous architectures. Furthermore, each CH is equipped with four directional receivers directed to the four coplanar ENs surrounding it for determining the EN from which the strongest EM signal is received. Similar to architecture A_5, the coordinate of the corresponding EN of the strongest EM signal is subsequently received by the BS and then used to estimate the location of target using (1). Here, (x_i, y_i, z_i) is once again the coordinate of ith EN contributing the strongest EM signal.

For the clarity, Table 2 summarizes the key features of the proposed underwater target detection and tracking architectures.

4.3. Performance Metrics

4.3.1. Error in Location Estimation. If (x, y, z) is the actual location of a target, then the absolute distance r between the actual and the estimated location becomes a metric for location estimation error and can be given by

$$r = \sqrt{(x_{est} - x)^2 + (y_{est} - y)^2 + (z_{est} - z)^2}. \quad (3)$$

Then the normalized mean square error (NMSE) of location estimation defined as the mean of r^2 normalized by the square of the minimum distance between two ENs can be given by

$$\text{NMSE}_D = \frac{\sum_{i=1}^{M} r_i^2}{M \left[\min\left(\Delta_x, \Delta_y, \Delta_z\right)\right]^2}, \quad (4)$$

where r_i is the location estimation error of ith simulations and M is the number of Monte Carlo simulations.

4.3.2. Error in Estimated Distance Travelled. If (x_1, y_1, z_1) and (x_2, y_2, z_2) are two different subsequent actual locations of a moving target and if the corresponding estimated locations of the target are (x_{e1}, y_{e1}, z_{e1}) and (x_{e2}, y_{e2}, z_{e2}), then the actual and the estimated travelled distance denoted by r_a and r_e can be calculated as

$$\begin{aligned} r_a &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \\ r_e &= \sqrt{(x_{e1} - x_{e2})^2 + (y_{e1} - y_{e2})^2 + (z_{e1} - z_{e2})^2}. \end{aligned} \quad (5)$$

Then the NMSE of the estimated travelled distance can be defined as

$$\text{NMSE}_T = \frac{\sum_{i=1}^{M} (r_{i,a} - r_{i,e})^2}{M \left[\min\left(\Delta_x, \Delta_y, \Delta_z\right)\right]^2}, \quad (6)$$

where $r_{i,a}$ and $r_{i,e}$ are the actual and the estimated travelled distance respectively of ith simulations.

4.3.3. Error in Travelled Direction Estimation. Error in estimated horizontal direction of travelling denoted as θ_{XY} and error in estimated vertical direction of travelling denoted as θ_{YZ} can be defined as follows:

$$\begin{aligned} \theta_{XY} &= \tan^{-1}\left(\frac{y_{e1} - y_{e2}}{x_{e1} - x_{e2}}\right) - \tan^{-1}\left(\frac{y_1 - y_2}{x_1 - x_2}\right), \\ \theta_{YZ} &= \tan^{-1}\left(\frac{z_{e1} - z_{e2}}{y_{e1} - y_{e2}}\right) - \tan^{-1}\left(\frac{z_1 - z_2}{y_1 - y_2}\right). \end{aligned} \quad (7)$$

Then the root mean square error (RMSE) in the estimated direction of travelling can be defined as

$$\text{RMSE}_{XY} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\theta_{i,XY}^2}, \quad \text{RMSE}_{YZ} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\theta_{i,YZ}^2}. \tag{8}$$

4.3.4. Propagation Delay. Propagation delay of a communication system can be evaluated by dividing the total distance a signal travel with the propagation speed of the signal. The propagation speed of EM wave in underwater environment can be represented by the following equation [12]:

$$C_w = \sqrt{\frac{f \times 10^7}{\sigma}}, \text{ m/s}, \tag{9}$$

where f is the transmission frequency in Hz and σ is the conductivity of water in S/m. Typical value for $\sigma = 0.01$ S/m for fresh water, which is much higher for seawater due to the higher salinity [12]. On the other hand, speed of acoustic signal in water can be approximately taken equal to 1,500 m/s [12].

4.4. Performance Evaluation. We evaluate the performance of the proposed target detection and tracking architectures using MATLAB based Monte Carlo simulation platform. The results presented in this section are obtained through averaging over $M = 10{,}000$ independent simulations. Performance evaluations are carried out considering only the underwater part of the complete surveillance system, which implies that any information reaching the SS will reach the ground BS as well. The parameters used for the simulations are chosen in reference to various references of RF based underwater communications [12, 40–42]. Without losing the generality, a 3D UWSN with $D_X = D_Y = D_Z = D$ and $\Delta_X = \Delta_Y = \Delta_Z = \Delta$ is considered for simulations. An underwater EM wave propagation model with path loss $20 \times (\log_{10} e) \times 2\pi \times \sqrt{(\sigma \times f \times 10^{-7})}$ dB/m and shadow fading is used [12]. Shadow fading is modeled as log-normally distributed random variable with a mean and standard deviation equal to 0 dB and ζ dB, respectively. Conductivity $\sigma = 4$ S/m is used for emulating a typical salty sea water environment [12, 40]. Transmission frequency equal to 6 kHz and 3 kHz is used for ENs and CHs, respectively [12, 41]. On the other hand, transmit power of EN and CH is assumed equal to 100 mW and 1 W, respectively [42]. Unless otherwise stated, a network with $D = 400$ m, $\Delta = 20$ m, and $P_{\text{th}} = -60$ dBm is considered for the simulations.

4.4.1. Hardware Requirement. Figure 7 compares the architectures in terms of the number of CHs per EN considering an infinitely long UWSN with an infinitely large number of segments along the three axes. The ratios will be smaller for smaller sized networks and thus the figure represents the worst case scenario requiring the maximum number of CHs per EN. Although the number of ENs in the total surveillance area is equal for all the five architectures, CHs are placed more sparsely in A_3, A_4 and A_5 compared to A_1 and A_2. This leads to the lower number of CHs per EN for the former three architectures.

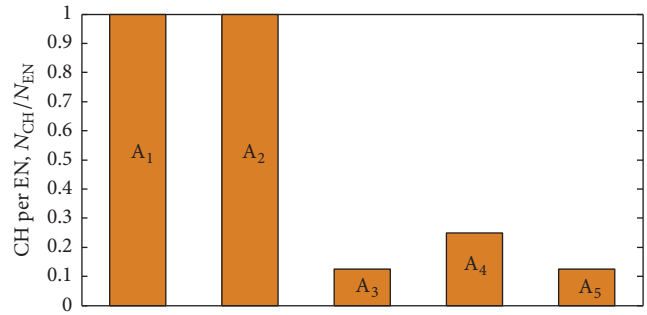


Figure 7: Number of CHs per EN for an infinitely long UWSN with an infinitely large number of segments along all the three axes (i.e., $N_X \to \infty$, $N_Y \to \infty$, and $N_Z \to \infty$).

4.4.2. Accuracy in Localizing Objects. Performance of the target detection architectures in terms of NMSE of the estimated location for two different detection thresholds with and without shadow fading is illustrated in Figure 8. Normalization is done using $\Delta = 20$ m and shadow fading is simulated using $\zeta = 8$ dB. Network length is increased by keeping the relative positions and the number of ENs and CHs the same as those of a UWSN with $D = 400$ m and $\Delta = 20$ m leading to increased distances among the nodes in the networks. Several insights can now be identified by observing Figure 8. *Firstly*, the figure clearly shows increasing trends of NMSE of all the architectures with the increase of the network length. With the increased distance between any two nodes, fewer number of ENs can communicate with the corresponding CHs and fewer CHs can communicate with the CHs in the upper layer leading to reduced accuracy and increased NMSE. *Secondly*, the maximum network size is much smaller (630 m for $P_{\text{th}} = -90$ dBm) for A_3 and A_5 than that for A_1, A_2, and A_4 (1275 m for $P_{\text{th}} = -90$ dBm), which is the direct result of larger distance between adjacent CHs in A_3 and A_5. Furthermore, comparing Figures 8(a)–8(c), it is found that with the increase of detection threshold from −90 dBm to −30 dBm, the maximum network size is reduced from 1280 m to 640 m for A_1, A_2, and A_4 and below 400 m for A_3 and A_5. *Thirdly*, the deteriorating impact of shadow fading on the performance of the architectures in terms of increased NMSE is clearly demonstrated in the figure. However, up to a certain network size, the performance of the architectures with and without fading environment is the same, which is due to the close proximity of network nodes such that the received signal power is above the detection threshold even in the presence of severe fading. Beyond this certain network size (around 1140–1170 m for A_1, A_2, and A_4), the performance gap between the two scenarios increases sharply. Moreover, as seen in Figures 8(a)-8(b), the impact of shadow fading on the performance of A_3 and A_5 is not visible. This is due to the much lower

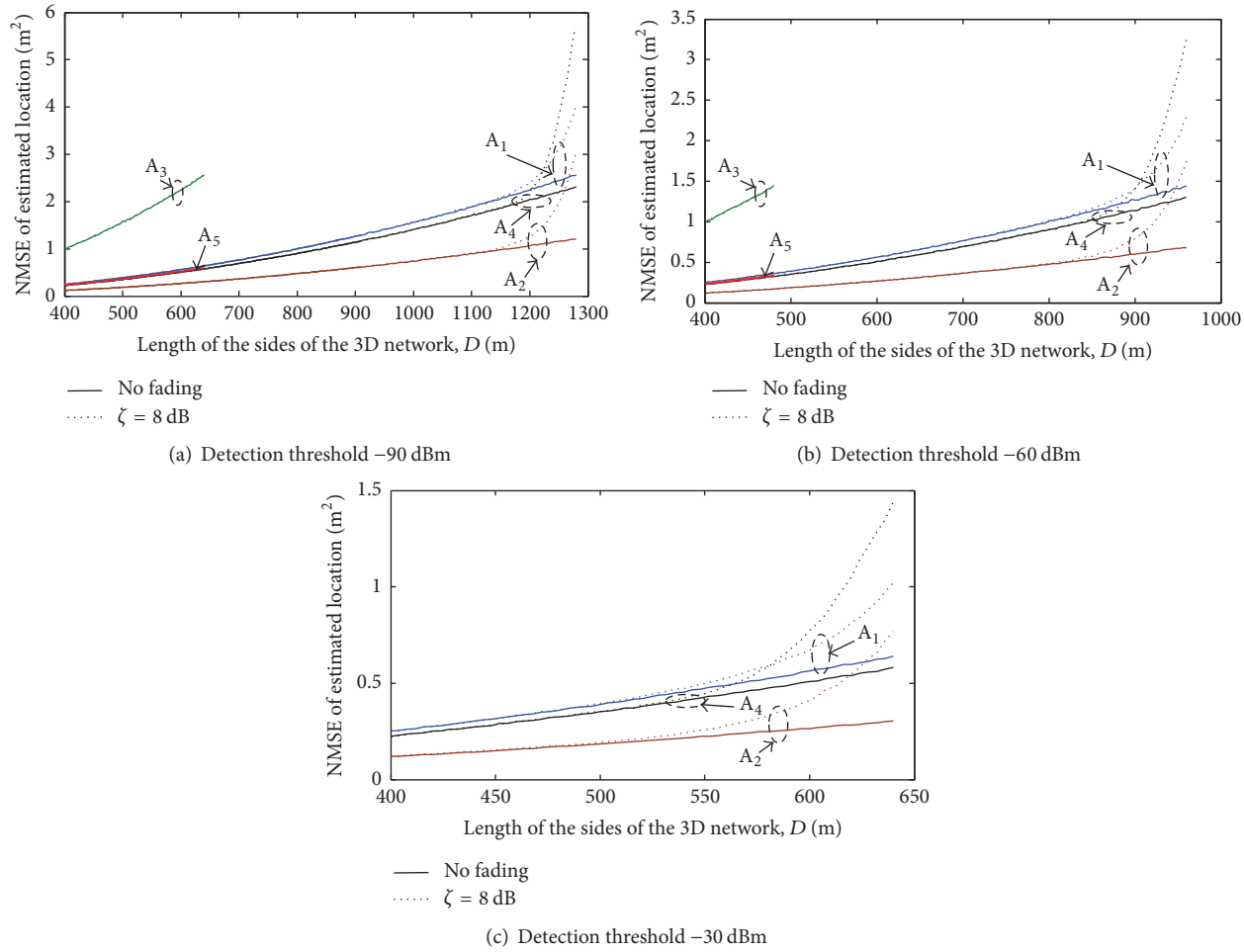
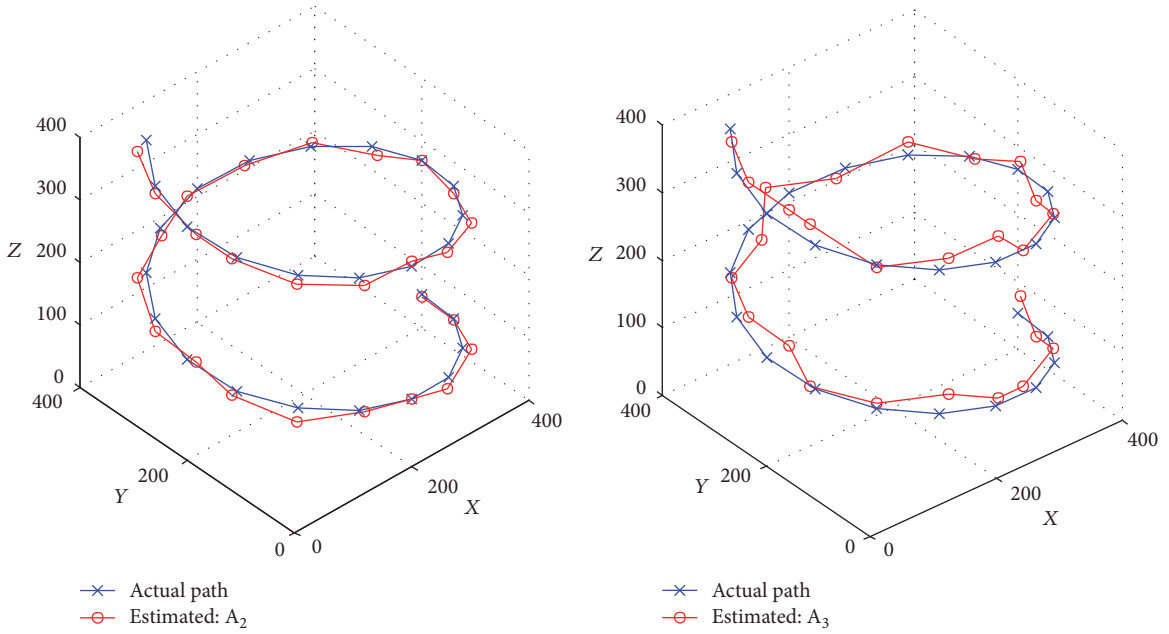


FIGURE 8: NMSE of the estimated location of a target with network size keeping the number of networking nodes unchanged.

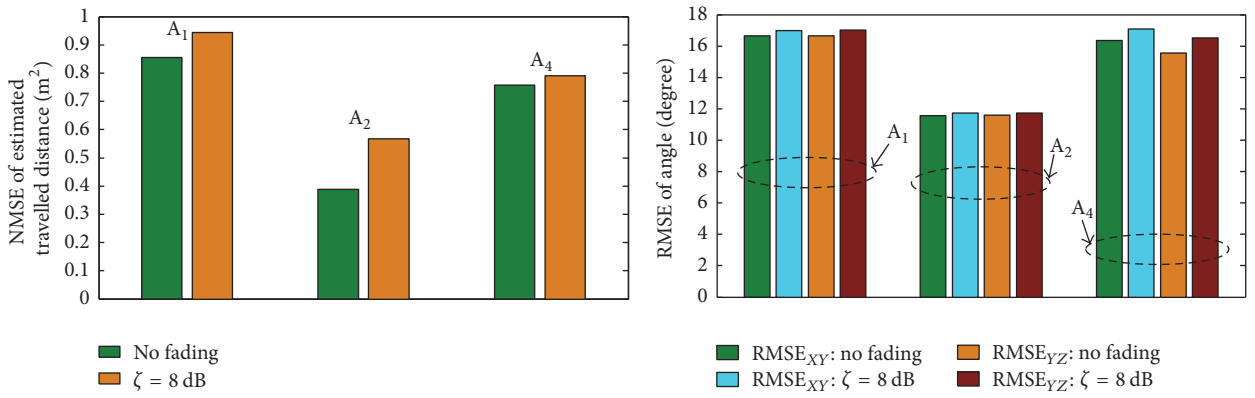
maximum network size (630 m for $P_{th} = -90$ dBm and 475 m for $P_{th} = -60$ dBm) compared to the other architectures up to which shadow fading does not degrade the received power by the network nodes to a value lower than P_{th}. *Finally*, it can readily be identified that the architecture A_3 has the worst accuracy, which is primarily due to the sparse location of CHs, whereas the best accuracy is achieved for A_2. It is also evident that with much fewer CHs, A_4 and A_5 have improved performance than that of A_1. The reason behind this better accuracy is the use of directional antennas integrated in the CHs and the position estimations from the locations of ENs. However, this additional feature may increase the size, computational complexity, and energy consumption in CHs. On the other hand, despite the use of the same topology, the better accuracy of A_2 compared to that of A_1 is directly attributed to the further fine-tuning of the initially estimated location using the information of received power. Thus it can be inferred that if the information of received power is utilized, the accuracy of A_4 and A_5 would be significantly improved, which can be further investigated.

4.4.3. Accuracy in Tracking Objects. On the other hand, for demonstrating the performance of the proposed architectures in tracking moving targets, a sample path of the intruder and the corresponding tracked path by A_2 and A_3 under no fading environment are illustrated in Figures 9(a)-9(b). From the visual inspection, it is clear that architecture A_2 has better accuracy than that of A_3 in tracking the path of the target, which is also supported by Figures 8(a)-8(b). On the other hand, the NMSE of the estimated travelled path and the RMSE of the angles in the XY and YZ planes with and without shadow fading are illustrated in Figure 10. For understanding the impact of shadow fading, the network is configured using $D = 900$ m (correspondingly $\Delta = 45$ m). The figure does not include the results of A_3 and A_5 as 900 m is well above the feasible network size of these two architectures. A detection threshold $P_{th} = -60$ dBm is considered for the simulations. From the figure, A_2 is found to have the best accuracy in both estimating travelled distance and the travelled direction as evident from the plot of NMSE and RMSE, respectively, which is also supported by Figure 8. Shadow fading can significantly reduce the accuracy in travelled distance estimation as evident from Figure 10(a). However, though shadow fading has little impact on the accuracy in estimating the travelled direction in A_4, negligible impact is seen in A_1



(a) Architecture A_2 (b) Architecture A_3

Figure 9: Sample travel path and the estimated path of a moving target with no fading.



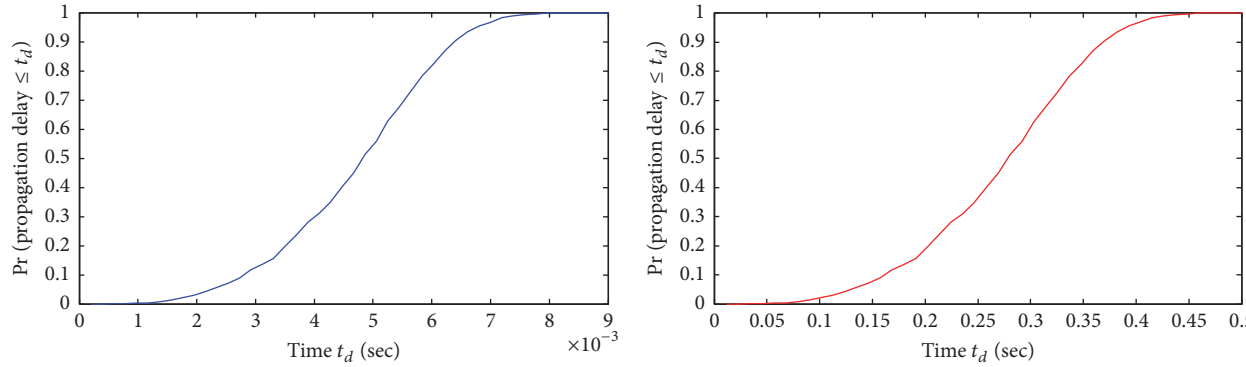
(a) NMSE of the estimation of travelled path (b) RMSE of angles in XY and YZ planes

Figure 10: Performance of the architectures in tracking a moving target with $P_{th} = -60$ dBm.

and A_2 because of the symmetric nature of their network topology.

4.4.4. Response Time. Finally, Figures 11(a)-11(b) present the cumulative distribution function (CDF) of the propagation time of signal from the sensing ENs to the SS for both the EM and acoustic wave-based systems. It is clearly seen that the propagation delay for EM wave-based system is of several orders of magnitude smaller than that in an acoustic wave-based one leading to fast intruder detection. The reason behind this significantly lower navigation delay of EM-based system is the fundamental characteristic of high propagation speed of EM waves compared to that of acoustic waves.

4.4.5. Nodes Mobility. For evaluating the performance of the proposed architectures, we have considered that the nodes are static with respect to each other. If we consider the relative displacement of nodes at their positions, accuracy of the proposed architectures will degrade and maximum allowable network size will decrease. That is, the performance will be affected in a similar way of shadow fading. Nevertheless, the proposed algorithms are equally applicable for mobile node scenario as well. It is worthwhile to mention that the proposed architectures with fixed node positions are also suitable for many practical applications where nodes mobility can be ignored and the relative displacement of the nodes are negligible. Such applications include 3D fence around sea beaches for detecting and tracking sharks, seaports and harbor for surveillance, offshore gas rigs, a moving grid, or fence fixed with a moving vessel. However, mobility of networking nodes is a critical issue for underwater networking, which depends on many factors including the water current patterns, moving



(a) EM wave-based (b) Acoustic wave-based

Figure 11: CDF of propagation delays from EN to SS.

ships and vessels, and movement of underwater objects (e.g., fish, AUV, and submarines). Advanced techniques can be integrated with the proposed systems for overcoming performance degradation due to node mobility, which is left for future works.

5. Conclusions

There is a great demand for highly sophisticated yet economically viable solutions for underwater surveillance and exploration of maritime resources. In terms of underwater target tracking, UWSNs are considered as a tangible, low cost solution. Existing acoustic wave-based UWSN systems fail to meet today's growing demand for fast response and higher data rates. EM wave-based communication systems on the other hand have great potential for providing such requirements. This paper investigates the challenges of using EM wave-based UWSNs and evaluates several EM wave-based UWSN architectures on the performance of localization and target tracking. For nations that border the ocean, the need for faster and smarter underwater communication networks becomes even more critical. For instance, from industry, military, scientific, and environmental points of view, it is extremely vital to have fast, robust, scalable, and adaptive underwater communications. *In our future works, we will focus on developing efficient node deployment strategies with different objectives for EM wave-based UWSNs for underwater target localization and tracking. These strategies will integrate sophisticated techniques for improving network robustness with mobile nodes, multipath propagation, and water dynamics.*

Competing Interests

The authors declare that they have no competing interests.

References

[1] J. Partan, J. Kurose, and B. N. Levine, "A survey of practical issues in underwater networks," in *Proceedings of the 1st ACM International Workshop on Underwater Networks (WUWNet '06)*, pp. 17–24, Los Angeles, Calif, USA, September 2006.

[2] R. Headrick and L. Freitag, "Growth of underwater communication technology in the U.S. Navy," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 80–82, 2009.

[3] J.-H. Cui, J. Kong, M. Gerla, and S. Zhou, "The challenges of building scalable mobile underwater wireless sensor networks for aquatic applications," *IEEE Network*, vol. 20, no. 3, pp. 12–18, 2006.

[4] L. Liu, S. Zhou, and J.-H. Cui, "Prospects and problems of wireless communication for underwater sensor networks," *Wireless Communications and Mobile Computing*, vol. 8, no. 8, pp. 977–994, 2008.

[5] M. Asif, M. Rizal, and A. Yahya, "An active contour for underwater target tracking and navigation," in *Proceedings of the International Conference on Man-Machine Systems*, pp. 1–6, Langkawi Islands, Malaysia, September 2006.

[6] E. Dalberg, A. Lauberts, R. K. Lennartsson, M. J. Levonen, and L. Persson, "Underwater target tracking by means of acoustic and electromagnetic data fusion," in *Proceedings of the 9th International Conference on Information Fusion (FUSION '06)*, July 2006.

[7] M. I. Pettersson, V. Zetterberg, and I. Claesson, "Detection and imaging of moving targets in wideband as using fast time back projection combined with space-time processing," in *Proceedings of the MTS/IEEE Oceans*, pp. 2388–2393, Washington, DC, USA, September 2005.

[8] Q. Zhang, M. Liu, S. Zhang, and H. Chen, "Node topology effect on target tracking based on underwater wireless sensor networks," in *Proceedings of the 17th International Conference on Information Fusion (FUSION '14)*, Salamanca, Spain, July 2014.

[9] S. Al-Dharrab, M. Uysal, and T. Duman, "Cooperative underwater acoustic communications," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 146–153, 2013.

[10] P. Casari and M. Zorzi, "Protocol design issues in underwater acoustic networks," *Computer Communications*, vol. 34, no. 17, pp. 2013–2025, 2011.

[11] J. G. Proakis, E. M. Sozer, J. A. Rice, and M. Stojanovic, "Shallow water acoustic networks," *IEEE Communications Magazine*, vol. 39, no. 11, pp. 114–119, 2001.

[12] X. Che, I. Wells, G. Dickers, P. Kear, and X. Gong, "Re-evaluation of RF electromagnetic communication in underwater sensor networks," *IEEE Communications Magazine*, vol. 48, no. 12, pp. 143–151, 2010.

[13] K. Chen, M. Ma, E. Cheng, F. Yuan, and W. Su, "A survey on MAC protocols for underwater wireless sensor networks," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 3, pp. 1433–1447, 2014.

[14] G. A. Shah, "A survey on medium access control in underwater acoustic sensor networks," in *Proceedings of the International Conference on Advanced Information Networking and Applications Workshops (WAINA '09)*, pp. 1178–1183, May 2009.

[15] I. F. Akyildiz, D. Pompili, and T. Melodia, "State of the art in protocol research for underwater acoustic sensor networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 4, pp. 11–22, 2007.

[16] D. Pompili, T. Melodia, and I. F. Akyildiz, "Deployment analysis in underwater acoustic wireless sensor networks," in *Proceedings of the First ACM International Workshop on Underwater Networks (WUWNet '06)*, pp. 48–55, Los Angeles, Calif, USA, September 2006.

[17] M. Arik and O. B. Akan, "Collaborative mobile target imaging in UWB wireless radar sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 6, pp. 950–961, 2010.

[18] A. O. Bicen, A. B. Sahin, and O. B. Akan, "Spectrum-aware underwater networks: cognitive acoustic communications," *IEEE Vehicular Technology Magazine*, vol. 7, no. 2, pp. 34–40, 2012.

[19] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges," *Ad Hoc Networks*, vol. 3, no. 3, pp. 257–279, 2005.

[20] J. Heidemann, W. Ye, J. Wills, A. Syed, and Y. Li, "Research challenges and applications for underwater sensor networking," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '06)*, vol. 4, pp. 228–235, Las Vegas, Nev, USA, April 2006.

[21] Q. Zhang, C. Zhang, M. Liu, and S. Zhang, "Local node selection for target tracking based on underwater wireless sensor networks," *International Journal of Systems Science*, vol. 46, no. 16, pp. 2918–2927, 2015.

[22] M. Erol, L. F. M. Vieira, and M. Gerla, "AUV-aided localization for underwater sensor networks," in *Proceedings of the 2nd Annual International Conference on Wireless Algorithms, Systems, and Applications (WASA '07)*, pp. 44–54, IEEE, Chicago, Ill, USA, August 2007.

[23] J. E. Faugstadmo, "Underwater wireless sensor networks," in *Proceedings of the 4th International Conference on Sensor Technologies and Applications (SENSORCOMM '10)*, Venice, Italy, July 2010.

[24] S. Zhou and P. Willett, "Submarine location estimation via a network of detection-only sensors," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 3104–3115, 2007.

[25] Q. Liang and X. Cheng, "Underwater acoustic sensor networks: target size detection and performance analysis," *Ad Hoc Networks*, vol. 7, no. 4, pp. 803–808, 2009.

[26] Y. Huang, W. Liang, H.-B. Yu, and Y. Xiao, "Target tracking based on a distributed particle filter in underwater sensor networks," *Wireless Communications and Mobile Computing*, vol. 8, no. 8, pp. 1023–1033, 2008.

[27] X. Wang, M. Xu, H. Wang, Y. Wu, and H. Shi, "Combination of interacting multiple models with the particle filter for three-dimensional target tracking in underwater wireless sensor networks," *Mathematical Problems in Engineering*, vol. 2012, Article ID 829451, 16 pages, 2012.

[28] M. T. Isik and O. B. Akan, "A three dimensional localization algorithm for underwater acoustic sensor networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4457–4463, 2009.

[29] C. Yu, K. Lee, J. Choi, and Y. Seo, "Distributed single target tracking in underwater wireless sensor networks," in *Proceedings of the SICE Annual Conference*, pp. 1351–1356, Tokyo, Japan, August 2008.

[30] P. Djukic, Y. Zhou, and M. Toulgoat, "Localization for electromagnetic radio underwater sensor networks," in *Proceedings of the 5th International Conference on Sensor Technologies and Applications (SENSORCOMM '11)*, pp. 172–177, Nice, France, August 2011.

[31] Y. Zhu, X. Lu, L. Pu et al., "Aqua-Sim: an NS-2 based simulator for underwater sensor networks," in *Proceedings of the ACM International Conference on Underwater Networks and Systems*, pp. 1–2, Kaohsiung, Taiwan, November 2013.

[32] S. M. N. Alam and Z. J. Haas, "Coverage and connectivity in three-dimensional networks," in *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking (MOBICOM '06)*, pp. 346–357, Ithaca, NY, USA, September 2006.

[33] G. Han, C. Zhang, L. Shu, N. Sun, and Q. Li, "A survey on deployment algorithms in underwater acoustic sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2013, Article ID 314049, 11 pages, 2013.

[34] Q. Cao, T. Yan, T. Abdelzaher, and J. Stankovic, "Analysis of target detection performance for wireless sensor networks," in *Proceedings of the International Conference on Distributed Computing in Sensor Networks*, Los Angeles, Calif, USA, June 2005.

[35] A. M. Mahdy and J. M. Groenke, "Target tracking in marine wireless sensor networks," *International Journal on Advances in Networks and Services*, vol. 3, no. 1-2, pp. 103–113, 2010.

[36] C. Gui and P. Mohapatra, "Power conservation and quality of surveillance in target tracking sensor networks," in *Proceedings of the Tenth Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pp. 129–143, Philadelphia, Pa, USA, October 2004.

[37] A. Arora, P. Dutta, S. Bapat et al., "A line in the sand: a wireless sensor network for target detection, classification, and tracking," *Computer Networks*, vol. 46, no. 5, pp. 605–634, 2004.

[38] A. K. H. Tung, H. Lu, J. Han, and L. Feng, "Efficient mining of intertransaction association rules," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 1, pp. 43–56, 2003.

[39] E. Mazor, "Interacting multiple model methods in target tracking: a survey," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 103–123, 1998.

[40] A. I. Al-Shamma'a, A. Shaw, and S. Saman, "Propagation of electromagnetic waves at MHz frequencies through seawater," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 11, pp. 2843–2849, 2004.

[41] I. Wells, A. Davies, X. Che et al., "Node pattern simulation of an undersea sensor network using RF electromagnetic communications," in *Proceedings of the IEEE International Conference on Ultra Modern Telecommunications and Workshops (ICUMT '09)*, pp. 1–4, IEEE, St. Petersburg, Russia, October 2009.

[42] M. R. Frater, M. J. Ryan, and R. M. Dunbar, "Electromagnetic communications within swarms of autonomous underwater vehicles," in *Proceedings of the First ACM International Workshop on Underwater Networks (WUWNet '06)*, pp. 64–70, Los Angeles, Calif, USA, September 2006.

Hindawi
Journal of Sensors
Volume 2017, Article ID 4259652, 15 pages
http://dx.doi.org/10.1155/2017/4259652

Research Article

Semantical Markov Logic Network for Distributed Reasoning in Cyber-Physical Systems

Abdul-Wahid Mohammed,[1,2] Yang Xu,[1] Ming Liu,[1] and Haixiao Hu[1]

[1]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China*
[2]*School of Engineering, University for Development Studies, Tamale, Ghana*

Correspondence should be addressed to Yang Xu; xuyang@uestc.edu.cn

Received 28 May 2016; Accepted 27 November 2016; Published 24 January 2017

Academic Editor: José A. Somolinos

The challenges associated with developing accurate models for cyber-physical systems are attributable to the intrinsic concurrent and heterogeneous computations of these systems. Even though reasoning based on interconnected domain specific ontologies shows promise in enhancing modularity and joint functionality modelling, it has become necessary to build interoperable cyber-physical systems due to the growing pervasiveness of these systems. In this paper, we propose a semantically oriented distributed reasoning architecture for cyber-physical systems. This model accomplishes reasoning through a combination of heterogeneous models of computation. Using the flexibility of semantic agents as a formal representation for heterogeneous computational platforms, we define autonomous and intelligent agent-based reasoning procedure for distributed cyber-physical systems. Sensor networks underpin the semantic capabilities of this architecture, and semantic reasoning based on Markov logic networks is adopted to address uncertainty in modelling. To illustrate feasibility of this approach, we present a Markov logic based semantic event model for cyber-physical systems and discuss a case study of event handling and processing in a smart home.

1. Introduction

Cyber-physical systems (CPSs) [1–3] represent a novel research field with high prospects for providing synergy between the digital and physical worlds. These systems consist of interconnected components, which collaboratively execute tasks in order to bridge the gap between the digital and physical worlds [4]. With the growing complexity of tasks due to the combination of CPSs and Internet of Things (IoT) [5, 6] however, CPSs have become distributed, and it is necessary developing interoperable CPSs capable of enabling timely delivery of services. In this way, CPSs become real-time distributed with multiscale dynamics and networking for efficient, dependable, safe, and secure management of monitoring and control of objects in the physical domain [7].

Following the integration of CPSs and IoT, innovative concepts and approaches, such as service-oriented architecture (SOA) [8, 9], collaborative systems [10], and cloud computing [11], have become apparent in the development of CPSs. As devices in CPSs are required to interoperate at both cyber and physical scales, service-oriented CPSs [12, 13] so far look promising. By coupling the need for CPSs to interact with real world objects in real-time with optimal resource consumption however, using only the service-oriented approach is not enough to realise comprehensive distributed real-time CPSs that take into account the strong interdependencies between the cyber and physical components. Therefore, context-aware agent-based approach is more appealing since diverse attributes can be encapsulated within agents, and distributed pervasive computing with better coordination and interoperability among autonomous and heterogeneous agents is achievable. Essentially, context-awareness is indispensable in CPSs since sensing, resource discovery, adaptation, and augmentation are the key drivers of this novel technology [14, 15]. Additionally, with the changing dynamics of distributed CPSs domains, it is natural that partial observability is inherent in CPSs [16, 17]. Using ontology as underlying semantic technology therefore

requires uncertainty modelling techniques towards good model performance.

This paper proposes a context-based multiagent architecture for distributed reasoning in CPSs. Sensor networks in distributed physical environments provide the semantic capabilities of this model, and semantically annotated low-level contextual information merges with domain knowledge in a reasoning engine. Derived implicit knowledge through semantic reasoning and together with the annotated data enables distributed software agents operating in the cyberspace to provide decision support for actuation information. Each agent is capable of interacting with the physical environment and sharing some principal commonalities with the other agents. As such, the agents are capable of exposing, consuming, and even at times processing collaborative services targeting laid down system goals. To incorporate uncertainty in modelling, semantic reasoning on this model incorporates the inferential power of Markov logic networks (MLN) [18] into event recognition to reduce inferential and computational loads. Finally, we discuss a case study of a smart home as a CPS, and results of our experiments show feasibility of this approach in modelling concurrent events in CPSs.

In summary, we describe in this paper our initial work in CPSs, which overcomes some of the major limitations in this research field. We provide four main contributions. First, we propose a multiagent architecture that can bridge the gap between the operations of the cyber and physical components of CPSs. Second, we describe a procedure that can be used to dynamically compose high-level system goals and underlying criteria from low-level contextual information. Third, we introduce a smart home ontology, which incorporates human actors in the physical space of CPSs as computing entities. Finally, we present a methodology based on MLN for event recognition in CPSs.

The rest of the paper is organised as follows: we present the state of the art and the study background and other preliminary information in Sections 2 and 3, respectively; Section 4 gives an agent model for CPSs, followed by our framework for distributed reasoning in CPSs in Section 5; uncertainty-based event recognition in CPSs using MLN is presented in Section 6; experiments and discussions of results are presented in Section 7; and finally, we conclude and propose future research in Section 8.

2. Related Work

A coherent distributed reasoning architecture is one that achieves an interoperable CPS to cope with the requirements of the physical and cyber components. Because of the lack of sound theoretical foundation for CPSs currently [19], most approaches successfully model either the physical component [20] or the cyber component [21], but not both. Salient studies towards comprehensive models for CPSs that take into account both the cyber and physical components include [8, 22, 23], which use service-oriented computing to achieve interoperable CPSs. Service-oriented approach alone, however, is not suitable for modelling real-time distributed CPSs with multiscale dynamics and networking. In this regard, an agent-based modelling is more appropriate for distributed complex systems such as CPSs [24, 25].

Because of the underlying sensor network in CPSs, semantic agent technologies are closely associated with our approach. As provided in [26], a semantic agent technology has been used to describe a battlefield information system, which uses information fusion processes to dynamically integrate sensor networks towards real-time context-based reasoning. To enable scalable sensor information access, an architecture and programming model for service-oriented sensor information platform has also been proposed [27]. This approach leverages an ontological abstraction of information to optimise use of resources in collecting, storing, and processing data. Timeliness and concurrency in distributed processing environments can also be enhanced using autonomous software agents. As such, use of autonomous semantic agents as a new software paradigm for distributed computing environments has been proposed [28].

Obviously, the complex dynamics of CPSs, coupled with the need to properly represent embedded computing and communication capabilities, motivate the use of semantics and distributed agents towards interoperable CPSs. As can be found in [29], a multiagent model for CPSs in which a distributed semantic agent model augments data acquisition process with ontological intelligence has been proposed. This model, however, provides no procedure for reasoning locally about individual components and globally about system-wide properties. But such semantics, for instance, event recognition, is very critical for distributed real-time CPSs and can essentially specify components of systems in terms of interfaces and observations [7, 30]. Additionally, ontology forming the underlying layer in the semantic agent-based model primarily supports certainty-based reasoning and as such requires techniques to address uncertainty in modelling.

In our approach therefore, we augment the semantic multiagent architecture with a robust reasoning mechanism, which can support both certainty and uncertainty-based reasoning. To promote concurrency and timeliness in the operations of CPSs, MLN is adopted as an uncertainty modelling framework and can compactly represent heterogeneous computations using a common set of rules. Essentially, this achieves a reasoning procedure for CPSs by leveraging advantages of both ontology and probabilistic graphical models to model both complexity and uncertainty, which reduces limitations of both the ontology and probabilistic graphical models.

3. Background and Preliminaries

We provide in this section aspects of the problem of real-time distributed reasoning in CPSs, Markov logic, and smart home as a case study of a CPS.

3.1. Problem Description. The growing pervasiveness of CPSs further keeps the cyber and physical components apart, and separately managing these components would not allow us to realise fully the benefits of these systems. Appropriate

techniques that would allow interoperability between these components are essential so that monitoring and actuation can be invoked remotely. In this regard, standardised interfaces that can achieve interoperable CPSs are desirable and should be guided by the following:

(1) Defining an architectural framework that supports interoperability and distributed reasoning in CPSs.

(2) Applying semantics to explicitly represent contextual information and providing an efficient data storage mechanism.

(3) Providing distributed reasoning procedure that creates more autonomy and intelligence in operations of CPSs.

(4) Incorporating uncertainty into modelling.

Following above challenges and the ability of semantic agents to be discoverable and autonomous, we pursue semantically oriented techniques, in which ontological intelligence is used to address the problem of uncertainty-based real-time distributed computing in CPSs.

3.2. Markov Logic Network. MLN [18] is an interface layer in artificial intelligence, which defines a first-order knowledge base in terms of first-order logic formulae and associated weights. Given a set of constants depicting objects of a domain, MLN defines a ground Markov network, which represents a probability distribution over possible worlds. Each world, basically, represents assignment of truth values to all ground atoms, and this distribution is a log-linear model given by

$$P(x) = \frac{1}{Z} \exp\left\{\sum_i w_i n_i(x)\right\} = \frac{1}{Z}\prod_i \phi_i\left(x_{\{i\}}\right)^{n_i(x)}, \quad (1)$$

where $n_i(x)$ is the amount of true grounding of first-order formula F_i in x, $x_{\{i\}}$ is the ith state of the predicates appearing in each formula, w_i is the weight of F_i, $\phi_i(x_{\{i\}}) = e^{w_i}$ is the potential function of each clique in the ground Markov network, and Z is a normalizing constant called the *partition function*. Each weight indicates the strength of a constraint that a formula represents and is directly proportional to the difference in the log probability between a world that satisfies the formula and one that does not.

Due to the varying number of constants that can represent the same knowledge base either in part or full, MLN allows the same formulae to be applicable under all circumstances and can be viewed as a template for constructing Markov networks. In this way, different sets of constants can produce different ground Markov networks using a common underlying MLN. This, ideally, is suitable for domains, such as CPSs, where the task of reasoning requires combining separate reasoning chunks, which need to be processed independently.

3.3. Case Study. As intelligence in the home gets more sophisticated, intelligent interconnection of distributed consumer hardware such as consoles, smart home servers, and smart phones running diverse functionalities like assistive health care and home automation constitutes a CPS. To demonstrate the heterogeneity, concurrency, and sensitivity to timing of CPSs, a case study of temperature event recognition is considered. This presents a scenario of using an ontology-based model to achieve an interoperable CPS that leverages a common event recognition model across different layers. Specifically, using a single computation of a car system, events pertaining to temperature conditions of the car user's home and that of the car engine can be achieved. Typically, this computational platform must support concurrent processing of events since temperature events for the home and the car can concur.

The fact that CPSs need to be sensitive is a challenging task in respect of false positives that can arise if uncertainty is not well managed. Apart from noisy sensor information and incomplete domain knowledge being primary sources of uncertainty, environmental factors are also potential sources of uncertainty that cannot be ignored in CPSs. For instance, a temperature event that considers optimal resource consumption in a smart home may trigger opening of windows on a cool sunny day for comfort of the home. This strategy, though, all things being equal, sounds ideal, but environmental factors such as air population and external noise can present a trade-off between comfortability and minimising resource consumption in the home. As such, the effectiveness of CPSs will be much appreciated if uncertainty, which is unavoidable in nature, is well managed in these systems.

4. Agent Network for CPSs

Building on the foundation of mobile agent network [31], we define a multiagent system residing in the cyberspace of CPSs. This model considers all aspects of agents' communication and operation including issues relating to performance of multiagent systems and CPSs.

4.1. Cyber Agent Model. A cyber agent model is defined by a triple CAM $= \langle A, D, N\rangle$, where A represents a community of agents depicting distributed computational environments in a CPS, D denotes specific domains of agents' services, and N defines networks of agents operating in the cyberspace. Essentially, agents in this definition can be stationary and mobile so as to suit the changing dynamics of CPSs domains. In this way, an agent A_i in a multiagent system, which is defined by $A = (A_1, \ldots, A_n)$, represents a specific computational platform and can perform tasks allowed by its domain. This specifies somewhat autonomy in the operations of these agents, and tasks can be encapsulated as agents' capability from the viewpoint of functionality and performance. As such, interactions between agents provide the needed communication and cooperation to bridge the gap between the operations of the cyber and physical components of CPSs.

Contextual reasoning paradigm [32] in which each agent depends on a domain specific ontology and can link with other agents through semantic mapping makes this design distributed. Since each computational platform provides a functionality for service execution, combining these multiple

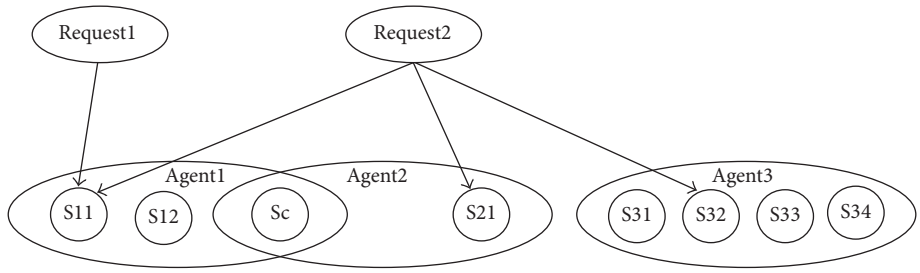


Figure 1: Service assignment graph for agents' domains.

ontologies through semantic mapping allows us to define joint functionalities that can be used in complex task operations. For a nonempty set I of indices used to identify agents associated with domain specific ontologies $\{D_i\}_{i\in I}$, we define the joint functionality of our multiagent system as a set of cross-layer services $JS = \{S_1, \ldots, S_n\}$. Thus, a set of services indexed by each domain is defined as $\{\mathcal{S}_i\}_{i\in I}$. Intuitively, $\mathcal{S}_i$ represents a service formalisation of the ith ontology.

We must recognise that each service S_i can be provided by one or more agents. To avoid conflicts in accessing services, an agent definition that explicitly specifies computational platforms and services provided is critical. In this way, we make services distinct by encapsulating each agent definition as a property of a computational platform in a CPS using the triple $agent_i = \langle name_i, deployment_i, service_i \rangle$. As we can see, this definition of an agent, apart from a property describing a given computation, also provides information about agents' services and domains for those services. A service domain is specified by a deployment property, which can be a physical address of a distributed environment in CPSs. For instance, given a set of agents' services domains $D = \{D_1, \ldots, D_n\}$, the set of services provided by agent $agent_i$ can be described by $S_i = \{S_{i1}, \ldots, S_{in}\}$, and each service within this domain can be invoked using the pair $address_i = \langle S_i, agent_i \rangle$. This means, within the cyberspace, the interactions between these agents define an undirected graph $N = (D, E)$, where E denotes an edge between any two domains of agents with overlapping functionalities.

4.2. Agent Cooperation in CPSs. For the multiagent system to ensure high fidelity between the physical and cyber components, each agent domain must trigger requests, which represent implicit knowledge inferred from the domain's low-level contextual information. Typically, these requests may involve complex tasks that can exceed capacity of a single operational domain and may require cross-domain services aggregation. In this case, agents operating in the cyberspace, through their interactions, can communicate and cooperatively assume specific roles to execute tasks.

The idea that agents' domains are distributed and can form a graph with overlapping functionality presents a complex network in which fast information sharing becomes necessary for large agent teams [33]. Relying on information importance as a determinant for service assignment based on performance requirements of agents' domains is ideal for fast information sharing and parallel execution of services. In this regard, service requests to agents can be either domain specific or across multiple domains. As shown in Figure 1, a special case in the assignment of services to agents is, for example, *Request1*, when a request is domain specific. In this case, all services can be deemed mutually exclusive, and communication and coordination among agents become less important. However, when services overlap or a given request requires a combination of cross-domain services, we face a problem of a multiagent autonomy denoted by *Request2*. But interestingly, instead of employing a planning agent for the assignment of services in this case, agents can combine their inherent intelligence with domain knowledge to negotiate for services in our design.

5. Distributed Reasoning in CPSs

Distributed computing systems in CPSs can employ pervasive computing techniques to provide autonomous, interoperable, computational elements that can be described, discovered, and orchestrated within and across different layers [15]. In this regard, ontology-oriented modelling of contexts offers a lot of advantages [34]. We present next a semantic agent-based architecture for CPSs. Additionally, semantics for formulation of high-level complex tasks with underlying criteria from low-level contextual information is discussed.

5.1. Semantic Multiagent Architecture for CPSs. This is a broker-centric multiagent architecture, which can support cross-layer service collaboration in CPSs. As shown in Figure 2, this architecture captures into perspective the modelling concerns of CPSs raised in [35]. Key components of this architecture are *data management module*, *context ontology module*, *semantic reasoning engine*, and a *confederation of semantic agents*. Detailed descriptions of these components are provided in the following subsections.

(1) Data Management Module. The key functionalities of this module are collection and transmission of data to storage areas. Raw context data are acquired from distributed sensor networks in the physical domain, and the heterogeneity of these data requires *semantic markups* that applications can easily understand. Through semantic annotation, the context data are transformed into semantic markups that can link to external definitions through unique *URIs* of ontology instances. For example, the semantic annotation of a temperature sensor data can be as in Listing 1.

```
<Device rdf: ID="&obs; TempSensor2">
<owl: sameAs rdf: resource ="&smh; TempSensorR2"/>
<smh: observedPhenomenom rdf: resource ="&smh; Temperature"/>
<smh: qtyValue rdf: datatype ="&xsd; float">24.5</smh: qtyValue>
<smh: qtyUnit rdf: resource ="smh; Celsius"/>
<smh: timeStamp>2015–07–28 16:52:30</timeStamp>
</Device>
```

Listing 1

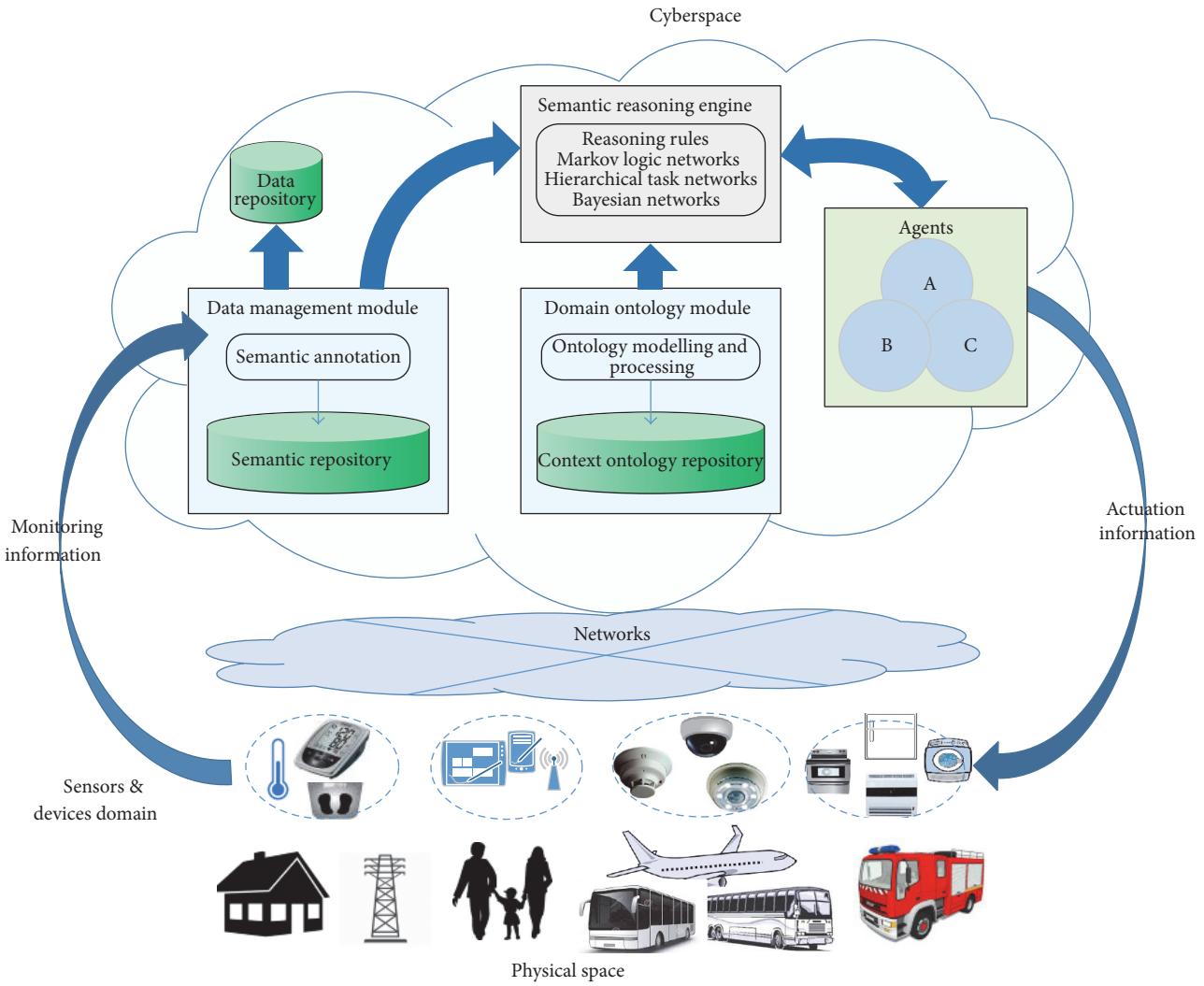


Figure 2: Multiagent context architecture for CPSs.

As can be seen clearly, each annotation contains a unique URI, such as *smh*, *TempSensor*2, of an ontological instance, description of the observed phenomenon, measured value of the phenomenon, unit of measurement used, and the timestamp for the observation. Apart from the timestamp and measurement value, which are XML Schema datatypes, the other attributes are resources which exist in an external definition.

Data storage in this architecture is achieved at two levels using different databases that must interoperate with each other. In the first instance, the raw sensor data is stored in such a way that it can be efficiently maintained and exported by disparate sources. Secondly, after the raw sensor data is semantically annotated, a storage mechanism that supports semantics is required for storing this annotated data. Thus, in our architecture, the annotated data are stored in a repository

as ontology instances and associated properties that machines can easily interpret. This repository is updated whenever a new context event occurs and can be augmented with Linked Data techniques to support semantic data integration at the instance level.

In Linked Data research [36, 37], dereferencing the URIs of resources through HTTP protocol can be exploited to incrementally obtain the description of resources. To further support integration of semantic data distributively, OWL axiom *owl:sameAs* has been used successfully at the instance level in Linked Data research. With such success in a distributed setting, this same technique can be adopted into CPSs for semantic integration. Thus, semantic markups in this paper are linked to external definitions using the owl:sameAs axiom. As can be seen in the example above, the use of this axiom shows that the two URIs refer to the same instance, thereby providing a mapping between the semantic repository and the context ontology repository.

(2) Context Ontology Module. Ontology modelling and processing occur in this module. In CPSs, computing entities and services of distributed intelligent environments can be grouped together, forming service-oriented ecosystems. Contexts in these domains can share some concepts in common, even though their detailed properties can differ significantly. Instead of completely modelling all contexts across different domains, the objective here is to model contexts using a *base ontology* and a *domain specific ontology*. Entities of the base ontology are extensible basic common concepts across different environments. The domain specific ontology, however, represents only those concepts that uniquely exist in each domain.

Specifically, in a smart home domain, the most fundamental concepts we have identified as extensible nodes of the base ontology are *user*, *deployment*, *service,* and *computing entity*. When these entities are linked together, a skeleton of contextual entity is formed, which allows context-based data acquisition. Figure 3 shows the context ontology model we propose for a smart home domain in this paper. The base ontology, which is extended by both a smart home and a smart institution, illustrates the advantage of knowledge reuse using ontology. In both cases, base entities such as Room and AdhocService are extended to meet specifications of the application domain. For instance, whilst we can specify bedroom and living room in a home, an institution can have rooms such as lecture room and conference room.

It is important to note that human as a computing entity in this model is a novel contribution that is ideal for design of CPSs. This essentially extends the service-oriented paradigm to incorporate human services in CPSs towards transmuting system components and behavioural practices [38]. Specially with our design, the role of humans as both actors and sources of contextual information in the physical domain can be explicitly represented and allows social awareness to be incorporated into CPSs. In this view, CPSs are well positioned to provide emotional intelligence [39] that will respond appropriately to people and situations.

```
<ReasonedData rdf: ID="&obs; TempSensor2">
<smh: domainURI rdf: resource ="&smh; room104"/>
<smh: criticalLevel rdf: resource ="smh; High"/>
<smh: alert>Fire in building</smh: qtyValue>
</ReasonedData>
```

Listing 2

(3) Semantic Reasoning Engine. This is the central component of our design in which high-level implicit knowledge can be inferred from sensed contextual information. Semantically annotated data and the context ontology are aggregated into a coherent model that semantic agents and physical objects can share. Both certainty-based and uncertainty models can be supported by this design. But the focus of this research is uncertain decision support in CPSs. Specifically, uncertainty-based reasoning about resources and events and dynamic formation of collaborative cross-layer services given high-level system goals with underlying criteria are the focus of this design.

It is worth noting that putting the reasoned data to use by the semantic agents requires a data structure that can easily integrate with the underlying semantic repository of this representation. Specifically, the semantics of the reasoned data, when used to feed these agents, must specify among other needs, the referenced domain of the inferred knowledge. The argument here is that since CPSs domains are highly distributed in nature, agents can cooperate effectively across different domains to achieve better computational intelligence if we semantically specify domains of inferred knowledge in the reasoned data. In line with this paper's objective, such a data structure can allow easy mashup of resources to solve complex problems. Thus, Listing 2 is an example of a semantic markup of a reasoned data.

As we can see, this example basically demonstrates that aside the domain of interest referenced using *domain-URI*, other elements fitting a given scenario are allowed. Among these additions that is also unavoidable is the high-level knowledge obtained through semantic reasoning. This knowledge in this example is specified using the element *alert* and mostly what users get as prompts. Obviously, such a data structure combined with the domain knowledge can allow the semantic agents gain enhanced computing capabilities.

(4) Semantic Agents. The semantic agents are distributed algorithms executing on multiple distributed computing entities in the physical space. To provide decision support for actuation information, these agents merge semantically annotated data with the reasoned data. By describing these agents as a community, they are ostensibly the control point of this architecture and can advertise their services in the reasoner through interactions and semantic reasoning. Thus, each agent's behaviour is well suited for its environment, and such behaviours are well suited for resource discovery through semantic reasoning.

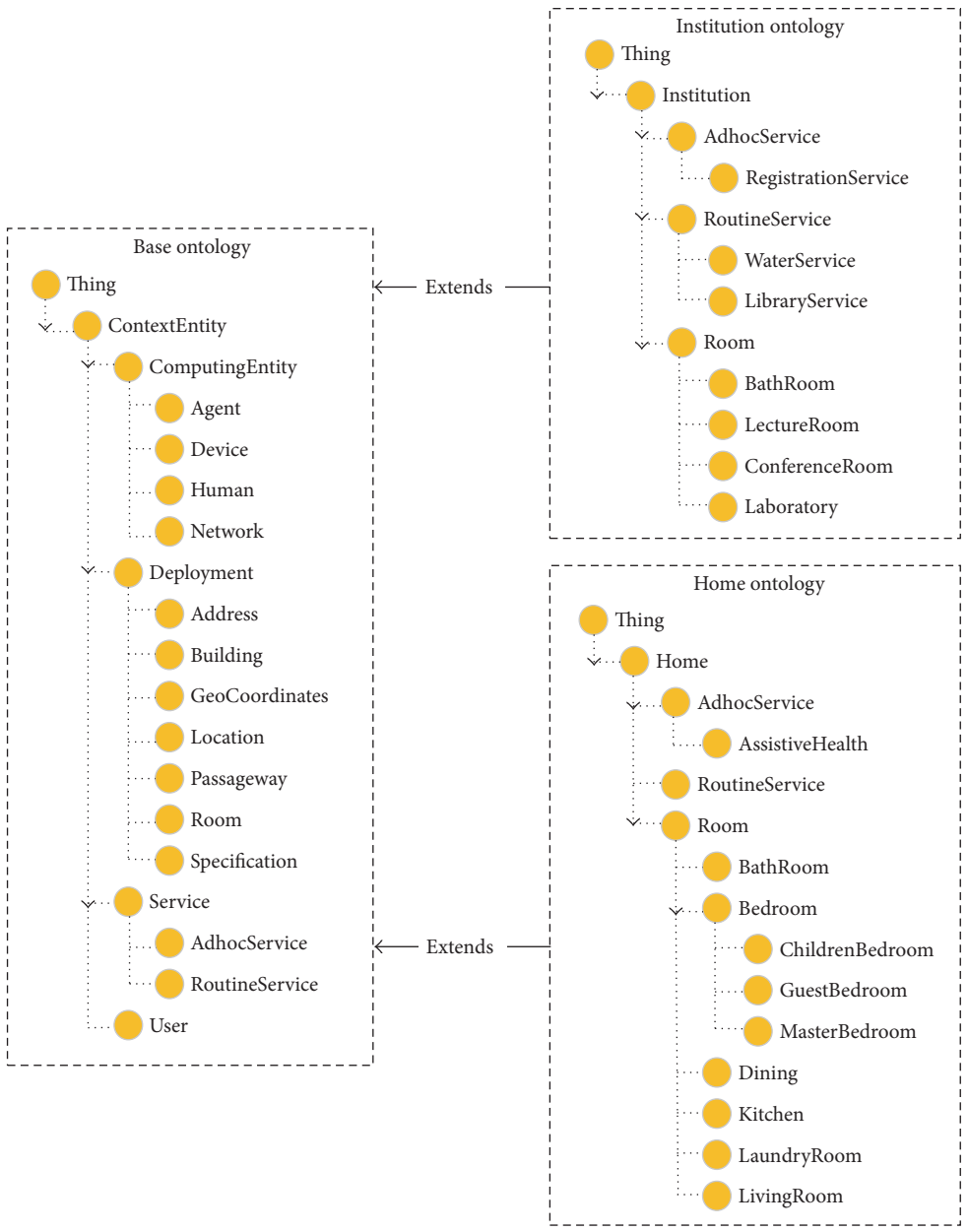


Figure 3: Smart home context model for CPSs.

5.2. Dynamic Composition of High-Level Complex Tasks. Facilitation of dynamic formation of collaborative services towards execution of complex tasks requires elicitation of high-level complex services from low-level information. This process can guarantee better quality CPSs and overcomes common engineering design flaws to provide right actuation information for the needs of physical objects. For instance, through low-level information, we can compose a task, such as *put off the fire*, as an event for handling fire outbreak in a CPS environment. However, this particular task, unlike some tasks, requires complex functionality and needs to be decomposed into primitive level tasks, which can then be serviced by specific resources. This is a challenging process and therefore requires a dedicated framework on how to figure out complex functionality from low-level contexts.

As shown in Figure 4, this approach is motivated by the established actuation relationship between the cyberspace and distributed CPS environments. For clarity of representation, the physical environment is categorised into *usage* and *context* environments. The usage environment describes processes performed by semantic agents, and how systems can achieve tasks in the environment. Because each agent's behaviour best suits its environment, the usage environment ostensibly contains specific objectives, which describe the activities of agents towards useful output. The context environment, as specified by the context acquisition process, is the

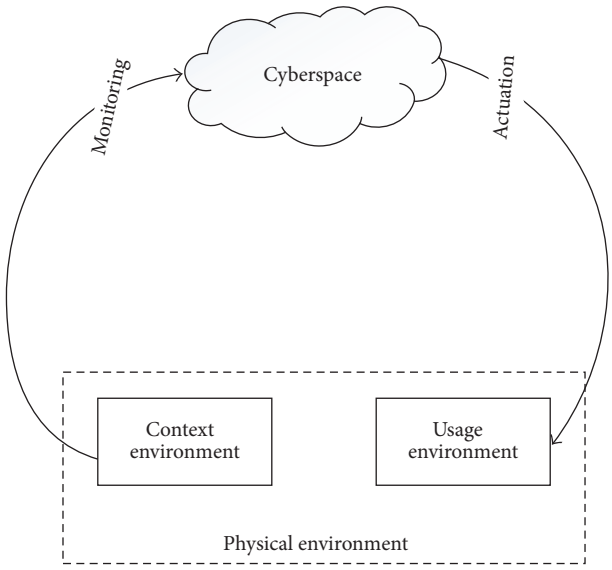


Figure 4: The relationships between a distributed CPS environment and the cyberspace.

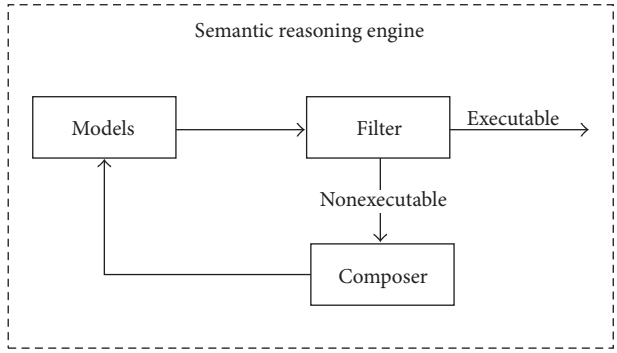


Figure 5: Logic flow of the semantic reasoner of Figure 2.

Input: in
Output: out
(1) **if** *executable* **then**
(2) *out* ← *statement*
(3) **else**
(4) *in* ← *PlanningProblem*(*statement*)
(5) goto step 1
(6) **end if**
(7) **return** out

Algorithm 1: Algorithm for composition and processing of composite context.

source of domain knowledge, which underpins the semantic capabilities of this approach.

In the distributed setting of CPSs, it is required that the cyberspace provides specifications that can address requirements of the physical environment. We can see from Figure 2 that these requirements in the cyberspace are models and processes that control objects in the physical environment and vice versa. This brings to the fore issues about *domain-driven* and *user-driven* requirements. Since the domain-driven requirements, which fundamentally hold the underlying semantics of this approach, have been discussed in the previous section, our focus now is the user-driven requirements. These requirements are enshrined in the actuation information and form the underlying idea towards the elicitation of high-level complex services from the domain-driven requirements. Therefore, understanding the relationship that the actuation information establishes between the cyberspace and the physical environment is essential towards mapping contextual information to high-level complex services.

High-level composite context can be derived from low-level contextual information through the semantic reasoner of Figure 2. As shown in Figure 5, the logic flow of the reasoning engine of this architecture consists of three main functional blocks: *models*; *filter*; and *composer*. Reasoned data from reasoning models are passed through a filtering process in which statements are categorised based on a predefined set of rules. All statements through this *filter* are either categorised as *executable* or *nonexecutable* but not both. An executable statement represents a single service phenomenon, such as *high temperature*, which can directly be executed either remotely or centrally by reducing the home's temperature through an air-conditioner. But when a statement requires aggregation of services in order to achieve its objective, it is filtered to be nonexecutable. An example of a nonexecutable statement is when sensors detect smoke and high temperature in a building, and the reasoned information is *fire outbreak in building*. Obviously, an appropriate action in this case is to put off the fire, which requires different services wrapped as capabilities of different physical objects in this approach. In this regard, this reasoned information requires further processing in the form of composition of appropriate services and constraints so that tasks can be appropriately scheduled among computing entities.

Whilst the executable statements directly feed the semantic agents, the nonexecutable statements are transformed into high-level composite contexts for further processing through the composer. Following the objective of this paper, these statements are composed into high-level tasks with underlying criteria. Specifically, this stage generates an HTN planning problem, which is passed as an input to the reasoner for automatic composition of collaborative services based on Algorithm 1. As we can see, the filtering process creates a cycle of execution of information whenever a nonexecutable statement is encountered. From lines 4 and 5, a nonexecutable statement results in a planning problem consisting of tasks and constraints. Generation of the planning problem is an instance of Algorithm 2, which takes a nonexecutable statement as an input. With this transformation, collaborative services can then be formed using a specified model and can directly feed the semantic agents. To support the needs of time-criticality in CPSs, this design promotes efficient use of computational resources by ensuring that all nonexecutable statements experience one-off processing. In essence, the

Input: in
Output: tasks, preconditions
(1) generate output
(2) **return** out

Algorithm 2: Planning problem.

set of rules provided in the model block ensures that all composite tasks are reduced to primitive tasks, which can be directly executed.

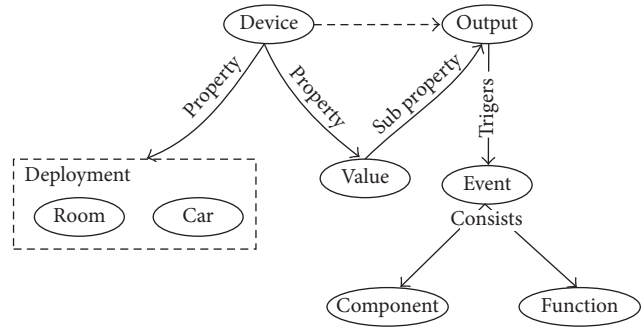


Figure 6: Context-based event model in CPSs.

6. Event Recognition Using Markov Logic

Context-based events are central in initiating activities in CPSs and naturally specify real-time demand responsiveness of systems' components in terms of interfaces and observations [39]. From the viewpoint of our multiagent architecture, events can stimulate services of one or more agents in the network, and it is therefore important to detect events of predefined operations that are desirable to both systems and users of CPSs.

An event ontology is required to augment the proposed context ontology in the previous section towards event-based reasoning. However, ontology in its classical form currently cannot represent and reason under uncertainty. In view of good modelling practice towards best performance of CPSs, we adopt MLN based event recognition to address uncertainty whilst keeping the structure of the underlying ontology intact. This exploits the view of MLN as a template for Markov networks so that only a part of OWL rules applicable to events are considered in the model construction. One obvious advantage is in the compact representation of model complexity, which can guarantee incorporation of rich domain knowledge for high sensitivity and good concurrent processing of events.

MLN allows existing knowledge bases in first-order logic to incorporate uncertainty in knowledge representation by adding weights to logic formulae. Since first-order logic underlines the fundamental theory of ontology, OWL rules for knowledge discovery can therefore be transformed into MLN weighted formulae towards event recognition.

6.1. Rules for Event Recognition. OWL rules form the underlying logical framework of our event recognition process, and the semantics we provide holistically capture the domain's interest phenomena in the rules. As shown in Figure 6, an event is described by a tuple $\langle component, function\rangle$, which denotes *event components* and *event semantic functions* [40]. We use event components to compose heterogeneous sensor data to form contextual information that match events' requirements. Thus, $component_i = \{o_{i1}, \ldots, o_{in}\}$, where o_{ij} represents observations driving the occurrence of an event. To be able to propagate logical constraints of events in rules, we use semantics of event functions, $\langle Stop, Change, Comparison\rangle$, to distinguish between categories of events. With this specification, Stop is a predicate we use to express action of an event that changes spontaneous states of natural phenomena. Thus, the predicate, $Stop(x, y)$, defined by

$$Active(y) \wedge (Action(x, y) \vee Affect(x, y)) \Longrightarrow \neg Active(y) \quad (2)$$

indicates that the state y changes when acted upon by an action of event x. In some cases of CPSs, the event may have effects on the state, as $Affect(x, y)$ indicates. In a fire scenario, for example, this predicate ensures the right event invocation that will put off the fire. We must note that the same predicate cannot be applied to routine events such as putting off the air-conditioner. Unlike in the fire case whereby the new state of affairs after putting off the fire may be perpetual, devices are only eligible for temporal state changes. As such, we use the unary predicate $Change(x)$ as the event function for achieving temporal state changes for events in CPSs. $Change(x)$ is defined as

$$\exists y \quad (x = y \Longrightarrow x \neq y), \quad (3)$$

where state x changes to new state y using three subfunctions. Specifically, a state change can be rise in degree of something using the predicate $Increase(x)$; reduction in degree of something using predicate $Decrease(x)$; and toggling between on and off modes of devices using the predicate $Switch(x)$. All these subfunctions operate by conditioning the current state against the new state. For instance, to increase the temperature of a room using an air-conditioner, the semantics of the predicate $Increase(x)$ is defined as

$$\exists x \quad Status(x) \wedge \exists (x, y)\, Value(y) \Longrightarrow \exists (> x, y)\, Value(y) \quad (4)$$

to indicate a change in status value of the air-conditioner when y is greater than x. Obviously, this predicate, like the other two, requires semantics that compares the current and new states in the change process.

The Comparison predicate is used to define conditions that describe state changes. We consider $LessThan(x, y)$, $LessThanEqual(x, y)$, $GreaterThan(x, y)$, $GreaterThanEqual(x, y)$, and $Equal(x, y)$ as the predicates for conditions for state changes. For example, the semantics for $LessThan(x, y)$ is defined as

$$\exists c, \quad Cond(c) \wedge Value(x) = c \wedge Value(y) < c \quad (5)$$

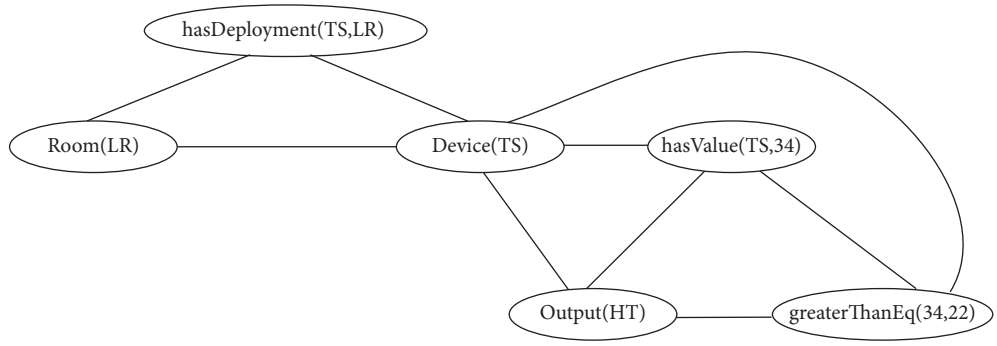


Figure 7: A section of ground Markov network constructed from the MLN for event recognition.

to express the condition that the value of x is c, and the value of x is less than the value of y. Therefore, event recognition based on component structures and semantic functions provide logical operations in formulae that form good basis for Markov logic based event recognition.

6.2. Translation of Rules into MLN. The first step towards conversion of OWL rules into MLN requires transformation of OWL rules into first-order logic formulae. As provided in [41], OWL classes and properties, respectively, represent *unary* and *binary* predicates in first-order logic and can be combined using logical connectives to form atomic formulae. For example, the first-order logic translation of a class Room is $Room(x)$, where x denotes instances of the given class. For a property hasDeployment, the equivalent first-order logic formula is

$$hasDeployment\,(x, y) \Longrightarrow Device\,(x) \wedge Room\,(y)\,, \quad \forall x, y, \tag{6}$$

where x and y, respectively, denote the domain and range classes of this property. Axiomatization of classes and property restrictions can also be translated into first-order logic. For instance, rdfs:subClassOf axiom can be translated into first-order logic as

$$Room\,(x) \Longrightarrow Bedroom\,(x) \quad \forall x \tag{7}$$

to indicate that bedRoom is a subclass of Room. On this basis, we can use logical connectives to compose the first-order logic formula of the concept class Device and its property as

$$Device\,(x) \wedge hasDeployment\,(x, y) \Longrightarrow Room\,(y) \quad \forall x, y. \tag{8}$$

Interestingly, the Alchemy [41] tool for MLN provides built-in functions that simplify the translation of logical conditions into MLN. For instance, the predicate $LessThan(x, y)$ in MLN is simply represented using the internal predicate of Alchemy $lessThan(int, int)$. Basically, this predicate tests if the first argument is less than the second argument.

Once we obtain the first-order logic translation of rules, the MLN is achieved by adding weights to each formula. The MLN together with a set of constants define a ground Markov network on which probabilistic reasoning can be performed. Figure 7 shows a section of the ground Markov network based on the MLN of our case study for event recognition. As we can see in this figure, links exist between any two ground terms appearing together in the same formula in the MLN. Hence, given a MLN and a set of constants, arbitrary queries such as the conditional probability that a formula holds given another formula in the MLN can be addressed.

6.3. Fuzzy Markov Logic Network. We recognise that the axiomatic notion of probability as presented in the last subsection is incapable of dealing with vague information in knowledge. This becomes apparent in MLN when multivalued clauses are encountered, and this presents a challenge beyond the classical notion of MLN. In this view, we provide a fuzzy notion of Markov logic called *fuzzy MLN* in which inference to queries requires the inference machinery of fuzzy logic.

The basic idea serving as a point of departure in fuzzy MLN lies in the fact that a formula F in first-order logic can be viewed as a collection of elastic constraints, which restrict the weights W associated with each grounding of its terms. To achieve this, we define a fuzzy membership function in terms of weights and ground terms of MLN clauses and obtain an extension of MLN to fuzzy MLN as

$$F_{\text{MLN}} = \left(\mu : F \longrightarrow W\right). \tag{9}$$

This represents a mapping of a set F of grounded first-order logic formula into a set of MLN weights W. The idea that different constants refer to different objects in MLN and a formula can contain more than one ground clause allows for separate assignments of weights to each ground clause in MLN. Essentially, this achieves fuzzy membership functions mapping ground MLN clauses into an ordered set of fuzzy pairs from which MLN inferences can be performed. Clearly, this fuzzy set is completely determined by the set of tuples

$$F_c = \left\{\left(c, \mu\,(c)\right) \mid c \in C\right\} \tag{10}$$

denoting the assignment of weights to each grounding of F for a set of constants C.

As shown in Figure 8, the membership function in fuzzy MLN, assuming without loss of generality that all weights are

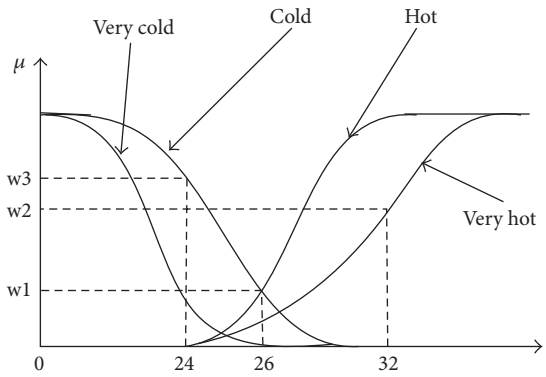


Figure 8: Membership function of fuzzy MLN.

positive, is a representation of the magnitude of participation of the weight of each ground term as an input. This associates different weights with the same formula for different ground terms and defines functional overlaps between these ground terms, which determines outcomes of rules. As we can see in this figure, a typical case of an overlap is the temperature value 26°C, which presents a case of a multivalued clause for different grounding of the same formula. This value exists in the interval of the *minimum criterion* of the two fuzzy sets defined by *cold* and *hot*

$$\mu_{cold \cap hot} = \min\left(\mu_{cold}, \mu_{hot}\right), \tag{11}$$

where μ_{cold} and μ_{hot}, respectively, define the membership functions of the states cold and hot. Intuitively, this can be described as to what degree a cold temperature 26°C is hot. Obviously, knowledge about true state of this temperature value smacks of vagueness and can be efficiently interpreted as a fuzzy constraint on a collection of ground terms.

In fuzzy MLN, unlike in classical fuzzy logic, the situation described above is easily handled by specifying ground clauses involving both cases in a training set. As Figure 9 depicts, this defines a ground Markov network in which the two outcomes are conditionally independent given the input. Each dotted circle in this figure indicates a weighted ground clause of the same formula, and either clause is a complement of the other, which in the absence of one clause defines the classical notion of MLN. This means learning the weight $w1$ as shown in Figure 8 in this particular case produces two values that define the fuzzy set for reasoning. Thus, fuzzy MLN leveraging these weights and together with a set of constants defines a ground Markov network, which can be reasoned upon using the inference machinery of fuzzy logic without employing any formal fuzzy logic semantics.

7. Results and Discussion

In this section, we present and discuss results of event recognition under uncertainty in a smart home as a CPS. The conditions for all experiments were designed to test performance of this approach using key intrinsic requirements of CPSs such as sensitivity to timing and concurrency. In this regard, thrust of our analyses bothers mainly about precision as a measure of sensitivity to timing of occurrence of single and concurrent events in CPSs.

To incorporate uncertainty into modelling for our experiments, we considered an MLN based event model rooted in the OWL ontology of this paper's case study. This ontology captures into perspective semantics of our event model, and the key properties include

(i) *hasDeployment*, which relates the concept *Device* to concepts *Room* and *Engine,*

(ii) *hasValue*, which relates the concept *Device* to a data value,

(iii) *hasOutput*, which relates a data value of a device to its semantic interpretation using the concept *Output,*

(iv) *hasEvent*, which relates the concept *Output* to the concept *event*.

Essentially, *Room* and *Engine*, which are subclasses of *Location,* define heterogeneous computational platforms for devices using the deployment property. Consequently, events associated with different platforms denote effects of interpretations of devices' values on those platforms and are stored in the ontology as type *event*.

For compact modelling towards expedited processing of events, OWL rules provided partial specification of domain concepts relevant for the construction of the MLN event model. With the ability of OWL to support heterogeneous processes, we used the same set of rules to represent different computational platforms in the MLN. Specifically, computations in relation to a home's indoor comfort index and operational safety condition of a car's engine were considered to be two likely synchronous events that a single computation can represent. The ostensible need for a single computation for distributed environments in CPSs can be seen in a case of driving towards home whereby the computation of the car's console monitors both the home condition and the engine temperature of the car. In this way, a distributed sensor network of the home and the vehicle engine provides contextual information for the computational intelligence. With the deployment property of devices in the underlying ontology, contextual information can be accurately filtered and applied according to domain specifications.

As shown in Table 1, five events were defined to represent the two distributed environments considered in our experiments. Essentially, the heterogeneity in these events is enshrined in the different conditions pertaining to temperature measurements in these environments. For instance, whilst we can specify a normal room temperature to be in the range 21°C–27°C, a normal operating temperature range of an engine is 180°C–205°C, which is high temperature in a case of the home and way beyond limits of human survival temperature. Clearly, these two cases represent vagueness in knowledge as both denote a normal temperature, and it is therefore important disambiguating between heterogeneous sensed information in modelling. In this regard, our event model can be described as a composite model designed to precede any recognition process with precise knowledge discovery. Aside event recognition, this same model can be used to perform semantic reasoning towards mashup of

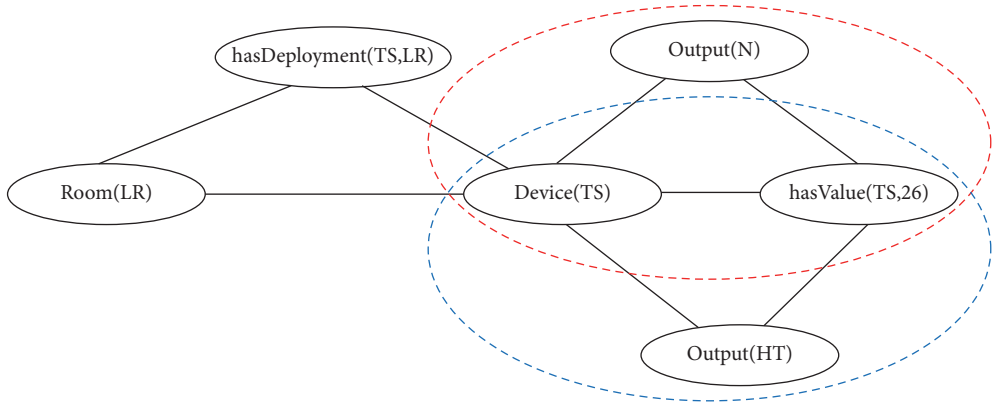


Figure 9: An example of ground Markov network of a fuzzy MLN.

Table 1: Event specification for home and engine domains.

Domain	Value range	Inferred knowledge	Event
Home	Less than 21°C	Low temperature	Increase temperature
	21°C–27°C	Normal temperature	No event
	Greater than 27°C	High temperature	Reduce temperature
Engine	180°C–205°C	Normal temperature	Normal operation
	Greater than 205°C	Overheat	Switch off engine

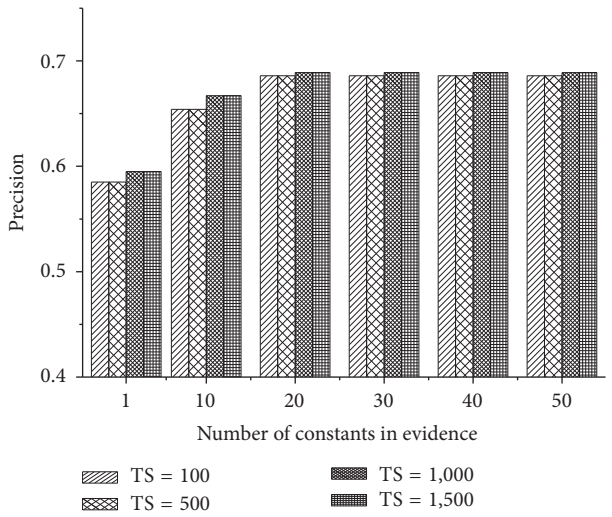


Figure 10: Precision of a single event recognition with increasing evidence.

resources. For instance, temperature sensors deployed at any of the two distributed environments in our case study can be inferred with this model.

We evaluated the performance of this model by considering both single event and multiple event recognition tasks. In either case, we varied the training set of the MLN from 100 constants to 1,500 constants. As we can see in Figure 10, the precision of a single event recognition improves as more ground terms are introduced into the training set. Looking down the column from left to right, we will notice that the effect of the number of constants in the training set stabilises at some point. We found this development interesting in our preliminary analysis because if we consider the columns representing the number of constants 100 and 500 or columns representing the number of constants 1,000 and 1,500 for the event of a single constant as evidence, one may be tempted to conclude that any two training sets differing by 500 constants give approximately the same results. This notion, however, is different when we look at the columns representing the number of constants 500 and 1,000. Consequently, we tried varying the number of constants in the evidence set as well to better understand this trend.

By increasing the constants of the evidence set to 50, we observe that the precision of the recognition increases from below 60% to about 70%. Whilst much improvement is achieved between 1 and 10 constants in evidence, the performance stabilises from 20 constants. This gives the intuition that a caveat can be established for the density of sensory information at a given location since more evidence after some point can be inadmissible in the recognition process. From this observation, we hypothesize that a coherent representation of events combining modalities can allow CPSs to efficiently monitor and control environments with multiple sensors.

Concurrent event recognition was also investigated using multiple events representing the two distributed domains under consideration. Similarly, we measured the precision for recognition of multiple events using training sets containing constants 100, 500, 1,000, and 1,500. As shown in Figure 11, the precision for recognition of multiple events follows the trend of the single event recognition. This observation is attributable to the fact that even though different constants

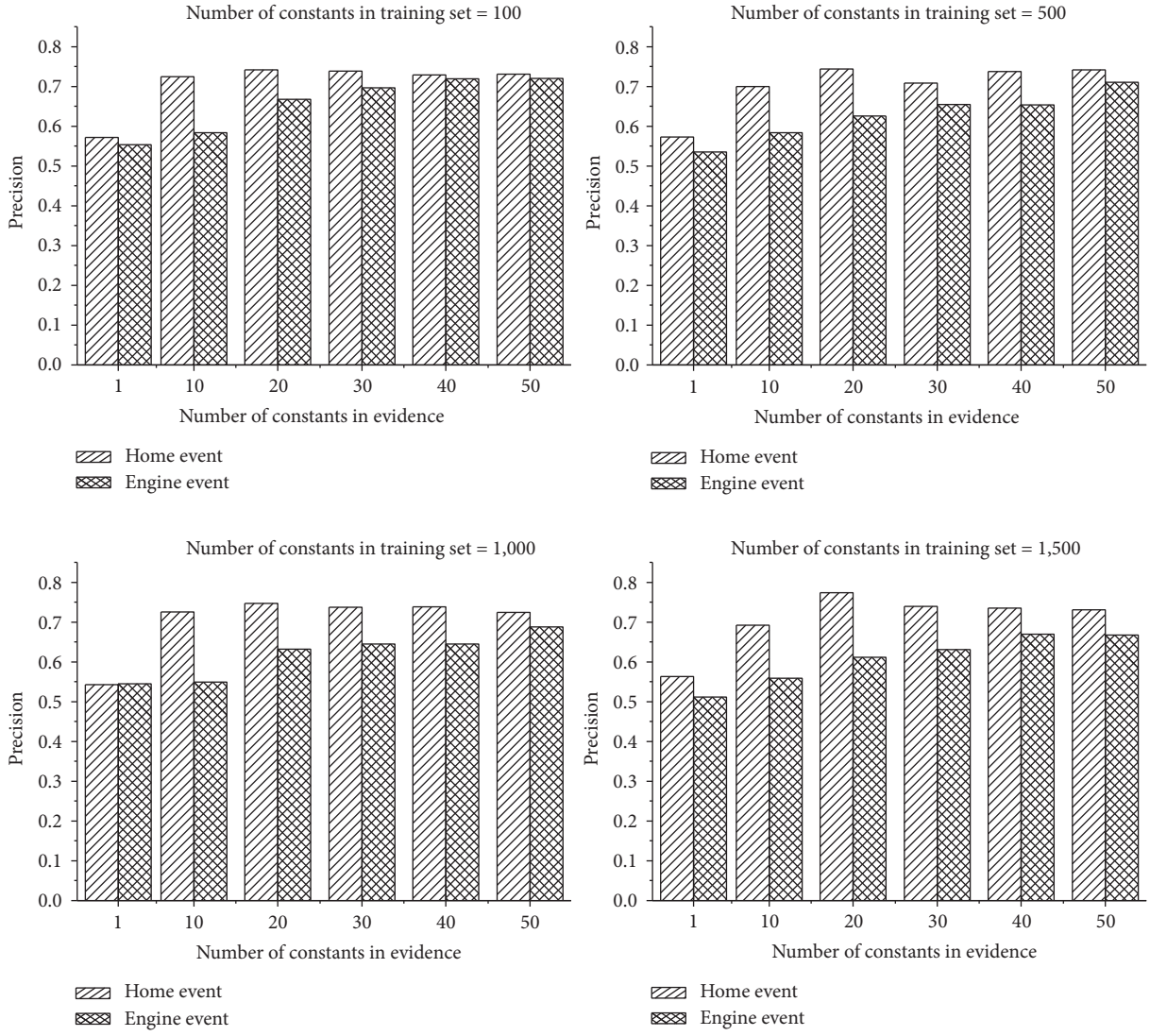


Figure 11: Precision of a multiple event recognition with increasing evidence.

applicable to the single event and multiple event cases generate different ground Markov networks with varying structures, the model performance is still guided by the common underlying MLN. However, we recognise that the monotonicity in the precision with increasing constants in evidence is not upheld fully in the multiple event recognition. Whereas the precision increases monotonically in the engine event, the home event does not follow this trend completely. But this new phenomenon is yet another indication that determining a caveat on the size of contextual data towards optimal precision in event recognition is paramount. Obviously, all training data sets hold that the precision of the home event is modal at 20 constants in evidence, which is consistent with the precision of single event recognition. Because the multiple event recognition process contains more combined constants than the single event recognition, the multiple event process performs better. In essence, this approach can support concurrency in operations of CPSs for collaborative processes.

Finally, we investigated a multivalued logic in which MLN clauses are somewhat fuzzified to represent partial knowledge. In mimicking rectangular fuzzy functions, we used the built-in Alchemy predicate greaterThanEq(x, y) to define a single MLN clause that represents both normal and high temperature event inputs. Weights of MLN clauses characterise subregions of high and normal temperature measurements. Overlapping regions, as enshrined in fuzzy logic [42], can be treated as a straightforward constants declaration in the training data set. For instance, temperature value 26°C was considered ambiguous in our experiments. So in the training data set, this temperature value was declared a few times to indicate that this same value could be described as normal and slightly warm. As we can see from Figure 12, we obtained impressive results with even least training data set. Using a training data set of 100 constants, the precision for a single fuzzy event recognition also improves with increasing number of constants in the evidence set. Overall, using MLN, event recognition in CPSs can be modelled to handle

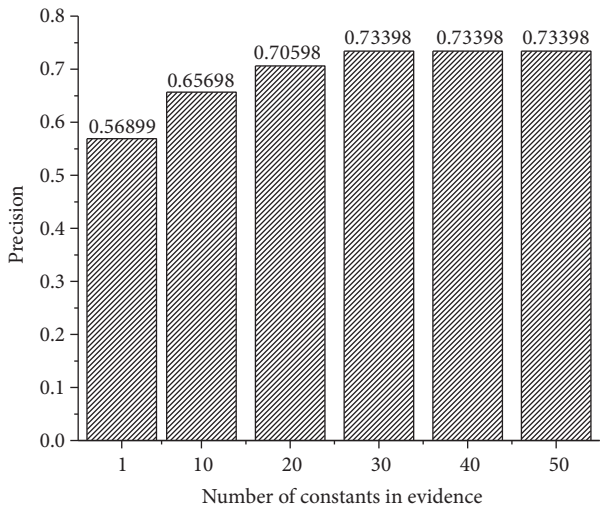


Figure 12: Precision of a fuzzy event recognition with increasing evidence.

both uncertainty and vagueness whilst modelling domain uncertainty.

8. Conclusion

In this paper, we proposed a context-aware multiagent architecture for distributed reasoning in cyber-physical systems. This architecture is rooted in service-oriented computing, and with the incorporation of services of semantic agents, the seamless integration of the cyber and physical components can be achieved. Ontological intelligence provides the underlying semantics of this approach, and together with Markov logic networks, we defined an uncertainty-based reasoning procedure for event recognition in cyber-physical systems. With the results of our experiments, it is convincing that this framework can be relied upon for concurrent processing of events in cyber-physical systems. Because these semantic agents are thought to be autonomous and intelligent in their operations, future work of this research shall consider agent communication techniques that can ensure good level of cooperation among these agents.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research was sponsored by NSFC 61370151 and 61202211, National Science and Technology Major Project of China 2015ZX03003012, Central University Basic Research Funds Foundation of China ZYGX2014J055, and Huawei Technology Foundation YB2013120141.

References

[1] R. Baheti and H. Gill, "Cyber-physical systems," *The Impact of Control Technology*, vol. 12, pp. 161–166, 2011.

[2] E. A. Lee, "Cyber physical systems: design challenges," in *Proceedings of the 11th IEEE International Symposium on Object Oriented Real-Time Distributed Computing (ISORC '08)*, pp. 363–369, Orlando, Fla, USA, May 2008.

[3] L. Sha, S. Gopalakrishnan, X. Liu, and Q. Wang, "Cyber-physical systems: a new frontier," in *Machine Learning in Cyber Trust*, pp. 3–13, Springer, Berlin, Germany, 2009.

[4] Y. Tan, M. C. Vuran, and S. Goddard, "Spatio-temporal event model for cyber-physical systems," in *Proceedings of the 29th IEEE International Conference on Distributed Computing Systems Workshops (ICDCS '09)*, pp. 44–50, Montreal, Canada, June 2009.

[5] L. Atzori, A. Iera, and G. Morabito, "The internet of things: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[6] G. Kortuem, F. Kawsar, V. Sundramoorthy, and D. Fitton, "Smart objects as building blocks for the internet of things," *IEEE Internet Computing*, vol. 14, no. 1, pp. 44–51, 2010.

[7] J. Singh, O. Hussain, E. Chang, and T. Dillon, "Event handling for distributed real-time cyber-physical systems," in *Proceedings of the 15th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC '12)*, pp. 23–30, Shenzhen, China, April 2012.

[8] S. Karnouskos, A. W. Colombo, T. Bangemann et al. et al., "The imcaesop architecture for cloud-based industrial cyber-physical systems," in *Industrial Cloud-Based Cyber-Physical Systems*, pp. 49–88, Springer, Berlin, Germany, 2014.

[9] A. Boyd, D. Noller, P. Peters et al., *Soa in Manufacturing Guide Book*, MESA International, IBM Corporation and Capgemini Co-Branded White Paper, 2008.

[10] R. Harrison and A. W. Colombo, "Collaborative automation from rigid coupling towards dynamic reconfigurable production systems," *IFAC Proceedings Volumes*, vol. 38, no. 1, pp. 184–192, 2005.

[11] M. Armbrust, A. Fox, R. Griffith et al., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[12] K.-J. Lin and M. Panahi, "A real-time service-oriented framework to support sustainable cyber-physical systems," in *Proceedings of the 8th IEEE International Conference on Industrial Informatics (INDIN '10)*, pp. 15–21, IEEE, Osaka, Japan, July 2010.

[13] D. D. Hoang, H.-Y. Paik, and C.-K. Kim, "Service-oriented middleware architectures for cyber-physical systems," *International Journal of Computer Science and Network Security*, vol. 12, no. 1, pp. 79–87, 2012.

[14] J. Pascoe, "Adding generic contextual capabilities to wearable computers," in *Proceedings of the Digest of Papers. Second International Symposium on Wearable Computers*, pp. 92–99, Pittsburgh, PA, USA, Oct. 1998.

[15] S. Noor and H. N. Minhas, "Context-aware perception for cyberphysical systems," in *Computational Intelligence for Decision Support in Cyber-Physical Systems*, pp. 149–167, Springer, 2014.

[16] Z. Song, Y. Chen, C. R. Sastry, and N. C. Tas, *Optimal Observation for Cyber-Physical Systems*, Springer, London, UK, 2009.

[17] L.-A. Tang, X. Yu, S. Kim et al., "Trustworthiness analysis of sensor data in cyber-physical systems," *Journal of Computer and System Sciences*, vol. 79, no. 3, pp. 383–401, 2013.

[18] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[19] J. Wan, H. Yan, H. Suo, and F. Li, "Advances in cyber-physical systems research," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 11, pp. 1891–1908, 2011.

[20] G. Quan, "An integrated simulation environment for cyber-physical system co-simulation," in *Proceedings of the National Workshop on High-Confidence Automotive Cyber-Physical Systems*, Troy, Mich, USA, April 2008.

[21] T. W. Hnat, T. I. Sookoor, P. Hooimeijer, W. Weimer, and K. Whitehouse, "MacroLab: a vector-based macroprogramming framework for cyber-physical systems," in *Proceedings of the 6th ACM Conference on Embedded Networked Sensor Systems (SenSys '08)*, pp. 225–238, ACM, Raleigh, NC, USA, November 2008.

[22] H. J. La and S. D. Kim, "A service-based approach to designing cyber physical systems," in *Proceedings of the 9th IEEE/ACIS International Conference on Computer and Information Science (ICIS '10)*, pp. 895–900, Potsdam, Germany, August 2010.

[23] C.-F. Lai, Y.-W. Ma, S.-Y. Chang, H.-C. Chao, and Y.-M. Huang, "OSGi-based services architecture for cyber-physical home control systems," *Computer Communications*, vol. 34, no. 2, pp. 184–191, 2011.

[24] P. Pederson, D. Dudenhoeffer, S. Hartley, and M. Permann, *Critical Infrastructure Interdependency Modeling: A Survey of Us and International Research*, Idaho National Laboratory, 2006.

[25] P. P. Datta, M. Christopher, and P. Allen, "Agent-based modelling of complex production/distribution systems to improve resilience," *International Journal of Logistics Research and Applications*, vol. 10, no. 3, pp. 187–203, 2007.

[26] G. Jiang, W. W. Chung, and G. Cybenko, "Semantic agent technologies for tactical sensor networks," in *Proceedings of the in SPIE Conference on AeroSense*, pp. 311–320, International Society for Optics and Photonics, 2003.

[27] J. Liu and F. Zhao, "Towards semantic services for sensor-rich information systems," in *Proceedings of the 2nd International Conference on Broadband Networks (BROADNETS '05)*, pp. 44–51, Boston, Mass, USA, October 2005.

[28] A. Elci and B. Rahnama, "Considerations on a new software architecture for distributed environments using autonomous semantic agents," in *Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC '05)*, vol. 2, pp. 133–138, IEEE, July 2005.

[29] J. Lin, S. Sedigh, and A. Miller, "Modeling cyber-physical systems with semantic agents," in *Proceedings of the 34th Annual IEEE International Computer Software and Applications Conference Workshops (COMPSACW '10)*, pp. 13–18, Seoul, Korea (South), July 2010.

[30] C. Talcott, "Cyber-physical systems and events," in *Software-Intensive Systems and New Computing Paradigms*, pp. 101–115, Springer, Berlin, Germany, 2008.

[31] I. Lovrek, "Context awareness in mobile software agent network," *Tehničke Znanosti*, vol. 513, no. 15, pp. 7–28, 2012.

[32] L. Serafini and A. Tamilin, "Drago: distributed reasoning architecture for the semantic web," in *The Semantic Web: Research and Applications*, pp. 361–376, Springer, 2005.

[33] Y. Zhang, Y. Xu, H. Hu, and X. Liu, "Semantical information graph model toward fast information valuation in large team work," in *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI '14)*, August 2014.

[34] X. H. Wang, D. Q. Zhang, T. Gu, and H. K. Pung, "Ontology based context modeling and reasoning using OWL," in *Proceedings of the 2nd IEEE Annual Conference on Pervasive Computing and Communications (PerCom '04)*, pp. 18–22, Orlando, Fla, USA, March 2004.

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

[36] L. Ding, J. Shinavier, T. Finin, and D. L. McGuinness, *Owl: Sameas and Linked Data: An Empirical Study*, 2010.

[37] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson, "When owl: sameas isn't the same: an analysis of identity in linked data," in *The Semantic Web—ISWC 2010*, vol. 6496 of *Lecture Notes in Computer Science*, pp. 305–320, Springer, Berlin, Germany, 2010.

[38] N. Bieberstein, S. Bose, L. Walker, and A. Lynch, "Impact of service-oriented architecture on enterprise systems, organizational structures, and individuals," *IBM Systems Journal*, vol. 44, no. 4, pp. 691–708, 2005.

[39] N. S. Schutte, J. M. Malouff, L. E. Hall et al., "Development and validation of a measure of emotional intelligence," *Personality and Individual Differences*, vol. 25, no. 2, pp. 167–177, 1998.

[40] K. Kaneiwa, M. Iwazume, and K. Fukuda, "An upper ontology for event classifications and relations," in *AI 2007: Advances in Artificial Intelligence: 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, December 2–6, 2007. Proceedings*, Lecture Notes in Computer Science, pp. 394–403, Springer, Berlin, Germany, 2007.

[41] S. K. P. S. M. Richardson, P. Domingos, and M. S. H. Poon, *The Alchemy System for Statistical Relational AI: User Manual*, 2007.

[42] L. A. Zadeh, G. J. Klir, and B. Yuan, *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers*, vol. 6, World Scientific, 1996.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 3953750, 18 pages
http://dx.doi.org/10.1155/2016/3953750

Research Article

Fiber-Reinforced Polymer-Packaged Optical Fiber Bragg Grating Strain Sensors for Infrastructures under Harsh Environment

Zhi Zhou, Zhenzhen Wang, and Lian Shao

Institute of Smart Structures, Dalian University of Technology, Ganjingzi, Liaoning, Dalian 116024, China

Correspondence should be addressed to Zhi Zhou; zhouzhi@dlut.edu.cn

Received 8 June 2016; Accepted 3 August 2016

Academic Editor: Rafael Morales

Optical fiber Bragg grating (FBG) has been recognized as an outstanding high-performance local monitoring sensor and is largely applied in structural health monitoring (SHM). This paper proposes a series of fiber-reinforced polymer- (FRP-) packaged optical fiber Bragg grating strain sensors to completely meet the requirements of rough civil engineering infrastructures, and their sensing performance under normal environment and harsh environment is experimentally investigated. It is experimentally and theoretically proved that FRP-packaged FBG strain sensors maintain excellent sensing performance as the bare FBG sensor under a harsh environment, and their durability is significantly enhanced due to the FRP materials. These FRP-packaged FBG strain sensors are successfully applied in the SHM system of Aizhai Bridge.

1. Introduction

Infrastructures, such as long-span bridges, high-rise buildings, large dams, nuclear power stations, and offshore platforms, will inevitably suffer damage accumulation and resistance degradation subjected to coupling actions of environmental corrosion, material aging, long-term loading, fatigue, and natural disaster hazards, even collapse, during their long time of service [1]. Therefore, in order to assure structural safety, integrity, suitability, and durability, a lot of infrastructures in service are in great need for intelligent health monitoring systems to evaluate their safety and rehabilitate and further control their damage. Due to the frequent disastrous lessons, more and more infrastructures have been equipped with long-term health monitoring systems during construction [2]. As one of the most important inventions in the measurement field in the late 20th century, optical fiber Bragg grating (FBG) has been greatly recognized and largely applied in long-term structural health monitoring (SHM) due to the fact that optical FBG shows distinguishing advantages: electromagnetic resistance, small size, resistance to corrosion, and so forth [3–11]. Fiber Bragg grating sensors further provide an absolute measurement that exhibits minimal drift with time, which performs measurement at a discrete position in the fiber, and several sensors can be multiplexed for a complex network connected to a single interrogator along a single fiber. As the main ingredient of the bare optical fiber is SiO_2 and the outer diameter is only 125 μm, the shear capacity of the optical fiber is so poor. Due to its fragility, it is rather difficult to be applied directly in the rough civil engineering infrastructures and harsh environments without packaging. Therefore, it is an important issue to develop packaging techniques for bare FBG strain sensors, which can be well protected inside the matrix component and less likely to be damaged by external infringement. The key problem of this development focuses on the selection of packaging materials for sensors form of different layout process and performance requirements, in order to ensure the that packaged FBG sensor possesses excellent durability, linearity, repeatability, and measurement range for long-term monitoring of civil engineering. There are three solutions for realizing the combination of FBG sensors and packaging materials. Firstly, metallic materials can be chosen as the packaging materials to combine with FBG sensors by an adhesive interlayer. Because of the plastic properties under large strain conditions and corrosion of metallic materials, as well as the creep and aging characteristics of the adhesive interlayer, sensors developed by this encapsulation technique are deficient in durability, linearity, and repeatability, in addition to the small measurement scale (less than 2000 $\mu\varepsilon$).

Secondly, the FBG sensors are firstly clamped with the ended mechanical anchorage and then packaged by the additional protection. It is inevitable that the creep of OFBG sensors would occur under sustained loads due to mechanical clamping; therefore, the measurement scale is just as the ultimate strain of the FBG sensor. In addition to this, the durability is restricted by the material of the clamping devices. Thirdly, advanced composites can be introduced as packaging materials, such as fiber-reinforced polymer (FRP), including carbon fiber-reinforced polymer (CFRP), aramid fiber-reinforced polymer (AFRP), basalt fiber-reinforced polymer (BFRP), and glass fiber-reinforced polymer (GFRP). FRP composites are originated from the design of large, high-performance structures in the aerospace industry. A fundamental issue in the implementation of an OFBG-based SHM system in composite structures aerospace area is the embedment of sensors during manufacturing. Tsutsui et al. applied small-diameter optical fiber sensors to stiffened composite panels for the detection of impact damage [12]. Ryu et al. have used multiplexed and multichanneled built-in FBG sensors to monitor the buckling behavior of a composite wing box [13]. Takeda et al. used FBG sensors to monitor damage due to compressive load in CFRP stiffened panels [14]. Tserpes et al. developed an integrated methodology for monitoring strain and damage in CFRP fuselage panels, and embedment of fiber sensors in the panel during manufacturing was done so as to minimize risk of fiber breaking during manufacturing and impact testing and to effectively capture strains being representative of the damage developed in the panel [15].

In this paper, the bonding mechanism of the combination of FRP and FBG sensors is explored firstly. And then an experiment program is conducted for investigating the durability properties of BFRP bars and GFRP bars, which are used as typical encapsulation materials. Thirdly, series of FRP-packaged FBG strain sensors are developed for infrastructures under harsh environment, and their sensing performance under normal environment and harsh environment is experimentally investigated. Finally, the practical application of these FRP-packaged FBG strain sensors in Aizhai Bridge SHM system is briefly introduced.

2. Bonding Mechanism between FRP and FBG Sensors

As we know, the sensing properties of FRP-packaged optical fiber strain sensors are determined by the FBG and packaging materials. Due to the small proportion of FBG, the basic sensing performance properties, such as linearity, repeatability, and measurement scale, and the main durability index are directly influenced by the basic mechanical and chemical properties of packaging materials. Fiber-reinforced polymer composites provide a reliable means for the development of high-performance packaged optical fiber strain sensors, owing to their linear-elastic material constitutive properties, namely, the notion that the elastic modulus remains constant until failure (as shown in Figure 1), excellent fatigue performance, and durability. Particularly for that, the whole-process pseudoelasticity of FRP composites ensures the perfect linearity and repeatability of the FRP-packaged optical

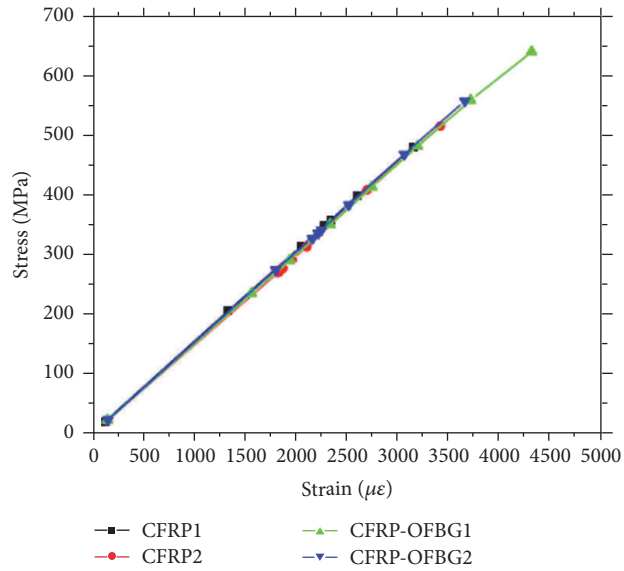
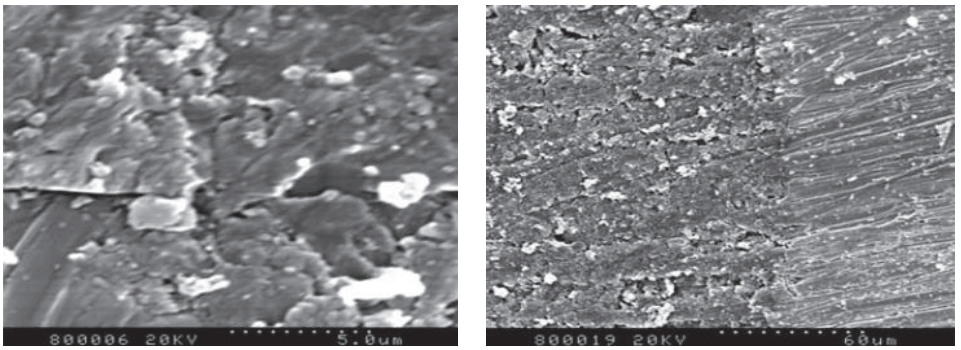


Figure 1: The constitutive curve of CFRP bar.

fiber strain sensors in the whole measurement scale. The fundamental basis of taking FRP composites as packaging materials is further explained as follows.

2.1. The Compatibility between FBG and FRP Composites. From the point of view of material component, fiber-reinforced polymer composites commonly consist of glass fibers (or carbon fibers, aramid fibers, basalt fiber, and hybrid fiber), resins, additives, and so forth. Silica, as the basic composition of the optical fiber, is also the main material composition of glass fibers. Thus, the glass fibers are infiltrated in the resin and cure for molding easily, which is the same as optical fibers. It is shown from the SEM photo (Figure 2) of bare FBG and FRP that the bare fiber FRP combined well with FRP composites and worked together in the manner of full interaction.

2.2. Effect of FRP on the Sensing Characteristics of FBG Sensors. The main indexes of sensing elements are highlighted in linearity, measurement range, repeatability, and so forth, and the macroscopic constitutive property of FRP materials is just linear elastic, which ensures that the excellent linearity of the FBG sensor would be maintained after encapsulating. In addition to that, the FBG is precompressed and induced by the shrinkage of the resin in the curing process of thermosetting FRP composites, due to their perfect bonding with FRP composites, and the precompressed section will be the extended portion for the tensile strain measurement range, compared with the bare FBG sensor under the initially unstressed state. According to the experiment results shown in Figure 3, the range of GFRP-packaged FBG sensor can reach up to 7000~8000 $\mu\varepsilon$. By contrast, the range of bare FBG sensors is only 3000~4000 $\mu\varepsilon$. Thus, it can be seen that FRP composites increase the measurement scale significantly without changing the sensing performance of the FBG sensor (Figure 5).



(a) The interface of bare FBG and CFRP (b) The interface of bare FBG and CFRP

Figure 2: SEM photo of bare FBG and FRP.

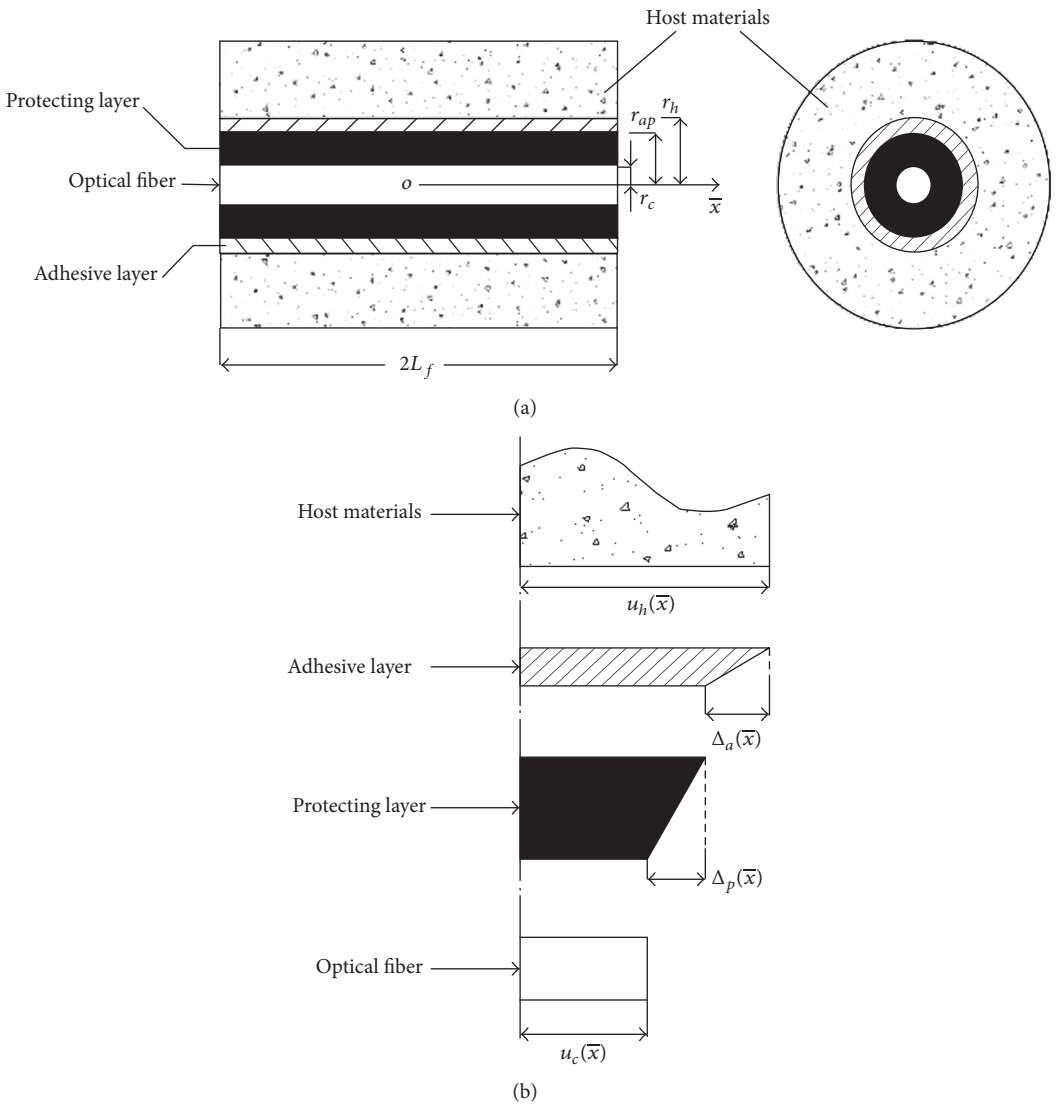


Figure 3: Cylindrical model of optical fiber strain sensing and relationship of deformation.

2.3. Strain Transfer of the FRP-Packaged Optical Fiber Strain Sensors. Embedded FRP-packaged FBG sensors are usually constituted by FRP encapsulation layer and sensing fiber. Deformation of host materials induced by external action passes through FRP encapsulation layer at first and then arrives at the FBG sensor, with a part of strain consumed by the FRP encapsulation layer in the strain transferring process, which causes the strain transfer error between the target strain of structures $\varepsilon_h(0)$ and measurement of FBG sensors $\varepsilon_c(\overline{x})$. In order to estimate the error between the measurement of FBG sensors and the strain of structures, as well as correcting the strain transfer error and improving the testing accuracy of sensors, a cylinder within the effective working length of optical fibers consisting of optical fiber, protecting layer, adhesive layer, and host materials was chosen as the mechanical analysis model to investigate the strain transfer mechanism in optical fiber sensing. Basic assumptions were introduced as follows: (1) optical fiber, protecting layer, adhesive layer, and host materials were considered to be linear elastic and isotropic; (2) all of the adhesive interfaces were continuous and satisfied the deformation compatibility condition; (3) temperature effects were ignored; (4) optical fiber is not sensitive to lateral stress, and thus the lateral stress and shear stress were ignored. According to the hypothesis of displacement continuity, the displacement at arbitrary point in the cylindrical model can be expressed as in the following formula:

$$u(\overline{x}) = \begin{cases} u_c(\overline{x}) & 0 \le r \le r_c \\ u_p(r, \overline{x}) & r_c < r \le r_{ap} \\ u_a(r, \overline{x}) & r_{ap} < r \le r_h \\ u_h(\overline{x}) & r > r_h. \end{cases} \tag{1}$$

In formula (1), $u_c(\overline{x})$, $u_p(r, \overline{x})$, $u_a(r, \overline{x})$, and $u_h(\overline{x})$ are represented as the displacement of the optical fiber, protecting layer, adhesive layer, and host materials, respectively. The displacement compatibility equations at the adhesive interface are shown in formula (2). Due to the presence of the protective layer and adhesive layer, there is relative displacement between the optical fiber and host materials, which is induced by the shear deformation of the protective layer and the adhesive layer. The quantity relationships between the relative displacement for the optical fiber and host materials and shear deformation in the protective layer and the adhesive layer are given in formulas (3)~(5). The relative displacement at each interface is shown in Figure 3(b). Hence,

$$\begin{aligned} u_c(\overline{x}) &= u_p(r_c, \overline{x}) \\ u_p(r_c, \overline{x}) &= u_a(r_c, \overline{x}) \\ u_a(r_c, \overline{x}) &= u_h(\overline{x}) \end{aligned} \tag{2}$$

$$u_h(\overline{x}) - u_c(\overline{x}) = \Delta_a(\overline{x}) + \Delta_p(\overline{x}) \tag{3}$$

$$u_a(r_h, \overline{x}) - u_a(r_{ap}, \overline{x}) = \Delta_a(\overline{x}) \tag{4}$$

$$u_p(r_{ap}, \overline{x}) - u_p(r_c, \overline{x}) = \Delta_p(\overline{x}). \tag{5}$$

When $\overline{x} = 0$,

$$\varepsilon_c(r, 0) = \varepsilon_a(r, 0) = \varepsilon_{ap}(r, 0) = \varepsilon_h(r, 0). \tag{6}$$

The axial force equilibria for the optical fiber infinitesimal, protecting layer infinitesimal, and adhesive layer infinitesimal are shown in Figures 4(a)–4(c), and the axial force equilibrium equations are

$$\sum F_{\overline{x}} = 0 \tag{7}$$

$$\frac{d\sigma_c(\overline{x})}{d\overline{x}} = -\frac{2\tau_{pc}(r_c, \overline{x})}{r_c} \tag{8}$$

$$\frac{d\sigma_p(\overline{x})}{d\overline{x}} = \frac{2\left[\tau_{pc}(r_c, \overline{x}) r_c - \tau_{ap}(r_{ap}, \overline{x}) r_{ap}\right]}{r_{ap}^2 - r_c^2} \tag{9}$$

$$\frac{d\sigma_a(\overline{x})}{d\overline{x}} = \frac{2\left[\tau_{ap}(r_{ap}, \overline{x}) r_c - \tau_h(r_h, \overline{x}) r_h\right]}{r_h^2 - r_{ap}^2}. \tag{10}$$

Deformation compatibility equations in the protective layer and adhesive layer can be approximately expressed as [16]

$$\begin{aligned} \tau(r, \overline{x}) &= \frac{r_{ap}}{r}\tau_{ap}(r_{ap}, \overline{x}) \quad r_c \le r \le r_{ap} \\ \tau(r, \overline{x}) &= \frac{r_h}{r}\tau_h(r_h, \overline{x}) \quad r_{ap} \le r \le r_h. \end{aligned} \tag{11}$$

When $r = r_c$ and $r = r_{ap}$,

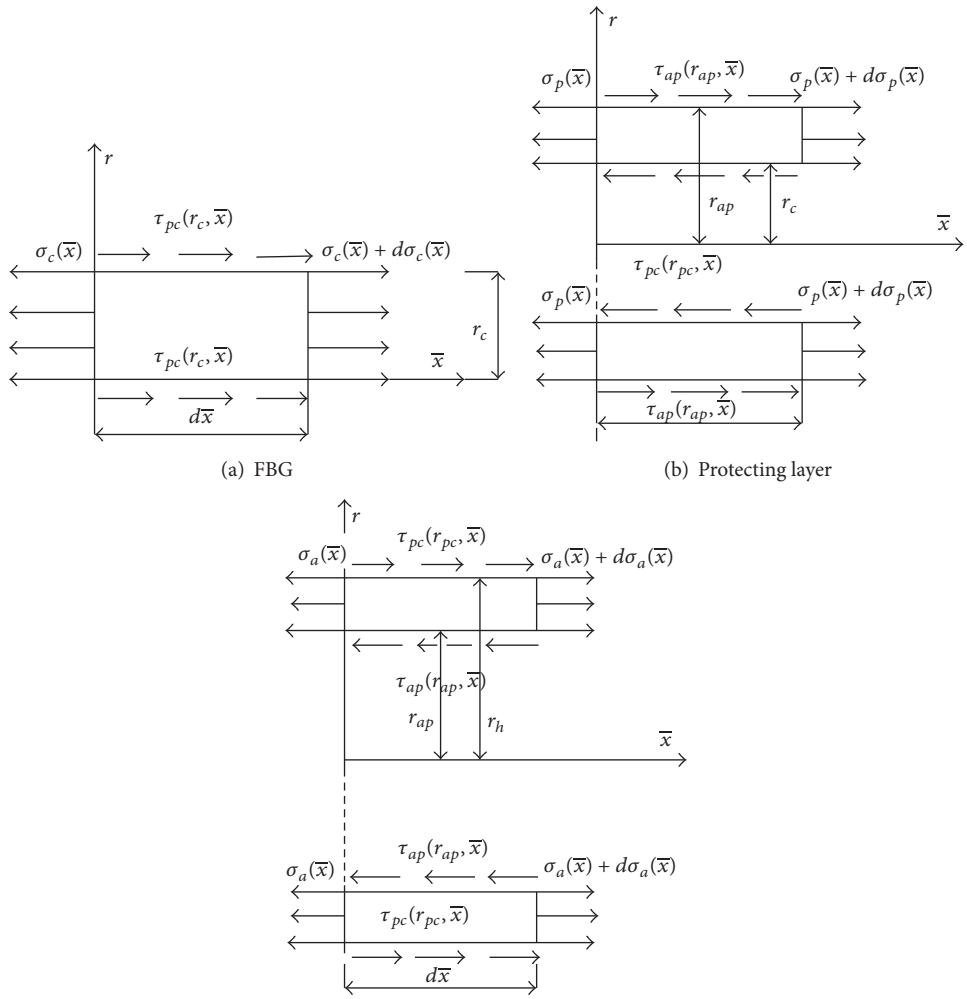
$$\begin{aligned} \frac{d\sigma_p(\overline{x})}{d\overline{x}} &= 0, \\ \frac{d\sigma_a(\overline{x})}{d\overline{x}} &= 0. \end{aligned} \tag{12}$$

Therefore,

$$\begin{aligned} \sigma_p(\overline{x}) &= \text{cons tan } t, \\ \sigma_a(\overline{x}) &= \text{cons tan } t. \end{aligned} \tag{13}$$

Physical equations and geometric equations for optical fiber, protecting layer, and adhesive layer are as follows:

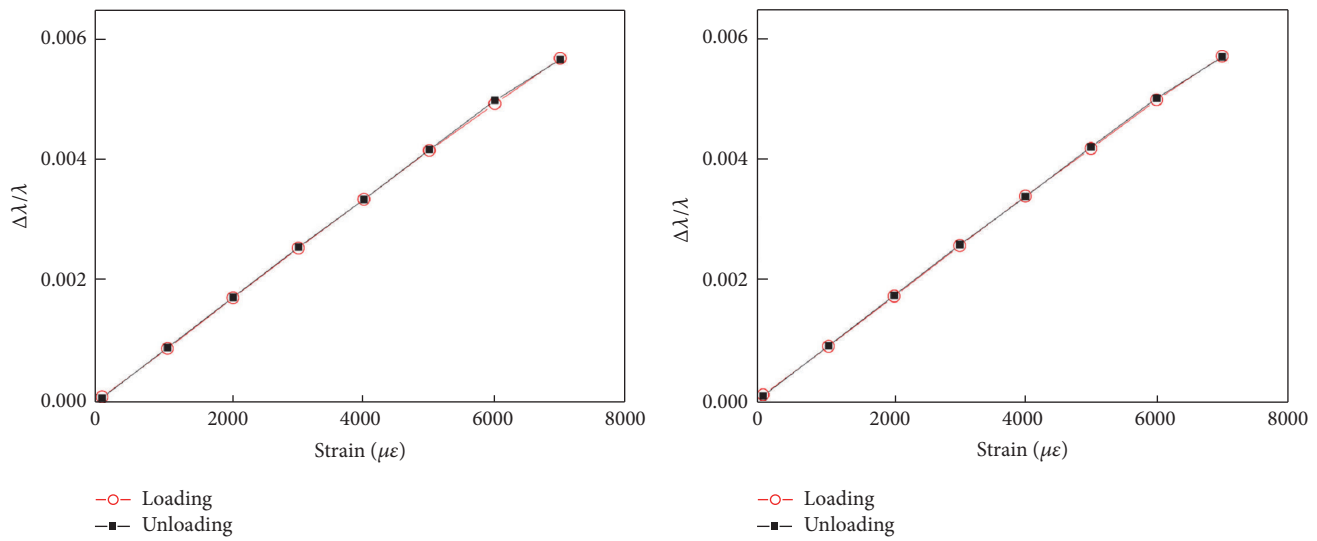
$$\begin{aligned} \varepsilon_h(\overline{x}) &= \frac{\sigma_h(\overline{x})}{E_h} \\ \varepsilon_c(\overline{x}) &= \frac{\sigma_c(\overline{x})}{E_c} \\ \gamma_a(r, \overline{x}) &= \frac{\tau(r, \overline{x})}{G_a} \quad r_{ap} \le r \le r_h \\ \gamma_p(r, \overline{x}) &= \frac{\tau(r, \overline{x})}{G_p} \quad r_c \le r \le r_{ap}. \end{aligned} \tag{14}$$



(a) FBG (b) Protecting layer

(c) Adhesive layer

FIGURE 4: Schematic diagram of an infinitesimal.



(a) The strain sensing sensitivity of OFBG-GFRP bars (b) The range and repeatability of OFBG-GFRP bars

FIGURE 5: The effect of FRP on FBG.

Therefore,

$$\begin{aligned} u_h(\overline{x}) &= \int_0^{\overline{x}} \varepsilon_h(\overline{x})\, d\overline{x} = \int_0^{\overline{x}} \frac{\sigma_h(\overline{x})}{E_h} d\overline{x} \\ u_c(\overline{x}) &= \int_0^{\overline{x}} \varepsilon_c(\overline{x})\, d\overline{x} = \int_0^{\overline{x}} \frac{\sigma_c(\overline{x})}{E_c} d\overline{x} \\ \Delta_a(\overline{x}) &= \int_{r_{ap}}^{r_h} \gamma_a(r,\overline{x})\, dr = \frac{1}{G_a}\int_{r_{ap}}^{r_h} \tau_a(r,\overline{x})\, dr \\ \Delta_p(\overline{x}) &= \int_{r_c}^{r_{ap}} \gamma_p(r,\overline{x})\, dr = \frac{1}{G_a}\int_{r_c}^{r_{ap}} \tau_a(r,\overline{x})\, dr. \end{aligned} \tag{15}$$

Formula (16) can be derived by means of substituting formulas (11) and (15) into formula (3), whose first derivative and second derivative are shown in formulas (17) and (18), respectively. The differential equation of interfacial shear stress $\tau_{ap}(r_{ap},\overline{x})$ in formula (18) is obtained based on formula (8) and the assumption that the axial stress of host materials $\sigma_h(\overline{x})$ is considered to be constant in case of short measured length.

$$\begin{aligned} &\int_0^{\overline{x}} \frac{\sigma_h(\overline{x})}{E_h} d\overline{x} \\ &= r_{ap}\tau_{ap}\left(r_{ap},\overline{x}\right)\left[\frac{1}{G_a}\ln\frac{r_h}{r_{ap}} + \frac{1}{G_p}\ln\frac{r_{ap}}{r_c}\right] \\ &+ \int_0^{\overline{x}} \frac{\sigma_c(\overline{x})}{E_c} d\overline{x} \end{aligned} \tag{16}$$

$$\frac{\sigma_h(\overline{x})}{E_h} = r_{ap}\left[\frac{1}{G_a}\ln\frac{r_h}{r_{ap}} + \frac{1}{G_p}\ln\frac{r_{ap}}{r_c}\right]\frac{\partial\tau_{ap}\left(r_{ap},\overline{x}\right)}{\partial\overline{x}} + \frac{\sigma_c(\overline{x})}{E_c} \tag{17}$$

$$\left[\frac{1}{G_a}\ln\frac{r_h}{r_{ap}} + \frac{1}{G_p}\ln\frac{r_{ap}}{r_c}\right]\frac{\partial^2\tau_{ap}^2\left(r_{ap},\overline{x}\right)}{\partial^2\overline{x}} - \frac{2}{E_c r_c^2}\tau_{ap}\left(r_{ap},\overline{x}\right) = 0, \tag{18}$$

where

$$\lambda_1^2 = \frac{2}{E_c r_c^2\left[\left(1/G_a\right)\ln\left(r_h/r_{ap}\right) + \left(1/G_p\right)\ln\left(r_{ap}/r_c\right)\right]}. \tag{19}$$

Therefore, formula (18) can be simplified as

$$\frac{\partial\tau_{ap}^2\left(r_{ap},\overline{x}\right)}{\partial^2\overline{x}} - \lambda_1^2\tau_{ap}\left(r_{ap},\overline{x}\right) = 0. \tag{20}$$

The general solution is obtained as follows:

$$\tau_{ap}\left(r_{ap},\overline{x}\right) = A\cosh\left(\lambda_1\overline{x}\right) + B\sinh\left(\lambda_1\overline{x}\right). \tag{21}$$

The axial force of the optical fiber can be expressed as follows:

$$N_c(\overline{x}) = \int_A \sigma_c(\overline{x})\, dx, \tag{22}$$

noting that $\sigma_c(0) = \sigma_c$.

And then,

$$\sigma_c(\overline{x}) = \sigma_c - \frac{2r_{ap}}{r_c^2}\int_0^{\overline{x}} \tau_{ap}\left(r_{ap},\overline{x}\right) d\overline{x}. \tag{23}$$

Therefore,

$$N_c(\overline{x}) = \pi r_c^2\sigma_c - 2\pi r_{ap}\int_0^{\overline{x}} \tau_{ap}\left(r_{ap},\overline{x}\right) d\overline{x}. \tag{24}$$

Introducing formula (21) into formula (24), the axial force of the optical fiber can be expressed in the following:

$$N_c(\overline{x}) = \pi r_c^2\sigma_c - 2\pi r_{ap}\frac{1}{\lambda}\left[A\,\mathrm{sh}\left(\lambda\overline{x}\right) + B\,\mathrm{ch}\left(\lambda\overline{x}\right) + B\right]. \tag{25}$$

On the basis of formula (6) and the axial force of the optical fiber being zero at $\overline{x} = l_f$, the following boundary conditions are introduced to calculate integration constants A and B:

$$\begin{aligned} N_c(0) &= \sigma_h\pi r_c^2\frac{E_c}{E_h} \\ N_c\left(l_f\right) &= 0 \\ A &= \frac{\sigma_c r_c^2\lambda}{2r_{ap}\mathrm{sh}\left(\lambda l_f\right)}, \\ B &= 0. \end{aligned} \tag{26}$$

Thus, the interfacial shear stress can be obtained as follows:

$$\tau_{ap}\left(r_{ap},\overline{x}\right) = \frac{\sigma_c r_c^2\lambda}{2r_{ap}\mathrm{sh}\left(\lambda l_f\right)}\mathrm{ch}\left(\lambda l_f\right) \tag{27}$$

$$\begin{aligned} \varepsilon_c(\overline{x}) &= \varepsilon_c(0)\times\left[1 - \frac{\mathrm{sh}\left(\lambda\overline{x}\right)}{\mathrm{sh}\left(\lambda l_f\right)}\right] \\ &= \varepsilon_h(0)\times\left[1 - \frac{\mathrm{sh}\left(\lambda\overline{x}\right)}{\mathrm{sh}\left(\lambda l_f\right)}\right]. \end{aligned} \tag{28}$$

In (28), $\varepsilon_h(0)$ is the axial strain of the host material at $\overline{x} = 0$.

This is the generic expression of strain transfer mechanism for the embedded optical fiber sensors; as for embedded FRP-OFBG strain sensors, adjustments shown as follows should be made with specific application $r_{ap} \to r_c$, $G_a \to G_{\mathrm{FRP}}$; therefore, formula (19) is converted to the following formula:

$$\lambda_1^2 = \frac{2G_{\mathrm{FRP}}}{E_c r_c^2\ln\left(r_h/r_c\right)}. \tag{29}$$

We defined the average strain measurement of FBG sensors $\overline{\varepsilon}_c(\overline{x})$ and the average strain of structures $\overline{\varepsilon}_h(\overline{x})$. The interfacial

Table 1: Interlaminar shear strength of BFRP and GFRP rebar under salt and acid condition under different age.

FRP bars	Corrosion solution	Items	20 days	240 days
BFRP	Salt	Average value	45.1	31.2
		Variance	2.2	2.5
		Deterioration rate	−9.8	−37.6
	Acid	Average value	43.3	29.4
		Variance	2.4	3.3
		Deterioration rate	−13.4	−41.2
	Blank specimen	Interlaminar shear strength (MPa)	50	
		Variance (MPa)	1.5	
GFRP	Salt	Average value	48.2	39.6
		Variance	1.8	2.5
		Deterioration rate	−6.2	−23
	Acid	Average value	46.8	37.9
		Variance	2.1	3.2
		Deterioration rate	−8.9	−26.3
	Blank specimen	Interlaminar shear strength (MPa)	50	
		Variance (MPa)	1.5	

strain transferring error rate η and the error correction factor k are expressed as follows:

$$\begin{aligned} \bar{\varepsilon}_c(\overline{x}) &= \frac{\int_0^{l_f} \varepsilon_c(\overline{x})\, d\overline{x}}{L_f} \\ \bar{\varepsilon}_h(\overline{x}) &= \frac{\int_0^{l_f} \varepsilon_h(\overline{x})\, d\overline{x}}{L_f} \\ \eta &= \frac{\left|\bar{\varepsilon}_c(\overline{x}) - \bar{\varepsilon}_h(\overline{x})\right|}{\bar{\varepsilon}_h(\overline{x})} = \frac{\cosh\left(\lambda_1 L_f\right) - 1}{\lambda_1 L_f \sinh\left(\lambda_1 L_f\right)} \\ k &= \frac{1}{1-\eta}. \end{aligned} \tag{30}$$

In (29), G_{FRP} is the shear modulus of the FRP composites; E_c is the elastic modulus of the optical FBG; r_h is the distance between optical fiber and matrix material; r_c is the outer diameter of the optical fiber; $2L_f$ is the effective working length of the optical fiber.

Calculations proved that the interfacial strain transferring error rate of CFRP-OFBG bars for the shear modulus being greater than 12 GPa and the outer diameter being Φ4~Φ10 mm is 1.92~2.16%, and the corresponding error correction factor k is 1.02~1.022. By contrast, the interfacial strain transferring error rate of GFRP-OFBG bars for the shear modulus being greater than 4.9 GPa and the outer diameter being Φ4~Φ10 mm is 3.11~3.5%, and the corresponding error correction factor k is 1.034~1.036. From the results provided in this section, we can conclude that the test accuracy of FRP-packaged optical fiber strain sensors is sufficient for civil engineering structures with significant material discrete characteristic and can be applied in the practical structures directly without any error correction.

3. Corrosion Durability Test of the FRP Bars

Corrosion durability tests of the GFRP bars and BFRP bars under the condition of acid and alkali salt were conducted in this section, with the comparative study of degeneration of interlaminar shear properties and tensile properties after corrosion. The corrosion solution with different ingredient and mixing ratio was exploited to simulate the acid, alkali, and salt corrosion environment in the practical civil engineering.

3.1. Variation in Tensile Properties of FRP Bars. The ultimate tensile strength and tensile modulus of the BFRP bars and GFRP bars after being corroded in the acid, alkali, and salt solution for 20 days and 240 days are summarized in Table 1.

The variations of the ultimate tensile strength and tensile modulus of the BFRP bars and GFRP bars with corrosive time are presented in Figures 6 and 7. Since BFRP bars were severely damaged in the alkali solution in 20 days, the tensile strength of both FRP bars was not measured due to the lack of a control group. It is shown that both of the ultimate tensile strengths increase with the corrosive time in the acid and salt solution. For a particular corrosive time, loss of the ultimate tensile strength of BFRP bars is more than that of GFRP bars in either acid solution or salt solution; besides, loss of the ultimate tensile strength in the acid solution is more than in the salt solution for both of the FRP bars. Whereas the tensile modulus of BFRP bars increases with the corrosive time, on the contrary, the tensile modulus of GFRP bars decreases with the corrosive time. In addition to that, the change rate of the tensile modulus for both FRP bars in the acid solution is more than that in the salt solution. Although the corrosion damage of basalt fibers does not occur, the interface between fibers and the resin is corrosively damaged by the acid solution and the salt solution, which cannot transfer the tension stress between fibers effectively, inducing stress concentration in a portion of fibers, followed by the decrease of ultimate tensile

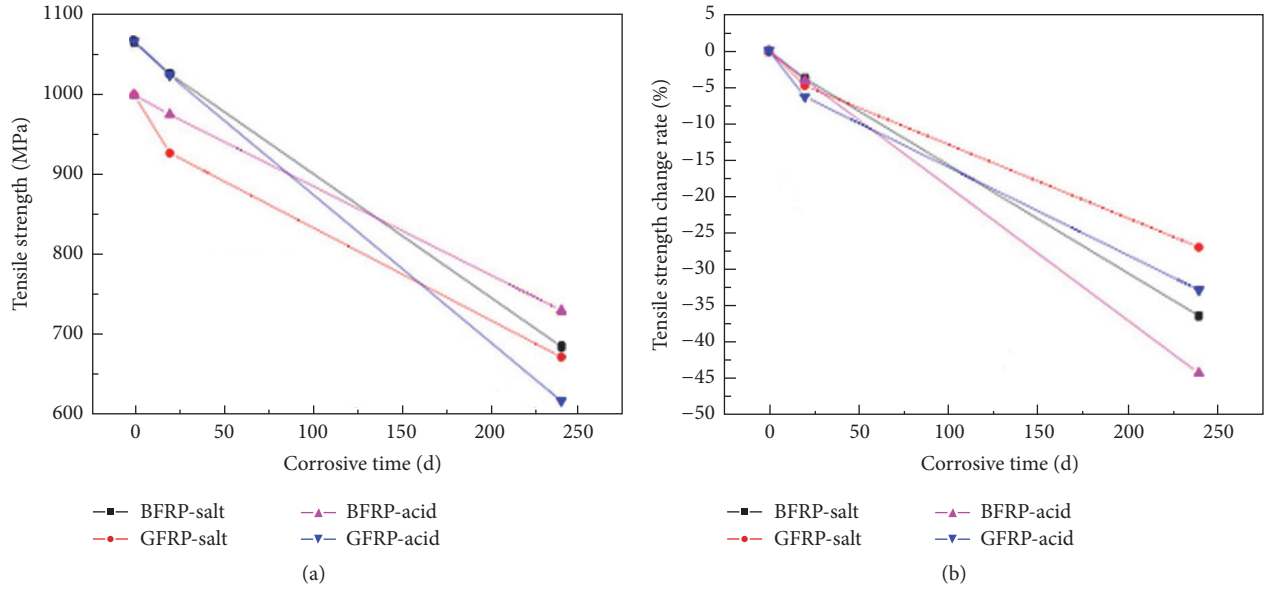


Figure 6: Curves of tensile strength versus age of BFRP and GFRP rebar under acid and salt condition.

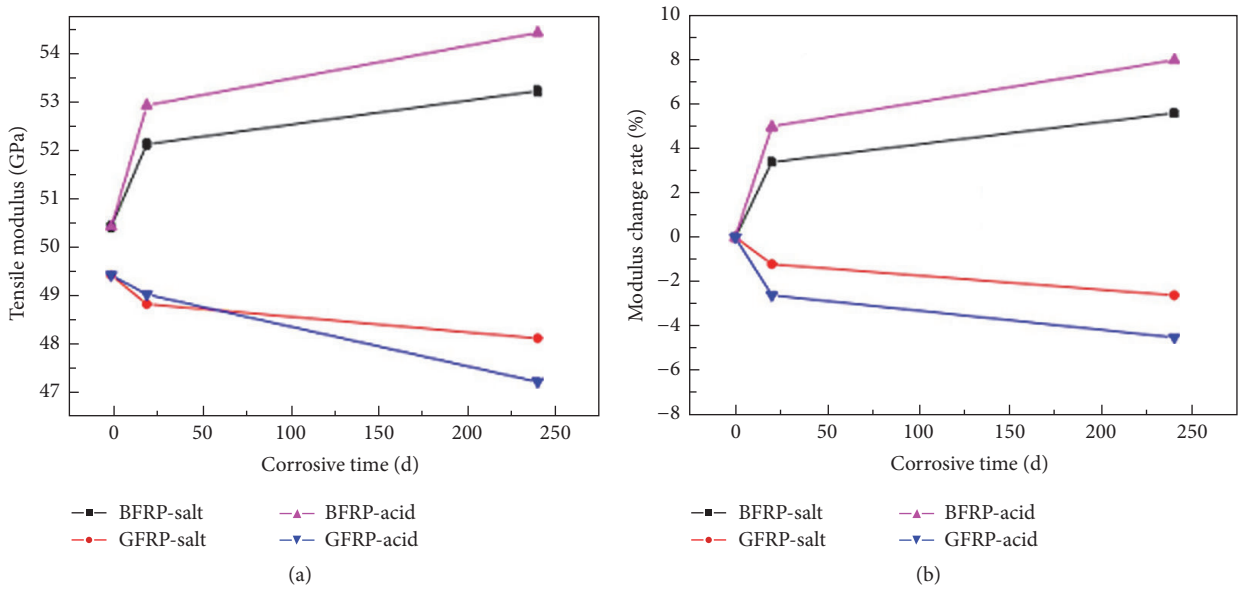


Figure 7: Curves of tensile modulus versus age of BFRP and GFRP rebar under acid and salt condition.

strength of BFRP bars. In addition, the relative slip between fibers and resin is also induced by the deterioration of the adhesive interface.

3.2. Variation in Interlaminar Shear Strength of FRP Bars. The interlaminar shear strength of BFRP and GFRP rebar under salt and acid condition under different age is summarized in Table 1.

The interlaminar shear strength of FRP bars is the important parameter reflecting the bonding properties between the resin and the fibers throughout the short beam shear test. The interlaminar shear strength of BFRP bars and GFRP bars decreases with corrosive time, and degeneration of the interlaminar shear strength for both FRP bars in the acid solution is more than that in the salt solution. In addition, loss of the interlaminar shear strength of BFRP bars is more than in the GFRP bars in both acid solution and salt solution. The variation of interlaminar shear strength mentioned above reveals the deterioration of the adhesive interface between fibers and the resin induced by the corrosive solution, as shown in Figure 8.

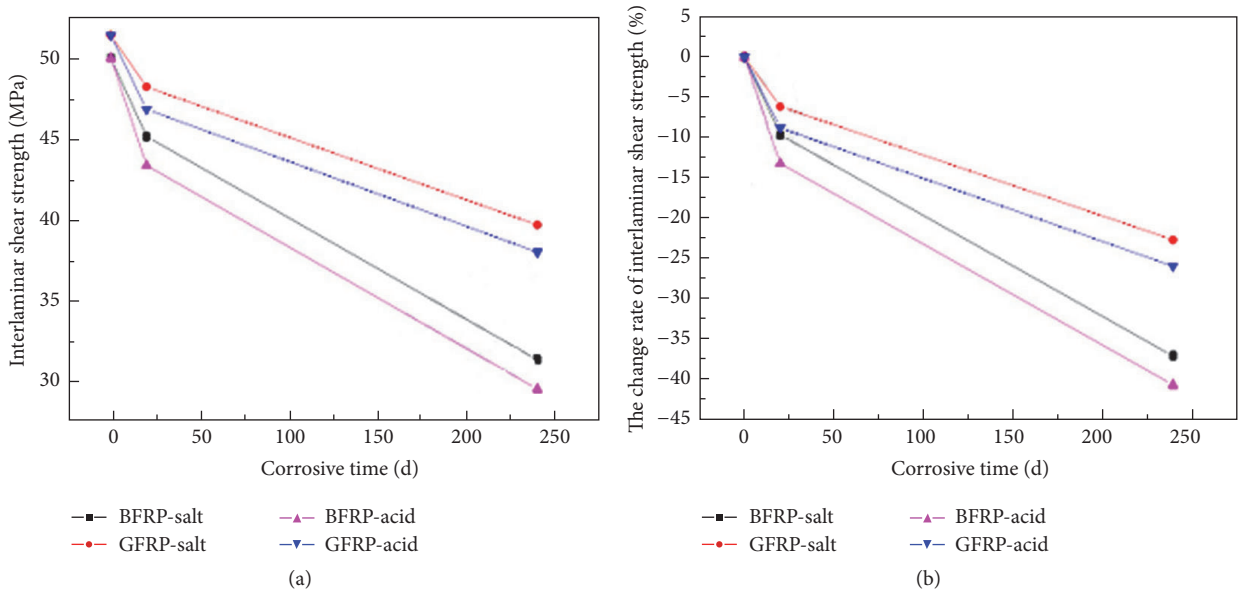


Figure 8: Curves of interlaminar shear strength and age in BFRP and GFRP rebar under acid and salt condition.

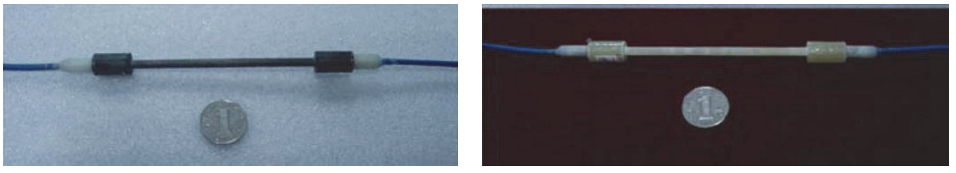


Figure 9: Embeddable FRP-FBG strain sensor.

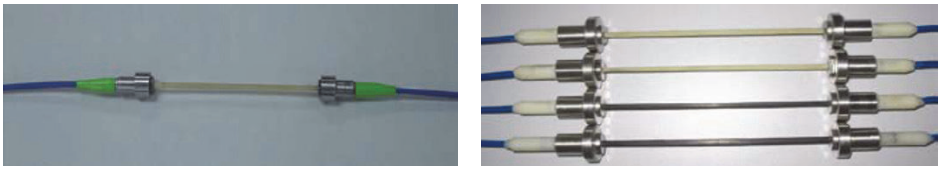


Figure 10: Embeddable FRP strain sensor with fixtures at two ends.

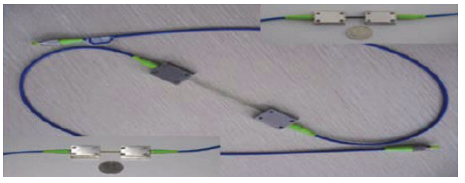


Figure 11: Surface-welded FRP strain sensor.

4. FRP-Packaged Fiber Bragg Grating (FBG) Strain Sensors

4.1. Package Structure of FRP-Packaged FBG Strain Sensor. Based on research achievements on the FRP-OFBG bars (Figures 20 and 21), a series of embedded FRP-packaged fiber Bragg grating strain sensors, FRP-packaged fiber Bragg grating strain sensors with expanding ends, surface-welding FRP-packaged fiber Bragg grating strain sensors, long gauge embedded FRP-packaged fiber Bragg grating strain sensors, and 3D FRP-packaged fiber Bragg grating strain sensors were developed for the demand of engineering test, shown as Figures 9–13.

Figure 12: Long gauge embeddable FRP strain sensor.

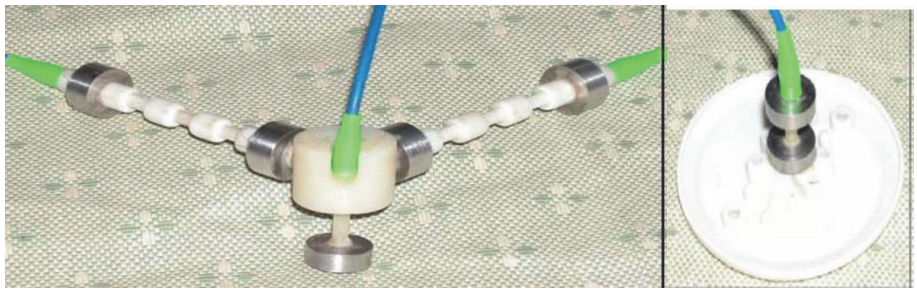


Figure 13: Three-dimensional FBG-packaged strain sensor.

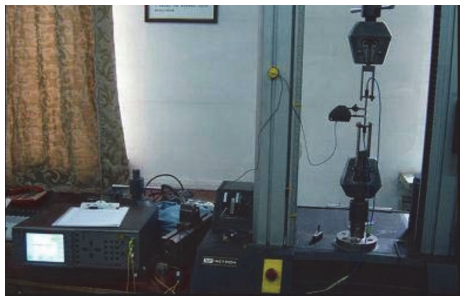


Figure 14: Test of FRP-packaged FBG strain sensor.

4.2. Sensing Performance of FRP-Packaged Fiber Grating Strain Sensors. Sensing performance tests in the presented study were conducted on the material testing machine Instron-5569, and the center wavelength of fiber Bragg grating was demodulated by SI-720 produced by MOI company with accuracy of 1 pm. Strain of FRP-packaged fiber grating strain sensors was measured by the extensometer, with the resolution of 1 $\mu\varepsilon$. In order to eliminate the effect on the variation of center wavelength of temperature, laboratory temperature was maintained as constant at 25°C by air conditioning. The test setup is shown in Figure 14. 3~5 loading-unloading cycles were conducted on each type of sensor mentioned earlier, and their variation was recorded in center wavelength and strain, respectively, for analyzing the sensing performance index, shown as Figure 15.

It is shown that, from the test results, the measurement scale of this series of FRP-packaged fiber grating strain sensors is more than 5000 $\mu\varepsilon$ (the maximum strain can reach 12000 $\mu\varepsilon$), the test accuracy is 1 $\mu\varepsilon$, the repeatability is less than 1.0%, linearity is less than 1.0%, and hysteresis is less than 0.5%.

4.3. Sensing Performance of FRP-Packaged Fiber Grating Strain Sensors in Harsh Environment

4.3.1. Strain Sensing Performance in High-Temperature Environments. The service temperature ranged from −40°C to 60°C for most civil engineering structures, but higher temperature occurs in some structural elements; therefore, it is meaningful to research the sensing performance of FRP-packaged fiber grating strain sensors under high-temperature conditions, due to the undesirable high temperature performance of FRP material.

This experiment was conducted on the MTS testing machine, as shown in Figure 14. Strain and variation of center wavelength were measured by the high-temperature extensometer and SI-720 optical fiber grating demodulator produced by MOI company, and the test setup is shown in

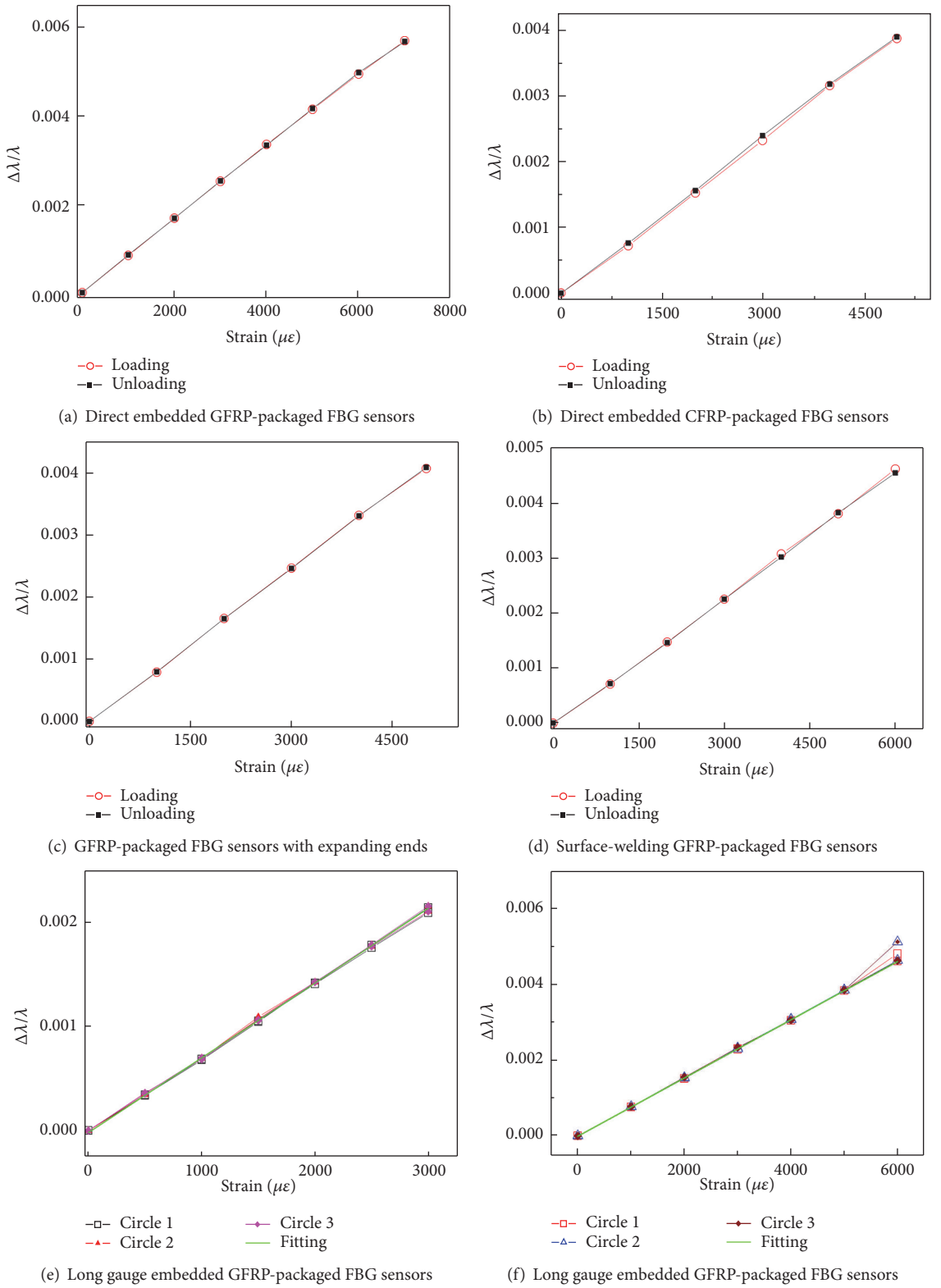


Figure 15: Properties of the FRP-packaged FBG strain sensors.

Figure 16: Test setup of GFRP-packaged FBG at high temperature.

Figure 16. In the presented experiment, the sensing performances, such as linearity and repeatability, at 20°C, 40°C, 60°C, 80°C, 100°C, 120°C, and 140°C, which were controlled by the high-temperature furnace, were investigated. Experiment results for one of the testing specimens are shown in Figure 17.

It can be seen from the experimental results that the strain sensitivity of the GFRP-packaged FBG sensors at 40°C, 60°C, 80°C, 100°C, 120°C, and 140°C is around 1.2 pm/$\mu\varepsilon$, being close to that at 20°C normal environment. Therefore, we can conclude that the performance of FRP-packaged optical fiber strain sensors does not degenerate under the ambient temperature of 140°C.

4.3.2. Fatigue Performance Experiment of FRP-Packaged Optical Fiber Strain Sensors. In order to examine the stability and reliability of FRP-packaged optical fiber strain sensors applied on the long-term monitoring of civil engineering structures, fatigue performance experiment of this series of FRP-packaged optical fiber strain sensors was performed on the fatigue testing machine MTS-810, with the test frequency selected as 50 Hz. For the purpose of accelerating the fatigue test, the strain amplitude was determined as 2000 ± 1000 $\mu\varepsilon$ and 3000 ± 1000 $\mu\varepsilon$. The center wavelength response of FBG sensors was recorded at intervals of a fixed time in order to compare variation of that at each moment. A significant change of the center wavelength variation at a moment with a comparison to that at the initial moment infers that the fatigue damage is accumulated in the FRP-packaged optical fiber strain sensors, and this phenomenon could be employed to evaluate fatigue reliability of FRP-packaged optical fiber strain sensors. In this paper, two surface-welding GFRP-packaged optical fiber strain sensors and a direct embedded CFRP-packaged optical fiber strain sensor were randomly selected as testing specimens to investigate the sensing performance after high-cycle fatigue; besides, three GFRP-packaged optical fiber strain sensors with expanding ends were also randomly chosen to examine the sensing performance after low-cycle fatigue. During the experiment, the center wavelengths of FBG sensors were recorded by SI-425 optical fiber grating demodulator produced by Micron Optics company, in which the instrument wavelength resolution is 5 pm and scanning frequency is 250 Hz. The fatigue test setup is shown in Figure 18. Part of the experiment results of all the specimens are summarized in Table 2. Figure 19 shows segmenting time history curves of one specimen.

Table 2: The testing results at low cycle fatigue.

Specimen	Strain amplitude ($\mu\varepsilon$)	Cycles (10000 times)	Wavelength variation (nm)
Surface welding GFRP (1)	2000~3000	100	7
Surface welding GFRP (2)	2000~3000	100	9
Surface welding GFRP (2)	2000~4000	160	6
Direct embedded CFRP	2000~3000	100	7

We can see from experimental results that there is no strain decrease of FRP-packaged FBG strain sensors during the fatigue test and that no member resistance reduction and no obvious damage occurred during the test. After multiple fatigue cycles, none of the significant drifts of the fiber grating center wavelength appeared. It is shown that FRP composites protect the sensing element effectively, and the stability of FRP-packaged FBG strain sensors kept excellent, which is suited for civil engineering applications.

Moreover, the fatigue tests of FRP-packaged FBG strain sensors at high strain amplitude were conducted, and a comparison was made between the sensors' strain sensing performance after certain fatigue circles and that of the control group. In this test, the strain amplitude of sensor 1 is 3000 ± 2000 $\mu\varepsilon$ and that of sensor 2 is 4500 ± 1500 $\mu\varepsilon$; test results are shown in Figures 20(a)–20(d).

It can be seen that FRP-packaged strain sensors, which have experienced a certain number of fatigue cycles, retain the original strain sensing performance, such as good linearity and repeatability, as well as a longer fatigue life.

4.3.3. Corrosion Durability Test of FRP-Packaged Optical Fiber Bragg Grating Strain Sensors. In order to verify the sensor's durability index, different types of FRP-packaged optical fiber Bragg grating (FBG) strain sensors were placed in salt spray chamber, with the working temperature of 35°C and the salt

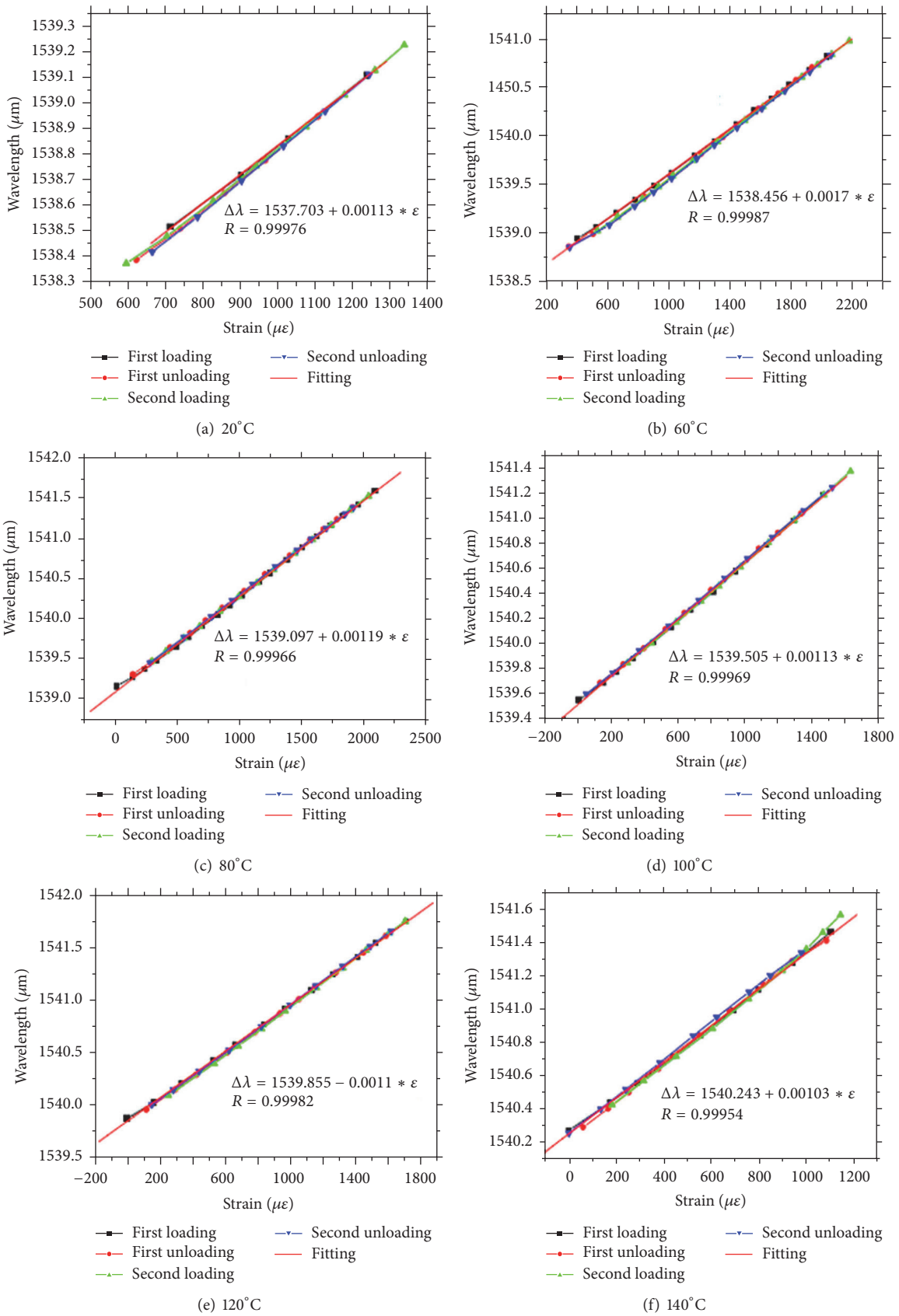


Figure 17: Relationship between wavelength changes and strain at different temperatures of sensor one.

Figure 18: Fatigue test of FRP-packaged FBG strain sensor.

Table 3: Monitoring items and technical parameters of FRP-packaged optical fiber Bragg grating sensors.

Object	Monitoring items	Sensors				Numbers
		Method	Frequency	Accuracy	Scale	
Main-beam	Steel strain	FBG	20 Hz	1 $\mu\varepsilon$	±1000 $\mu\varepsilon$	40
Main-beam	Steel temperature	FBG	1 time/h	0.5°C	−20°C~+70°C	15
Main-tower	Concrete strain	FBG	20 Hz	1 $\mu\varepsilon$	±1000 $\mu\varepsilon$	8
Main-tower	Concrete temperature	FBG	1 time/h	0.5°C	−20°C~+70°C	8

spray of 3.5% NaCl solution, shown as Figure 21. It can be seen from the results of corrosion test that the corrosion resistance of the metal packaged FBG sensors is far less than the FRP-packaged FBG sensors. This series of FRP-packaged FBG strain sensors overcomes the insurmountable durability issues compared with the traditional adhesive packaged FBG strain sensors, possessing outstanding advantages, such as simple laying process in practical engineering, large measurement scale (more than 5000 $\mu\varepsilon$; the maximum can reach 10000 $\mu\varepsilon$), excellent durability, high accuracy (1~2 $\mu\varepsilon$), and low error correction factor. Furthermore, it can be easily customized according to custom.

Key performance indicators of the series of FRP-packaged optical fiber Bragg grating strain sensors are summarized as follows: measurement scale is more than 5000 $\mu\varepsilon$, test accuracy is 1 $\mu\varepsilon$, repeatability is less than 1%, linearity is less than 1.0%, hysteresis is less than 0.5%, and so forth. No fatigue properties are found under more than one million times of fatigue cycles, with strain amplitude of 3000 ± 1000 $\mu\varepsilon$.

This series of FRP-packaged FBG strain sensors is particularly suitable for the internal strain measurement of the concrete structures, besides the whole-process monitoring of civil engineering structures, including construction stage, completion test stage, and in-service stage. Still, it also can be easily used for road engineering, geotechnical engineering, and so forth throughout the reconstruction design.

5. Applications in Aizhai Bridge

Aizhai Bridge is a super long suspension bridge with separated towers and beams and with span arrangement of 242 m + 1176 m + 116 m. In order to obtain the information on structure behavior in service, a structural health monitoring (SHM) system was established for providing information about conditions such as strain, temperature, acceleration, deflection, wind velocity, cable force, and humidity. In this SHM system, surface-welding GFRP-packaged optical fiber strain and temperature sensors (shown as Figure 9) are used for strain and temperature measuring of the main beam and main tower, and the measuring point arrangement is shown in Figure 22. The monitoring items and technical parameters of FRP-packaged optical fiber Bragg grating sensors are listed in Table 3. The data, acquired and processed by FBG demodulator, is transferred to the data receiving server, in which the data can be retained, managed, and arithmetically processed.

6. Conclusions

In this study, a series of FRP-packaged optical fiber Bragg grating strain sensors to completely meet the requirements of the rough civil engineering infrastructures are introduced, and their sensing performances under normal environment and harsh environment are experimentally investigated; the following conclusions can be drawn:

(1) Based on stain transfer mechanism, it is theoretically proved that the testing accuracy of FRP-packaged optical fiber strain sensors is sufficient for civil engineering structures and can be applied in the practical structures directly without any error correction.

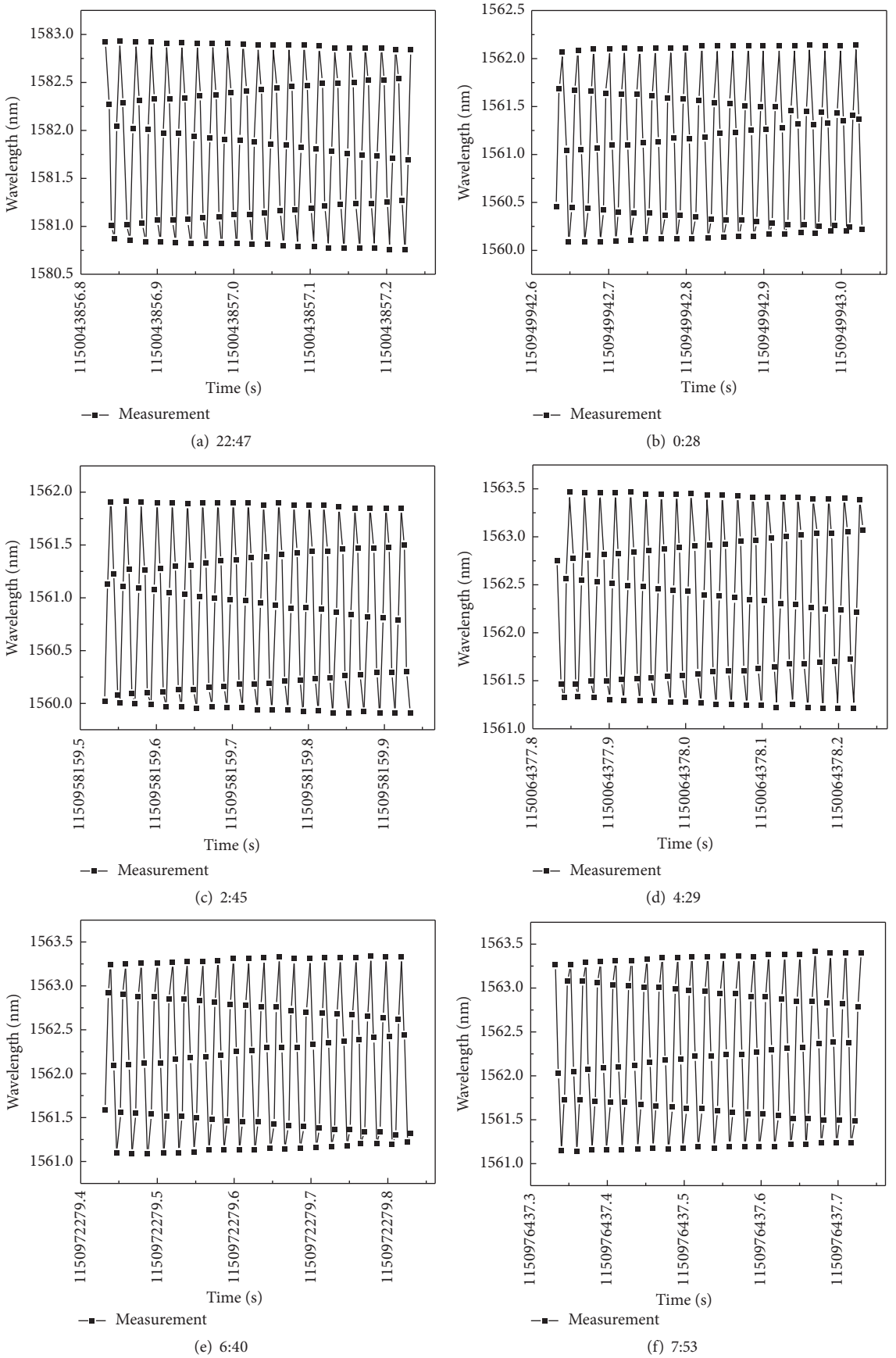


Figure 19: Response of the FBG wavelength at different times.

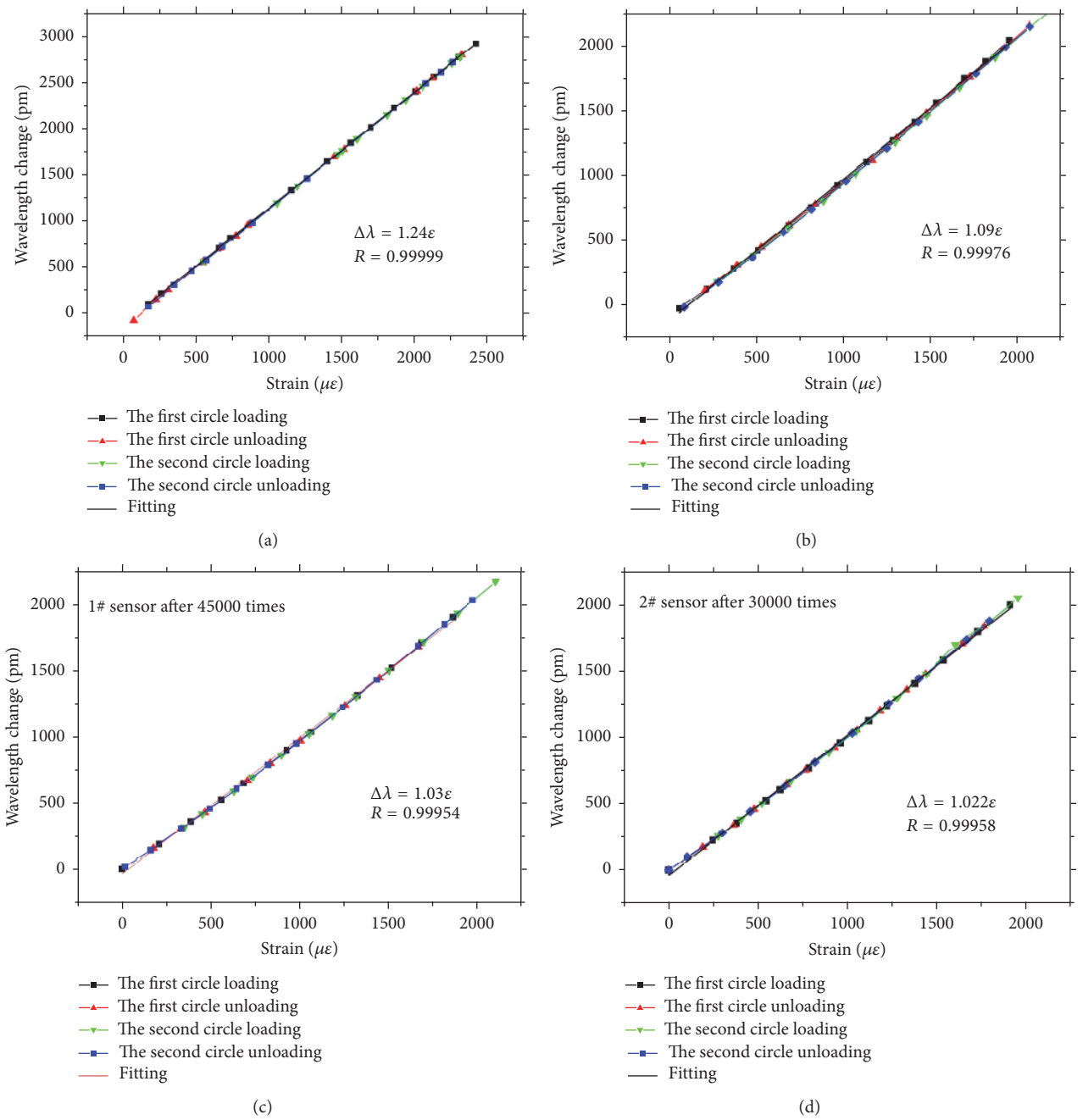


Figure 20: In situ diagram of FRP-packaged OFBG sensors installation.

(2) FRP composites do not change the sensing performance of the FBG sensor; furthermore, the measurement scale significantly increased to 8000~12000 $\mu\varepsilon$.

(3) FRP-packaged optical fiber FBG strain sensors do not degenerate under the ambient temperature of 140°C and maintain excellent linearity and repeatability as the bare FBG in the case of experiencing a certain number of fatigue cycles. Besides, they have superior corrosion resistance compared with metal packaged optical fiber strain sensors, which makes them particularly suitable for the internal strain measurement of the concrete structures and the whole-process monitoring for civil engineering structures, including construction stage, completion test stage, and in-service stage.

Competing Interests

The authors declare that they have no financial and personal relationships with other people or organizations that can inappropriately influence this work and that there is no professional or other personal interest of any nature or

Figure 21: Schematic diagram of FRP-packaged OFBG sensors installation.

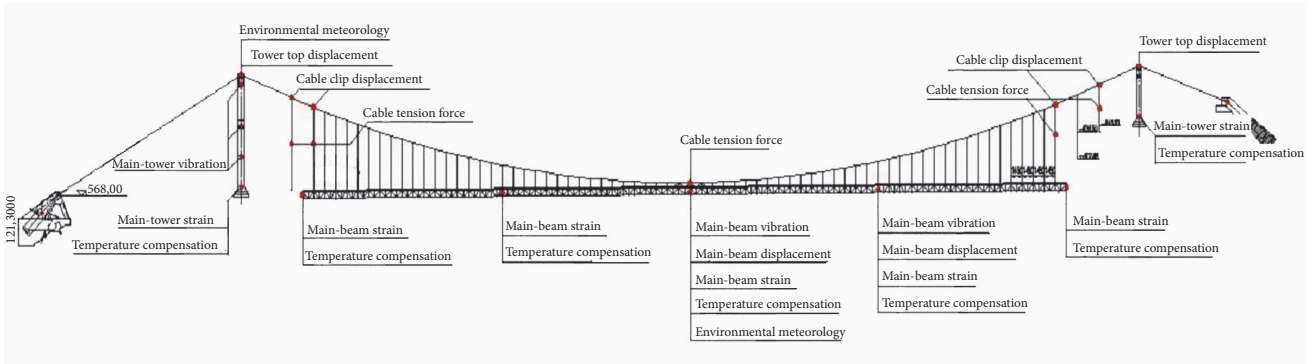


Figure 22: Measuring point arrangement of all kinds of sensors.

kind in any product, service, and/or company that could be construed as influencing the position presented in or the review of the manuscript.

References

[1] J. P. Ou, "Damage accumulation and safety evaluation for important large infrastructures," in *21st Century's Chinese Mechanics-9th Science Association Reports of 'Forum for Youth Scientist'*, pp. 179–189, Tsinghua University Press, Beijing, China, 1996.

[2] G. W. Housner, L. A. Bergman, T. K. Caughey et al., "Structural control: past, present, and future," *Journal of Engineering Mechanics*, vol. 123, no. 9, pp. 897–971, 1997.

[3] G. Meltz, W. Morey, and W. Glenn, "Formation of bragg gratings in optical fibers by a transverse holographic method," *Optics Letters*, vol. 14, no. 15, pp. 823–825, 1989.

[4] A. A. Mufti, "FRPs and FOSs lead to innovation in Canadian civil engineering structures," *Construction and Building Materials*, vol. 17, no. 6-7, pp. 379–387, 2003.

[5] Z. Zhou, T. Graver, and J. P. Ou, "Techniques of advanced FBG sensors: manufacturing, demodulation, encapsulation and their application in the structural health monitoring of bridges," *Pacific Science Review*, vol. 5, no. 1, pp. 116–121, 2003.

[6] Z. Zhou, B. Wang, and J. P. Ou, "Local damage detection of RC structures with distributive FRP-OFBG sensors," in *Proceedings of the 2nd International Workshop on Structural Health Monitoring of Innovative Civil Engineering Structures*, vol. 22-23, pp. 205–214, Winnipeg, Canada, September 2004.

[7] Z. Zhou and J. P. Ou, "Development of FBG sensors for Structural Health Monitoring in civil infrastructures," in *Proceeding of the North American Euro-Pacific Workshop 'Sensing Issues in Civil Structural Health Monitoring'*, Waikiki, Hawaii, USA, 2004.

[8] Z. Zhou, C. G. Lan, and J. P. Ou, "A novel ice-pressure sensor based on dual FBGs," in *Smart Structures and Materials 2005: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems*, vol. 5765 of *Proceedings of SPIE*, San Diego, Calif, USA, 2005.

[9] Z. Zhou, J. Liu, H. Li, and J. Ou, "A new kind of high durable traffic weighbridge based on FBG sensors," in *17th International Conference on Optical Fibre Sensors*, vol. 5855 of *Proceedings of SPIE*, pp. 735–738, Bruges, Belgium, May 2005.

[10] J. P. Ou and Z. Zhou, "Techniques of optical fiber Bragg grating smart sensors and intelligent monitoring systems of infrastructures," in *Proceedings of the 1st International Workshop on Advanced Smart Materials and Smart Structures Technology*, p. 23, Honolulu, Hawaii, USA, June 2003.

[11] Z. Zhou, J. P. Ou, and B. Wang, "Smart FRP-OFGB bars and their application in reinforced concrete beams," in *Proceedings of the 1st International Conference on Structural Health Monitoring and Intelligent Structure*, vol. 13–15, pp. 861–866, Tokyo, Japan, November 2003.

[12] H. Tsutsui, A. Kawamata, T. Sanda, and N. Takeda, "Detection of impact damage of stiffened composite panels using embedded small-diameter optical fibers," *Smart Materials and Structures*, vol. 13, no. 6, pp. 1284–1290, 2004.

[13] C.-Y. Ryu, J.-R. Lee, C.-G. Kim, and C.-S. Hong, "Buckling behavior monitoring of a composite wing box using multiplexed and multi-channeled built-in fiber Bragg grating strain sensors," *NDT & E International*, vol. 41, no. 7, pp. 534–543, 2008.

[14] S.-I. Takeda, Y. Aoki, and Y. Nagao, "Damage monitoring of CFRP stiffened panels under compressive load using FBG sensors," *Composite Structures*, vol. 94, no. 3, pp. 813–819, 2012.

[15] K. I. Tserpes, V. Karachalios, I. Giannopoulos, V. Prentzias, and R. Ruzek, "Strain and damage monitoring in CFRP fuselage panels using fiber Bragg grating sensors. Part I: design, manufacturing and impact testing," *Composite Structures*, vol. 107, pp. 726–736, 2014.

[16] T. C. Triantafillou, "Shear strengthening of reinforced concrete beams using epoxy-bonded FRP composites," *ACI Structural Journal*, vol. 95, no. 2, pp. 107–115, 1998.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 8085407, 8 pages
http://dx.doi.org/10.1155/2016/8085407

Research Article

Light-Weight and Versatile Monitor for a Self-Adaptive Software Framework for IoT Systems

Young-Joo Kim,[1] Jong-Soo Seok,[1] YungJoon Jung,[1] and Ok-Kyoon Ha[2]

[1]*Electronics and Telecommunications Research Institute, Embedded SW Platform Research Section, Deajeon 34129, Republic of Korea*
[2]*Department of Aeronautics & Software Engineering, Kyungwoon University, Gumi 39160, Republic of Korea*

Correspondence should be addressed to Ok-Kyoon Ha; okha@ikw.ac.kr

Received 15 April 2016; Accepted 7 November 2016

Academic Editor: Antonio Fernández-Caballero

Today, various Internet of Things (IoT) devices and applications are being developed. Such IoT devices have different hardware (HW) and software (SW) capabilities; therefore, most applications require customization when IoT devices are changed or new applications are created. However, the applications executed on these devices are not optimized for power and performance because IoT device systems do not provide suitable static and dynamic information about fast-changing system resources and applications. Therefore, this paper proposes a light-weight and versatile monitor for a self-adaptive software framework to automatically control system resources according to the system status. The monitor helps running applications guarantee low power consumption and high performance for an optimal environment. The proposed monitor has two components: a monitoring component, which provides real-time static and dynamic information about system resources and applications, and a controlling component, which supports real-time control of system resources. For the experimental verification, we created a video transport system based on IoT devices and measured the CPU utilization by dynamic voltage and frequency scaling (DVFS) for the monitor. The results demonstrate that, for up to 50 monitored processes, the monitor shows an average CPU utilization of approximately 4% in the three DVFS modes and demonstrates maximum optimization in the Performance mode of DVFS.

1. Introduction

Rapid growth in information and communications technology (ICT) has resulted in the development of various types of Internet of Things (IoT) devices and applications for the industry, home, and other sectors. However, such IoT devices have different hardware (HW) and software (SW) capabilities. The HW capability is mainly influenced by the number of CPU cores or the CPU clock speed. Further, battery capacity is important because IoT devices do not generally use external power. Therefore, many researchers have considered the relation between performance and power. For example, if a system allocates many CPU cores to a program, the program has high performance but its power consumption is not efficient. The SW capability of an IoT device is mainly determined by the number of running applications because running applications can affect system performance, power, and so forth. Hence, these running applications must be customized when the devices are changed or new applications are executed on the device. However, the applications are not optimized with respect to performance, power, and so forth because IoT device systems do not suitably provide static/dynamic information for fast-changing system resources and applications.

Therefore, in this manuscript, we propose a light-weight and versatile monitor for a self-adaptive software framework; the proposed monitor and the framework can automatically control system resources according to the system status. The proposed monitor can function with small-scale systems (e.g., IoT devices and embedded devices) and large-scale systems (e.g., PC and rich systems); the monitor has a light-weight design. In order to support the self-adaptive software framework, the monitor helps running applications to guarantee low power and high performance, thus creating an optimal environment. The proposed monitor has two components: a monitoring component and a controlling component. The monitoring component provides static and dynamic information about the systems and applications

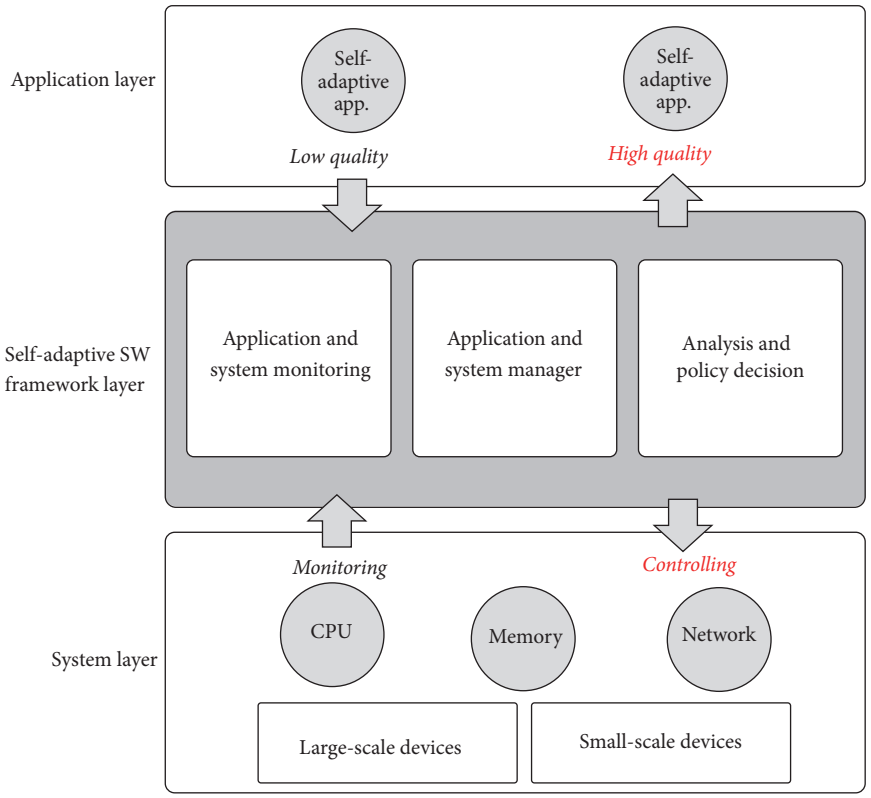


Figure 1: Concept of self-adaptive software framework.

in real-time. Static information is meaningful data that are already fixed in systems and applications, and dynamic information is meaningful data that change during the execution of systems and applications. The controlling component helps in the control of system resources (e.g., CPU, memory, network, etc.) in real-time. The control functions defined are CPU on/off, CPU frequency control, and network bandwidth control. For the experimental verification of the proposed monitor, we created a video transport system based on IoT devices and measured the CPU utilization for the monitor. The results showed that, for up to 50 monitored processes, the average CPU utilization of the monitor is approximately 4% in the three DVFS modes. Further, we observed that the monitor shows maximum optimization in the performance DVFS.

The remainder of this manuscript is structured as follows: In Section 2, we introduce the concept of a self-adaptive software framework. In Section 3, we describe the lightweight and versatile monitor for the self-adaptive software framework. In Section 4, we demonstrate the potential of the monitor through a self-developed QoS guarantee system. Finally, we state our conclusion and outline our directions for future research.

2. Self-Adaptive Software Framework

A self-adaptive software framework [1, 2], which is a middleware to guarantee optimal QoS for each application executing in a system, can manage and control running applications during their life cycle in a real-time and dynamic manner. In order to manage and control these applications, the framework provides monitoring functions. These functions are typically called adaptive applications. These adaptive applications include at least two modules: a QoS generator such as heartbeat [3] and a performance container with various algorithms. The QoS generator is inserted into a monitoring point of an application before the execution of the application; then, during the execution of the application, the QoS generator periodically reports the QoS for the application. The performance container uses one of two approaches: the first approach is to change the input parameters that influence the performance of the application, and the second approach is to create one or more algorithms that can change according to the performance. The self-adaptive software framework examines the reported QoS and then controls the adaptive applications. The framework uses the static and dynamic information about the application and the system resources to adjust the optimal QoS performance.

Generally, the self-adaptive software framework is composed of modules that monitor application or system information and control system resources. Figure 1 shows the overall structure of the self-adaptive software framework. As shown in this figure, the self-adaptive software framework layer is located between the application layer and the system layer. The framework consists of three modules: *an application and system monitoring* module, *an application and system*

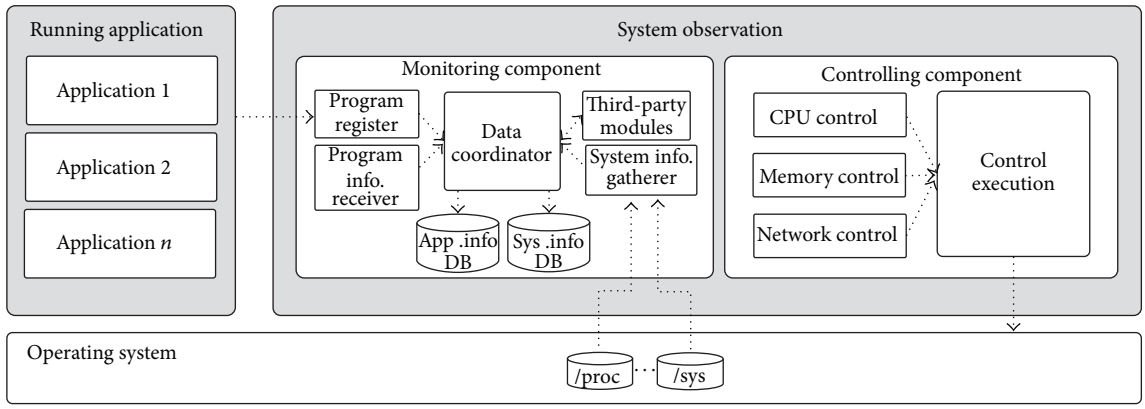


Figure 2: Block diagram of the light-weight and versatile monitor consisting of monitoring component.

manager module, and *an analysis and policy decision* module. In the application layer, when a self-adaptive application is running with low quality and the user requirement for the application is high quality, it is not easy for a general operating system to change the low-quality service to a high-quality service. However, the self-adaptive software framework can adjust the service quality by using the three modules. The application and system monitoring module gathers information from the system layer through monitoring, the application and system manager module provides the gathered information to the other modules, and the analysis and policy decision module controls the system resources or the flow of the application with the assistance of the manager. Our proposed monitor corresponds to the application and system monitoring module and the application and system manager module.

3. Self-Adaptive System Observation

In this manuscript, we present the *light-weight and versatile monitor*, which consists of the monitoring component and the controlling component. First, the monitor gathers static/dynamic information about the applications and systems; this information is required for the self-adaptive decision. Then, this information is provided to external modules and external devices. The proposed monitor is light-weight with respect to CPU utilization; therefore, it can be ported to diverse IoT devices such as embedded systems.

Figure 2 shows the block diagram of the light-weight and versatile monitor. In this figure, when applications are executed on a system, the monitor observes the status of these applications and the system in real-time and records this information in the *"application information DB"* and *"system information DB,"* respectively. The recorded information will be utilized to determine the optimization of the application to obtain high performance and low power consumption.

3.1. Monitoring Component. The monitoring component consists of five modules: *program register*, *program information receiver*, *system information gatherer*, *data coordinator*, and *third-party modules*. The program register and the program information receiver are responsible for collecting information about the application. During the execution of an application, two modules are connected to applications in TCP/IP. For example, when an adaptive application executes on a system, the application is registered in the light-weight and versatile monitor of the self-adaptive software framework by the program register; then, the program information receiver gathers static/dynamic information about the registered application. Table 1 provides a summary of the static/dynamic information for a running application.

Table 1: Application information (static/dynamic).

Static information	Dynamic information
Program ID	Allocated core
Program name	The number of threads
Program path	Thread list
Max QoS	Program status
Min QoS	Program time
Target QoS	Program/memory Utilization
Application sampling time	Network information (Tx, Rx)
Log file	Heartbeat rate

The system information gatherer collects static/dynamic information about the system resources in real-time. Further, the third-party modules receive information such as power, program characteristics, and internal kernel information from external modules or external devices. For example, system power and application power must be measured by external power equipment, and the measured power is transferred to the third-party modules of the self-adaptive system observation monitor. System static information represents the fixed values corresponding to the HW resources (e.g., CPU, memory, and network), and it is configured only once when a system functions. Generally, system static information has a unique value for the system when HW specifics remain unchanged. The value is determined by one-time data collection. System dynamic information represents values that change according to the system status. Most of these values can vary according to the system overhead, and they are updated periodically during the setup for information

Table 2: System information (static/dynamic).

Static information	Dynamic information
The number of cores	Core activity status
Max freq. of core	Core current freq.
Min freq. of core	The number of threads
Core MIPS	The number of processes
Available frequency	Core utilization
System memory	System memory
Network interface	Network packet (Tx, Rx)
DVFS	Power (CPU, GPU, and memory)

collection. The system static information shown in Table 2 includes CPU core, CPU frequency, memory, network, and DVFS. The system static information is the fixed information of the HW. If the HW capability is changed, the system static information is also updated. For this information, a data structure corresponding to each HW system is maintained separately. The system static information is updated only once when the light-weight and versatile monitor is executed. The system dynamic information shown in Table 2 includes the current state of the CPU core, CPU utilization, memory, network, and power. The dynamic information can be changed according to the current state of the system; further, users can change the resource values of the system, and thus, dynamic information is changed in real-time.

The data coordinator reorganizes the information collected from the program information receiver, the system information gatherer, and the third-party modules; then, this information is saved in the application information DB and the system information DB. The dynamic information about the application and systems is categorized into two data structures: instant data and calculation data. The instant data (e.g., CPU activity) can be used immediately in other modules, whereas the calculation data (e.g., utilization) cannot be used immediately owing to the need for additional operations. The static information about the applications and systems has a single data structure, which corresponds to instant data. These data are classified in real-time as application, system, or third-party data.

3.2. Controlling Component. The controlling component consists of four modules: *cpu control*, *memory control*, *network control*, and *control execution*. This component provides an environment to control system resources such as CPU, memory, and network. The various parameters of these system resources are as follows:

(i) CPU: core on/off, core frequency, thread affinity, DVFS (dynamic voltage frequency scaling) [4]

(ii) Memory: cache drop, minimum memory set

(iii) Network: bandwidth, packet drop

Control execution controls the CPU and memory by interacting with the proc file system. Control execution controls the network by using network control commands (e.g., tc command) or kernel model programs (e.g., network stack). Further, the controlling component provides monitoring interfaces and controlling interfaces such as libraries. Table 3 shows the monitoring interfaces and controlling interfaces. Thus, self-adaptive system observation uses the static/dynamic information about the applications and systems to enable optimal execution of applications.

Table 3: Monitoring interfaces and controlling interfaces.

Monitoring interface	Controlling interface
The number of cores	Core activity status
Max freq. of core	Core current freq.
Min freq. of core	The number of threads
Core MIPS	The number of processes
Available frequency	Core utilization
System memory	System memory
Network interface	Network packet

4. Experimental Verification

In this section, we present the experimental verification of the proposed monitor. We introduce a video transport system based on IoT devices; the monitor is applied to this system and the experimental results for the proposed light-weight monitor are presented.

4.1. Implementation. Figure 3 shows a video transport system with a self-adaptive SW framework including the light-weight and versatile monitor. The system consists of a three-tier structure (IoT devices ↔ set-top box (STB) ↔ host system). The IoT device that performs video capture and encode consists of an Intel Edison Board [5–8] with CAM; the STB, which performs video streaming, consists of an Embedded Board with Exynos 5422 [9] and contains the proposed monitor and a self-adaptive policy manager. This manuscript does not focus on the policy manager. The host system, which includes a user interface, consists of a mobile device with wireless communication such as Wi-Fi. The functioning of this system can be described as follows: video sources are generated from the Intel Edison Board with CAM in real-time; the generated sources are transferred to the STB through Wi-Fi; a video streaming server on the STB receives these video sources and decodes them in real-time; finally, the decoded sources are transferred to the mobile device through Wi-Fi.

In Figure 3, if a self-adaptive SW framework is not present in the STB, the STB cannot guarantee the transfer of a high bitrate video generated in the Intel Edison Board to the host system because the STB may convert the video into a low bitrate video owing to overhead. In typical systems, this behavior is observed. However, even in the presence of a self-adaptive SW framework in the STB, the high bitrate generated from the Intel Edison Board may not reach the IoT devices and the host system. For example, if the number of video sources increases steadily, it is not easy for a typical video transport system to provide the high bitrate videos generated

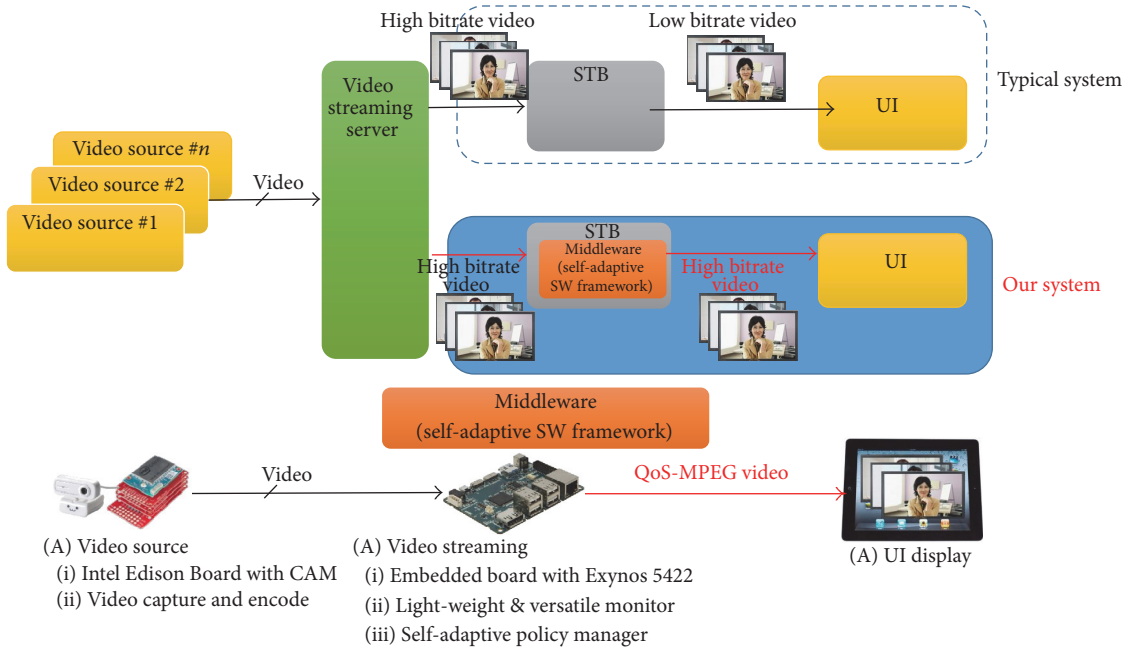


Figure 3: Video transport system with self-adaptive SW framework including light-weight and versatile monitor.

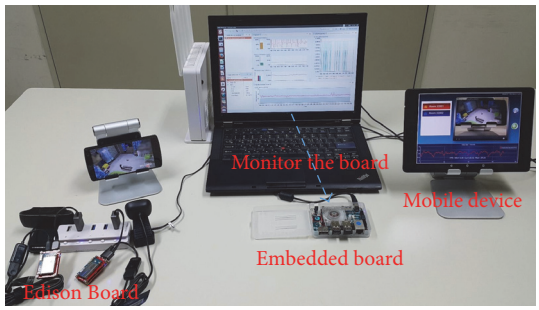


Figure 4: The screenshot of the implementation of the system in Figure 3.

in the Edison Board to the mobile device. The reason is that system resources (e.g., the number of cores, memory, and network bandwidth) allocated in the streaming server are insufficient; hence, the STB assigns low performance to running applications in a fair manner. However, the proposed video transport system can provide the high bitrate videos to the mobile device because our system has a middleware—the self-adaptive SW framework including the light-weight and versatile monitor. This middleware can handle system resources by using the application and system information collected by the monitor. Therefore, the proposed system always maintains the QoS defined by the user even in the case of many overheads.

Figure 4 demonstrates a system based on the implementation of Figure 3. In Figure 4, two Edison boards capture and encode video images. The embedded board receives these images and processes them. The light-weight and versatile monitor is ported to the board. The tool that runs in the notebook shows the information collected by the monitor in real-time; this information represents the static/dynamic information about the video streaming server and embedded board. The mobile device is provided with the original images of the target systems without loss of images. The tool and UI (user interface) show a regular QoS, indicated by the red color, in the graph.

4.2. Results and Analysis. In order to verify the light-weight characteristic of the proposed monitor, which functions on the Exynos 5422 embedded system, we measure the CPU utilization of the monitor by using our homebrew experimental application. The application automatically creates processes according to the input values and then initiates the execution of the processes. The input values are 1, 7, 15, 30, 50, 80, 100, 130, 170, and 200. CPU utilization is measured by DVFS governor (Interactive, Performance, and Ondemand). Figures 5, 6, and 7 show the CPU utilization corresponding to the number of running processes in the monitor for Interactive, Performance, and Ondemand modes of DVFS, respectively. The top graphs in each figure show the results corresponding to the conversion of CPU utilization into a percentage for 200 running processes. The CPU utilization is measured 500 times while the processes are executing. The bottom graphs in these figures show the average and error range of CPU utilization according to the number of monitored processes.

For the Interactive mode, the measured CPU utilization is shown as a percentage in the top graph of Figure 5. As shown in the dotted red rectangle of this graph, the percentage is approximately 20% for up to 50 processes. However, more than 80 processes exceed 30%, so that the suggested monitor can affect an embedded system (e.g., Exynos 5422) due to

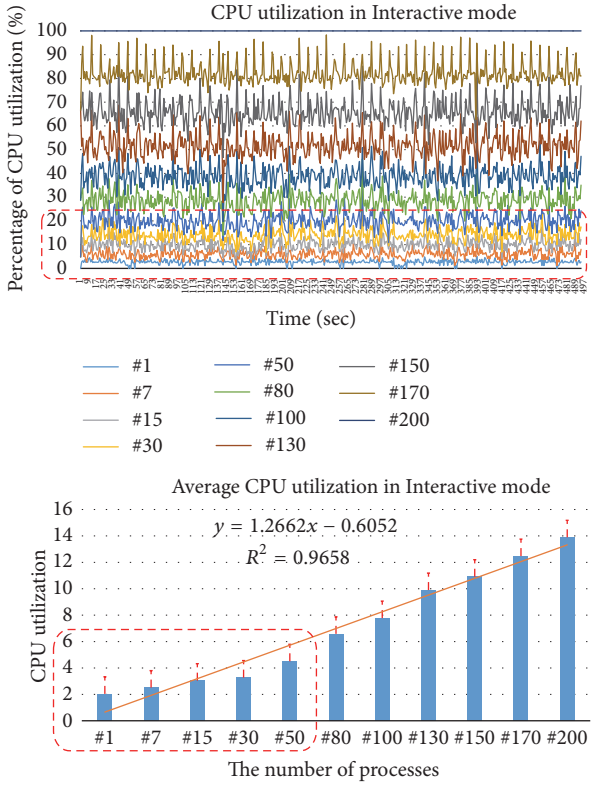


Figure 5: CPU utilization in Interactive mode.

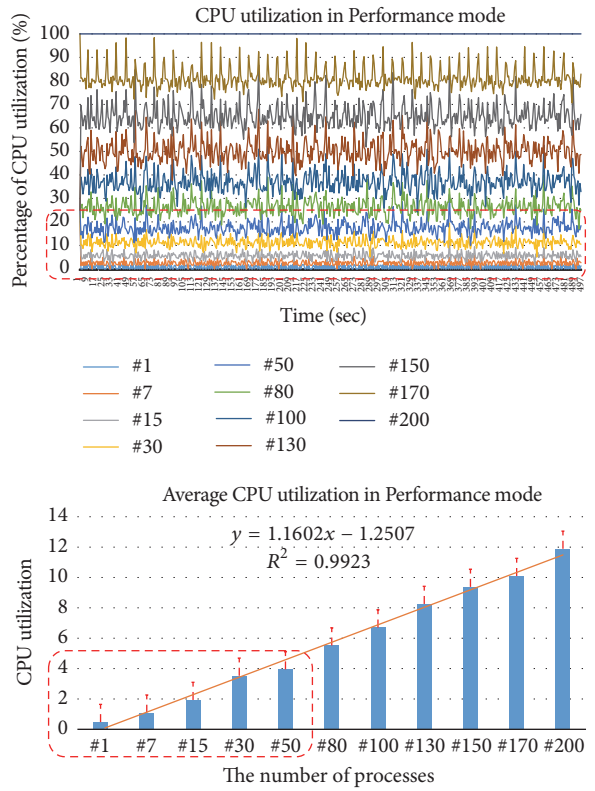


Figure 6: CPU utilization in Performance mode.

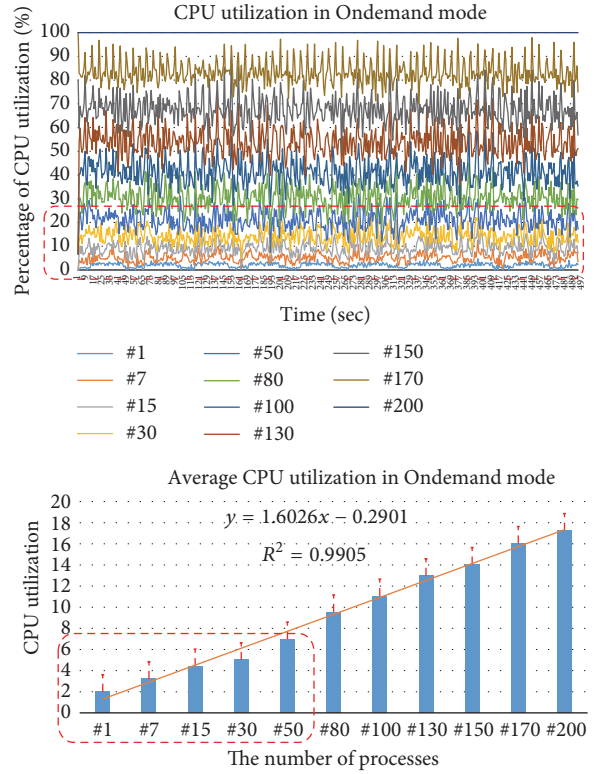


Figure 7: CPU utilization in Ondemand mode.

its monitoring overheads. The bottom graph of Figure 5 is the detailed experimental results for the aforementioned fact. The graph shows that the CPU utilization coefficient of the monitor working in the Interactive mode is 1.2662 ($y = 1.2662x - 0.6052$). The coefficient is calculated by the regression analysis result for the measured CPU utilization based on the number of processes. This result is correct because R^2, the reliability of the measured data, is 0.9658. In the bottom graph of Figure 5, the CPU utilization is approximately 4% for up to 50 processes as shown in the dotted red rectangle. Therefore, we can state that the proposed monitor has the light-weight characteristic because the monitor based on a self-adaptive SW framework is not necessary for a large number of processes (greater than 50). The results in Figures 6 and 7 are similar to the result of Figure 5. The CPU utilization coefficient is 1.1602 ($y = 1.1602x - 1.2507$) and R^2 is 0.9923. The CPU utilization is also approximately 4% for up to 50 processes. The CPU utilization coefficient in Figure 7 is 1.6026 ($y = 1.6026x - 0.2901$) and R^2 is 0.9905. However, the CPU utilization is approximately 5% for up to 50 processes. Through experimental results, the proposed monitor shows the best CPU utilization in the Performance mode and the second-best CPU utilization in the Interactive mode. The Ondemand mode has more overhead than the Performance and Interactive modes because the mode controls CPU frequency by checking CPU utilization periodically.

Figure 8 shows the CPU utilization corresponding to the number of processes (1, 7, 15, and 50) for the Interactive, Performance, and Ondemand modes. In all four graphs,

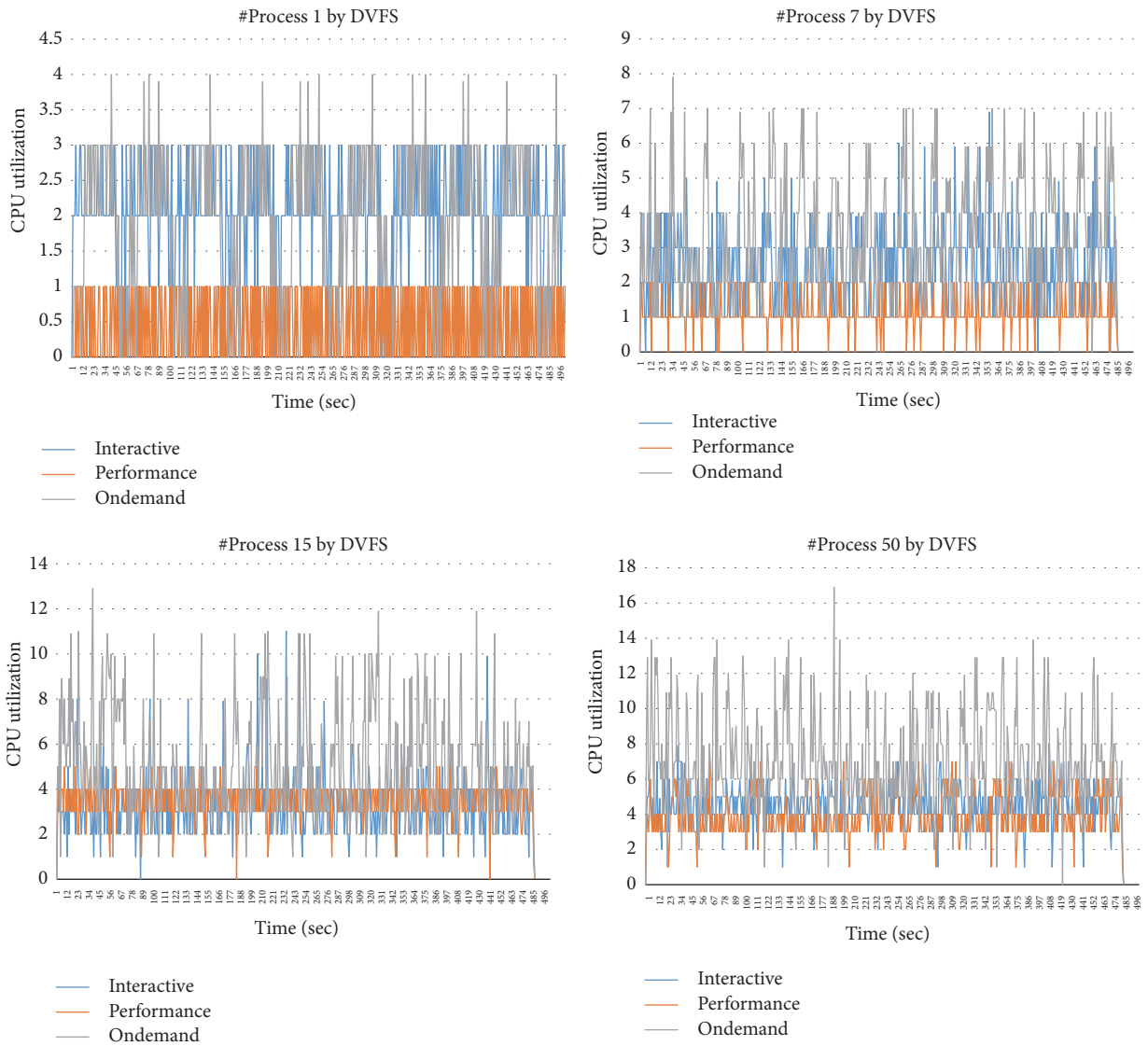


Figure 8: CPU utilization by DVFS (number of processes: 1, 7, 15, and 50).

the CPU utilization of the Ondemand mode fluctuates very widely and the utilization is also higher than that of Interactive and Performance modes. As the number of processes increases, the suggested monitor has much more overhead in the Ondemand mode. (e.g., "#Process 50 by DVFS" graph of Figure 8 indicates that the CPU utilization of the Ondemand mode is over 2 times that of the Interactive and Performance modes). That is, the monitor efficiently works in the Interactive and Performance modes. Overall, the Ondemand mode has greater overhead than the Interactive and Performance modes. Figure 9 explains the aforementioned logical reason well. In the Performance, the CPU utilization is 0.49 when the number of process is 1. In this case, the suggested monitor has the most performance. The CPU utilization is 3.95 when the number of processes is 50. The increase rate of the utilization is about 8.06. The increase gap is the largest among the CPU utilization of DVFS governor. In the Interactive mode, the CPU utilization is 2.04 when the number of process is 1 and the CPU utilization is 4.50 when the number of processes is 50. The increase rate of the utilization is about 2.21. The increase gap is the smallest, so that the suggested monitor has the most stable performance. In the Ondemand mode, the CPU utilization is 1.99 when the number of process is 1, and the CPU utilization is 7.03 when the number of processes is 50. In this case, the suggested monitor has the worst performance. The increase rate of the utilization is about 3.53. Therefore, the monitor has the light-weight overhead in the Performance mode and the monitor has the most stable overhead in the Interactive mode.

5. Conclusion

In this manuscript, we propose a light-weight and versatile monitor that can be used in a self-adaptive SW framework. This monitor can be used for large-scale devices (gateway and STB) and small-scale devices (Intel Edison Board and

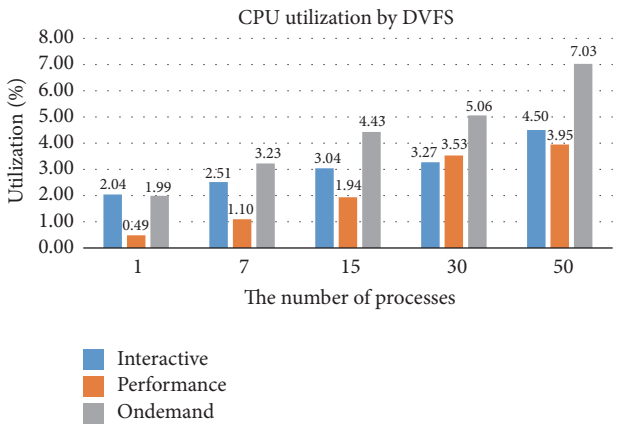


Figure 9: Average CPU utilization by number of processes.

IoT devices). The proposed monitor provides static/dynamic information about system resources and running applications to users in real-time; thus, the monitor helps a self-adaptive policy manager of the self-adaptive SW framework to optimally control system resources or applications. From our experiments, we determined that the monitor shows maximum optimization in the Performance mode of DVFS. The monitor shows 3.95% CPU utilization for up to 50 monitored processes. In future work, we will apply the proposed monitor to various hardware platforms and will demonstrate the superiority of the proposed monitor.

Disclosure

This paper is a revised and expanded version of the paper entitled "Design of Self-Adaptive System Observation over Internet of Things" presented at International Conference on Control and Automation (CA 2015), November 25, 2015, Jeju, Korea.

Competing Interests

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

Acknowledgments

This work was supported by ICT R&D program of MSIP/IITP (B0101-16-0661, the Research and Development of the Self-adaptive Software Framework for various IoT Devices) and also was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2060082).

References

[1] C. Bolchini, M. Carminati, A. Miele, and E. Quintarelli, "A framework to model self-adaptive computing systems," in *Proceedings of the NASA/ESA Conference on Adaptive Hardware and Systems (AHS '13)*, pp. 71–78, Torino, Italy, June 2013.

[2] H. Hoffmann, M. Maggio, M. D. Santambrogio, A. Leva, and A. Agarwal, "A generalized software framework for accurate and efficient management of performance goals," in *Proceedings of the 13th International Conference on Embedded Software (EMSOFT '13)*, pp. 1–10, IEEE, October 2013.

[3] H. Hoffmann, J. Eastep, and M. D. Santambrogio, "Application heartbeats: a generic interface for specifying program performance and goals in autonomous computing environments," in *Proceedings of the International Conference on Autonomic Computing (ICAC '10)*, pp. 79–88, June 2010.

[4] H. Hong, J. Lim, H. Lim, and S. Kang, "New thermal-aware voltage Island formation for 3D many-core processors," *ETRI Journal*, vol. 37, no. 1, pp. 118–127, 2015.

[5] https://www.sparkfun.com/categories/272.

[6] http://www.intel.com/content/www/us/en/do-it-yourself/edison.html.

[7] Intel Edison Boards, Intel Edison Breakout Board Hardware Guide, 2015 .

[8] Intel Edison Boards, Intel Edison Board Support Package—User Guide, Revision 001, 2014.

[9] http://www.samsung.com/semiconductor/minisite/Exynos/w/solution/mobile_ap/5422/.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 9784071, 12 pages
http://dx.doi.org/10.1155/2016/9784071

Research Article

Sensor-Based Model Driven Control Strategy for Precision Irrigation

Camilo Lozoya, Carlos Mendoza, Alberto Aguilar, Armando Román, and Rodolfo Castelló

Tecnologico de Monterrey, Ave. Heroico Colegio Militar 4700, 31300 Chihuahua, CHIH, Mexico

Correspondence should be addressed to Camilo Lozoya; camilo.lozoya@itesm.mx

Received 17 June 2016; Accepted 18 October 2016

Academic Editor: Rafael Morales

Improving the efficiency of the agricultural irrigation systems substantially contributes to sustainable water management. This improvement can be achieved through an automated irrigation system that includes a real-time control strategy based on the water, soil, and crop relationship. This paper presents a model driven control strategy applied to an irrigation system, in order to make an efficient use of water for large crop fields, that is, applying the correct amount of water in the correct place at the right moment. The proposed model uses a predictive algorithm that senses soil moisture and weather variables, to determine optimal amount of water required by the crop. This proposed approach is evaluated against a traditional irrigation system based on the empirical definition of time periods and against a basic soil moisture control system. Results indicate that the use of a model predictive control in an irrigation system achieves a higher efficiency and significantly reduce the water consumption.

1. Introduction

Agriculture represents the major water consumer globally; this sector uses approximately the 70% of the available fresh water resources worldwide, mostly in crop irrigation activities. The world average efficiency of agricultural irrigation is about the 50%–60%, mainly due to the inappropriate management of this natural resource [1]. A deficient water management causes not only the waste of this vital liquid, but also a significant reduction on crop productivity.

Precision irrigation refers to the management of the irrigation scheduling according to the crop requirements. The amount of water applied to the crop is based on measurements of soil, crop, and weather variables which reflects the status of the plant [2]. Among the main goals of precision irrigation are the increment of water efficiency, the reduction of energy consumption, and the maximization of crop productivity, by using technology such as wireless sensor networks, mobile devices, remote sensing, real-time control, and information systems.

Nowadays, most of the commercial automated irrigation systems offered by the market (Acclima, Rainbird, Watermark, Decagon) are programmed to irrigate at time intervals for predefined periods of time. The irrigation schedule is defined offline, and it is usually based on the user empirical knowledge on crop needs, soil characteristic, and weather factors [3]. Some farmers use crop evapotranspiration (et_c) data to determine the irrigation schedule. Evapotranspiration represents the water lost caused by soil surface evaporation and crop transpiration; therefore the amount of water applied is used to refill the water consumed by the plant and the environment. Most recently, there have appeared in the market new commercial automatic controllers that regulate the use of water, based on soil moisture measurement by implementing a closed-loop irrigation control. These controllers activate irrigation when sensors detect that soil moisture is under a predefined low limit threshold and deactivate irrigation if soil moisture is above a high limit [4, 5]. On-off control may also be implemented based on direct plant canopy measurements, such as the crop water stress index that can be obtained by measuring the air and canopy temperatures, as well as the atmospheric vapor pressure [6]. However, due to practical difficulties on obtaining real-time measurements from the canopy, it is difficult to find any commercial irrigation controllers based on direct plant measurements.

In recent years, agricultural engineers and control community have increased their attention to the analysis and implementation of real-time closed-loop irrigation control

systems, since the use of control techniques for precision irrigation has demonstrated the obtention of large amounts of water savings. Two major research approaches can be observed in this area: modelling and implementation.

For the first approach, the main focus is the analysis of the dynamics of soil, crop, and weather conditions in order to properly model the irrigation process. Reference [7] implements a wireless sensor network to automate a viticulture irrigation system; the work focuses on modelling the process dynamics using the water balance model proposed by [8]. In [2] a soil-plant-atmosphere model was developed to simulate water transport in a crop field and to design and test model-based irrigation control strategies such as PID (Proportional-Integral-Derivative) control and MPC (model predictive control). Reference [9] proposes a predictive control algorithm to schedule irrigation events and uses measured weather data to evaluate the simulation model. Reference [10] investigates how soil moisture sensors positioning and accuracy may affect the performance of soil moisture based surface drip irrigation scheduling under different conditions. However, in general those approaches do not deal with the implementation details when the systems are implemented for large crop fields.

The second research approach mainly deals with the integration of technology such as wireless sensor networks, real-time controllers, and information systems to implement automatic irrigation systems. In [11] an automated irrigation system, based on a wireless network of soil moisture and soil temperature sensors, was developed to optimize water use for agricultural crops; the system was evaluated in an organic sage greenhouse where either the moisture or the temperature activates the on-off irrigation control. In [12] different experiments on commercial plantations were conducted in order to manage the irrigation based on soil moisture measurements validating the feasibility to implement on-off control on woody and vegetable crops. In [13] a wireless sensor network was developed to acquire field soil property data (soil moisture, electrical conductivity, and soil temperature), compared to predefined thresholds, and as a result activate or deactivate the drip irrigation. In [14] an automated closed-loop irrigation control system was developed where irrigation decisions were site-specifically made based on feedback from soil water conditions, by controlling the on-off times for groups of sprinkler nozzles. Although these approaches consider detailed implementation issues, they do not include a model for the process dynamics and as a result controller design is simple and empirical.

The work presented in this paper describes the implementation of an automatic system for precision irrigation considering a model driven approach, where the process dynamics are experimentally identified and validated in order to design a predictive algorithm implemented over an embedded system platform to achieve optimization in the use of water for agricultural activities. This work extends the preliminary results obtained on [15] where a model predictive control strategy for closed-loop irrigation was simulated. In this paper, the mathematical model is refined and validated experimentally, the predictive algorithm is implemented over an embedded platform in order to automate a drip irrigation field, and the obtained results on the evaluated crop are compared against the typical commercially available approaches.

Model predictive control (MPC) is an optimal control strategy based on numerical optimization over a finite horizon, as denoted by [16]. MPC requires a heavy computational load to achieve optimization, where future control inputs and future process responses are predicted using a mathematical model and optimized according to a cost function. Model predictive control has been proposed as a suitable technique for large water distribution systems. In [17], MPC is used to generate flow control strategies from the sources of water to the consumer and irrigation areas to achieve safety volumes in dams and flow control stability. In [18], a hierarchical system to control an irrigation canal is proposed, where a centralized predictive controller controls the inflow to the canal and coordinates the local controllers by modifying their set-points. In [19], MPC is used to maintain the water level of navigation canals while reducing water levels deviations. In the proposed approach a predictive algorithm minimizes the control signal (effective irrigation) while keeping soil moisture under specific thresholds (avoiding water stress) and considers external disturbances (evapotranspiration) to predict the process dynamics.

The rest of this paper is structured as follows. Section 2 introduces the hydrological balance, the evapotranspiration, and the soil moisture concepts. Then, in Section 3 the formal definition of the process model is presented and the controller design strategy is defined. Section 4 presents the model parameter estimation based on direct measurements from soil moisture and weather conditions, and it also presents the results where the predictive model is evaluated against traditional irrigation systems based on the definition of time periods and against a basic soil moisture control system. Finally, Section 5 concludes the paper.

2. Preliminaries

2.1. Hydrological Balance. The process dynamics of an agricultural irrigation system can be described by using the hydrological balance model [8]. This model establishes that a change in water storage during a time period in a specific location is the result of water inflows (irrigation, rainfall, and capillary rise) minus the water outflows (evaporation, plant transpiration, water runoff, and deep percolation), as depicted in Figure 1.

Using soil moisture $\theta(t)$ in order to measure field water storage, then, the hydrological balance dynamics can be defined as

$$\dot{\theta}(t) = \mathrm{ir}(t) + \mathrm{rf}(t) + \mathrm{cr}(t) - \mathrm{et}_c(t) - \mathrm{dp}(t) - \mathrm{ro}(t), \quad (1)$$

where soil moisture variations in the root zone $\dot{\theta}(t)$ depends on effective irrigation $\mathrm{ir}(t)$, rainfall $\mathrm{rf}(t)$, capillary rise $\mathrm{cr}(t)$, crop evapotranspiration $\mathrm{et}_c(t)$, deep percolation $\mathrm{dp}(t)$, and water outflow due to runoff $\mathrm{ro}(t)$.

If a dry and plain land area for irrigation is considered, then $\mathrm{rf}(t)$ (assuming no rainfall), $\mathrm{cr}(t)$ (assuming no deep water available for capillary rise), and $\mathrm{ro}(t)$ (assuming no

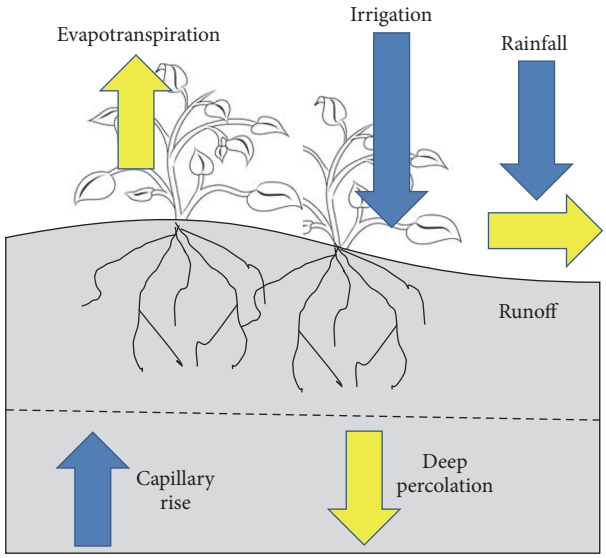


Figure 1: Components of hydrological balance for an irrigation system (figure obtained from Allen et al. [8]).

runoff due to plain land) terms can be removed from water balance, and simplified dynamics can be expressed as

$$\dot{\theta}(t) = \mathrm{ir}(t) - \mathrm{et}_c(t) - \mathrm{dp}(t), \quad (2)$$

where soil moisture variations $\dot{\theta}(t)$ depend just on the effective irrigation, crop evapotranspiration, and deep percolation.

2.2. Evapotranspiration. Crop evapotranspiration represents the water lost caused by soil surface evaporation and crop transpiration. The evapotranspiration rate is normally expressed in millimeters (mm) per unit of time (usually days). The rate expresses the amount of water lost from the cropped surface in units of water depth. An evapotranspiration of 1 mm/day is equivalent to a loss of 10,000 liters per hectare per day.

Crop evapotranspiration et_c depends on both weather factors and crop characteristics (crop type, development stage). According to [8], the water demand of any crop can be computed by multiplying the weather factors of the evapotranspiration with a coefficient that depends on the crop specific characteristics, as denoted by

$$\mathrm{et}_c(t) = K_c \mathrm{et}_o(t), \quad (3)$$

where K_c is the constant crop coefficient which depends on the crop type and its development stage; this constant is globally known and it is independent from the environmental conditions. Reference evapotranspiration $\mathrm{et}_o(t)$ depends only on weather parameters, and it can be obtained by using the FAO Penman-Monteith method [20], which requires measurements from solar radiation, wind speed, air temperature, and relative air humidity variables.

2.3. Soil Moisture. Soil moisture plays an important role in agriculture. Soil moisture refers to the amount of water in soil,

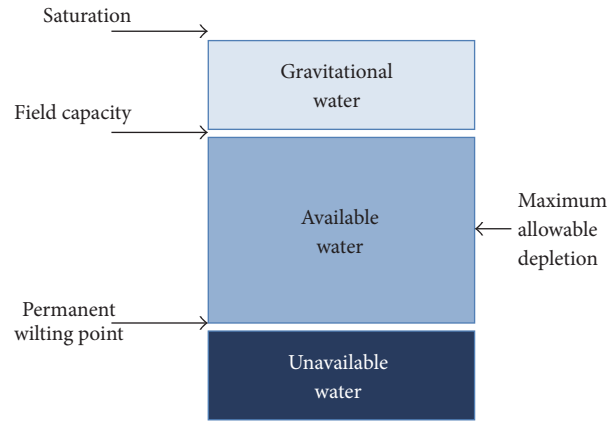


Figure 2: Volumetric water content parameters.

which is described as the volumetric water content (VWC). Volumetric water content (θ) indicates the percentage of water volume for a specific volume:

$$\theta = \frac{V_W}{V_T}, \quad (4)$$

where V_W is the water content in volume units for a specific sample and V_T is the total volume sample (soil + water + air).

In any crop, the soil moisture needs to be maintained above permanent wilting point and stay below field capacity. Permanent wilting point is the soil moisture level at which plants cannot longer absorb water from the soil. Field capacity is the quantity of water stored in a soil volume after drainage of gravitational water. The available water capacity of soil is the water that is available to the crop, and it represents the range of soil moisture values that lie above permanent wilting point and below the field capacity, as shown in Figure 2. The point below field capacity where crops become stressed is known as the maximum allowable depletion (MAD); below this level the crop is able to receive water from soil; however after a period of time it will become stressed. This value is expressed as a percent of the available water capacity and typically represents 50% for most of the crops.

Field capacity and permanent wilting point are heavily influenced by soil textural classes, [21]; for example, a silt loam type of soil (frequently used for agricultural purposes) has a typical range of values from 0.3 to 0.4 volumetric water content for the available water capacity.

3. Materials and Methods

3.1. Plant Dynamics Model. Based on the hydrological balance (2), the process dynamics for an irrigation system can be described as a block diagram with two inputs (effective irrigation and reference evapotranspiration) and one output (soil moisture), as shown in Figure 3.

Notice that reference evapotranspiration (et_o) is used instead of crop evapotranspiration (et_c), because et_o depends only on external weather parameters. On the other hand, et_c is the result of et_o that multiplies the crop coefficient (K_c) according to (3), so it is assumed that K_c is a constant that

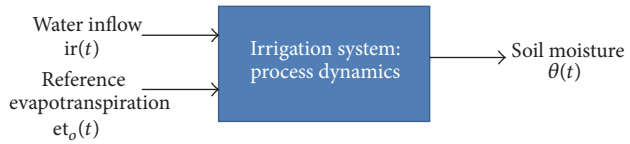


Figure 3: Irrigation system input/output definition.

belongs to the internal process dynamics of the irrigation system. Also notice that deep percolation $\mathrm{dp}(t)$ is not present in the block, since it is assumed that water percolation in an irrigation system is clearly proportional to soil moisture [7], then (2) can be rewritten as

$$\dot{\theta}(t) = \mathrm{ir}(t-\tau) - K_c \mathrm{et}_o(t) - c_0\theta(t), \tag{5}$$

where c_0 is a constant value denoting the proportional relation between soil moisture and deep percolation and τ represents the time-delay from the start of irrigation until the sensor detects a change in the soil moisture.

Since a discrete model is required, then by using the Euler approximation on soil moisture variations

$$\dot{\theta}(t) = \frac{\theta(kh+h) - \theta(kh)}{h}, \tag{6}$$

where h is the sampling interval. Using (6) in (5), the discrete time dynamics is given by

$$\theta(kh+h) = c_1\theta(kh-\tau) + c_2\mathrm{ir}(kh) - c_3\mathrm{et}_o(k); \tag{7}$$

without the loss of generality c_1, c_2, and c_3 can be used as the three discrete coefficients that absorb the previous coefficients h, K_c, and c_0. Also it is considered that τ is considerably larger than h.

Now (7) can be reformulated by using a first-order state-space representation as

$$\theta(kh+h) = [c_1][\theta(kh)] + \begin{bmatrix} c_2 & c_3 \end{bmatrix} \begin{bmatrix} \mathrm{ir}(kh-\tau) \\ -\mathrm{et}_o(kh) \end{bmatrix}, \tag{8}$$

where c_1, c_2, and c_3 are coefficients that define the dynamics of the process and can be obtained from direct measurements of the evapotranspiration, the soil moisture, and the effective irrigation.

3.2. Coefficients Estimation. Although soil moisture dynamics can be defined by a well-known stochastic differential equation (2), the estimation of soil moisture variations for large areas is highly complex, due to the large spatial variation in measurements along with the presence of processes (irrigation, percolation, evapotranspiration, etc.) that vary in space and time [22]. As a result soil moisture has a highly nonlinear behavior, since measurements of soil moisture can vary at spatial scale as small as meters. In addition, drainage rates depend on topographic variations, water movements depends on heterogeneity at scale that hardly can be quantified, and even evapotranspiration varies spatially and timely due to soil and vegetative variations. Therefore coefficients c_1, c_2, and c_3 from (8) are time-variant and difficult to estimate.

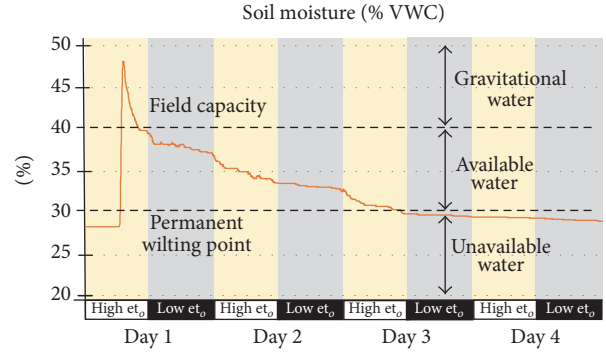


Figure 4: Soil moisture general dynamics.

However the soil moisture general dynamics is simple and intuitive; see Figure 4. Irrigation adds soil moisture up to a saturation level. Then excess water is rapidly drained until field capacity is reached. Below field capacity, moisture is withdrawn at a slower rate, depending on the crop evapotranspiration. Therefore, the soil moisture slope that represents the water depletion rate is higher during the day (high et_c) and lower during the night (low et_c). When soil moisture arrives to the wilting point, plants cannot longer extract water from soil. Below wilting point the rate of depletion is even slower and mostly depends on soil characteristics.

Therefore, in order to simplify the dynamic model, make an adequate parameter estimation, and reduce the complexity caused by the spatial and temporal variations in measurements, two considerations were carried out to conduct the model identification process:

(i) Soil moisture measurements may involve a large group of nodes distributed along a large piece of land; therefore in order to obtain a single representative soil moisture value for the irrigation area, a data aggregation method is required to summarize the information provided by a group of soil moisture sensors. According to [23] measuring accuracy is exponentially improved when increasing the number of soil moisture sensors.

(ii) Since crop-water-soil dynamics are different depending on the soil volumetric water content level specified in Figure 4, then parameter estimation for c_1, c_2, and c_3 is obtained for each level; that is, one set of parameters correspond to the level below the permanent wilting point, another to the available water level (between field capacity and permanent wilting point), and another one to the gravitational water level. Within each level parameters are considered time-invariant.

For system identification purposes (8) is rewritten as

$$\theta(kh+h) = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} \theta(kh) \\ \mathrm{ir}(kh-\tau) \\ -\mathrm{et}_o(kh) \end{bmatrix}; \tag{9}$$

then (9) can be simplified as

$$\theta(kh+h) = \gamma^T \phi(kh), \tag{10}$$

where

$$\phi(kh) = \begin{bmatrix} \theta(kh) \\ \text{ir}(kh - \tau) \\ -\text{et}_o(kh) \end{bmatrix} \tag{11}$$

is the regressor vector that can be obtained from direct measurements, $\theta(kh + h)$ is the known output, and $\gamma^T = [c_1 \ \ c_2 \ \ c_3]$ is the parameter vector to be estimated.

Suppose there is an estimate of the parameter vector $\widehat{\gamma}$; then at time kh the estimation can be obtained by

$$\widehat{\theta}(kh + h) = \widehat{\gamma}^T(kh)\,\phi(kh), \tag{12}$$

where the least square estimate of γ minimizes the cost function defined by

$$J_k = \sum_{i=1}^{k} \left[\theta(ih) - \widehat{\gamma}^T(kh)\,\phi(ih - h)\right]^2, \tag{13}$$

expanding the quadratic term

$$J_k = \sum_{i=1}^{k} \begin{bmatrix} y(ih)^2 + \widehat{\gamma}^T(kh)\,\phi(ih - h)\,\phi^T(ih - h)\,\widehat{\gamma}(kh) \\ -2y(ih)\,\phi(ih - h)\,\widehat{\theta}(kh) \end{bmatrix}; \tag{14}$$

since on least squares $\partial J_k/\partial\widehat{\gamma}(kh) = 0$, then applying partial derivative on (14) the parameter vector can be estimated by

$$\widehat{\gamma}(kh) = F(k) \sum_{i=1}^{k} \phi(ih - h)\,y(i), \tag{15}$$

where

$$F(k) = \left[\sum_{i=1}^{k} (ih - h)\,\phi^T(ih - h)\right]^{-1}. \tag{16}$$

Now, recursive least square is used to ease the implementation into a real-time algorithm. Recursive least square is an online implementation of least squares where the estimated parameter is predicted and corrected by using the current measurement [24], in the form of

$$\widehat{\gamma}(kh + h) = \widehat{\gamma}(kh) + F(kh + h)\,\phi(kh)\,\epsilon(kh + h), \tag{17}$$

where $\epsilon(kh + h) = y(kh + h) - \widehat{\gamma}^T(kh)\phi(kh)$ is the a priori estimate error and $F(kh + h)$ is the adaptation gain that can be updated by

$$F(kh + h) = F(kh) - \frac{F(kh)\,\phi(kh)\,\phi^T(kh)\,F(kh)}{1 + \phi^T(kh)\,F(kh)\,\phi(kh)}. \tag{18}$$

By using (17) and (18), an offline algorithm can be implemented in order to estimate coefficients c_1, c_2, and c_3, as shown in Algorithm 1.

This recursive algorithm is executed after a number of samples are directly obtained from the process; therefore $\theta(kh + h)$ and $\phi(kh)$ are known and available. The current estimate is equal to the previous estimate plus a correction term. The correction term is proportional to the deviation of the predicted value from what is actually observed. The adaptation gain is updated on each iteration to achieve fast convergence. At the end, coefficients are taken from the last estimation of the loop.

3.3. Model Predictive Control. To implement the controller for the closed-loop irrigation system, model predictive control (MPC) is used in order to minimize the control signal (effective irrigation) while keeping soil moisture under specific thresholds (avoiding water stress) and by considering external disturbances (reference evapotranspiration) (Algorithm 2). Figure 5 shows a feedback loop where the control objective is to keep within certain thresholds the soil water content in order to have a healthy and productive crop. Thus the process variable $y(kh)$ is the soil moisture, $r(kh)$ is the reference value (soil moisture set-point), and the error value $e(kh)$ is obtained as a result of the difference between the process value and the reference value.

The environmental factors affecting the irrigation systems are modelled as an external disturbance, so the reference evapotranspiration $\text{et}_o(kh)$ represents the disturbance signal affecting the process. By knowing the disturbance model, then the system may predict the disturbance effects and react before these effects affect the process output.

Given the dynamic model of a closed-loop irrigation system defined by (8), the controller knows the process dynamics due to the online estimated internal model obtained from (17) by using recursive least squares. Within the controller, a numerical optimization algorithm is executed based on the current error and the disturbance measurements; this information is applied to the internal model and an optimal solution is found over a finite horizon T_{FH}, which minimizes the following quadratic cost function based on the error and the control signal,

$$J(kh) = \sum_{i=0}^{T_{\text{FH}}-1} \begin{bmatrix} \left(e^T kh + ih \mid kh\right) Qe(kh + ih \mid kh) \\ +u^T(kh + ih \mid kh)\,Ru(kh + ih \mid kh) \end{bmatrix}, \tag{19}$$

where matrix Q is positive semidefinite and matrix R is positive definite and represents the weight given to the error and the control action, respectively, within the cost function. Also, $e(kh + ih \mid kh)$ and $u(kh + ih \mid kh)$ denote the predicted error and control effort, respectively, at time $kh+ih$ performed at kh.

The optimal input sequence for the problem of minimizing $J(kh)$ is denoted by

$$u^*(kh) = \arg\min_u J(kh), \tag{20}$$

subject to

$$\begin{aligned} u(kh) &= \{0, I_{\max}\}, \\ \theta(kh) &\geq \theta_{\min}, \\ \theta(kh) &\leq \theta_{\max}, \end{aligned} \tag{21}$$

$\widehat{\gamma}(0) = [c_1(0), c_2(0), c_3(0)]$
$F(0) = \sigma$
for $k = 0; k \leq$ number_of_samples; $k++$ **do**
  $y(kh + h) = \theta(kh + h)$
  $\epsilon(kh + h) =$ obtain_error $(y(kh + h), \widehat{\gamma}(kh), \phi(kh))$
  $F(kh + h) =$ calculate_adaptation_gain $(F(kh), \phi(kh))$
  $\widehat{\gamma}(kh + h) =$ estimate $(\widehat{\gamma}(kh), F(kh + h), \phi(kh), \epsilon(kh + h))$
end
$[c_1, c_2, c_3] = \widehat{\gamma}(kh)$

Algorithm 1: Coefficients estimation.

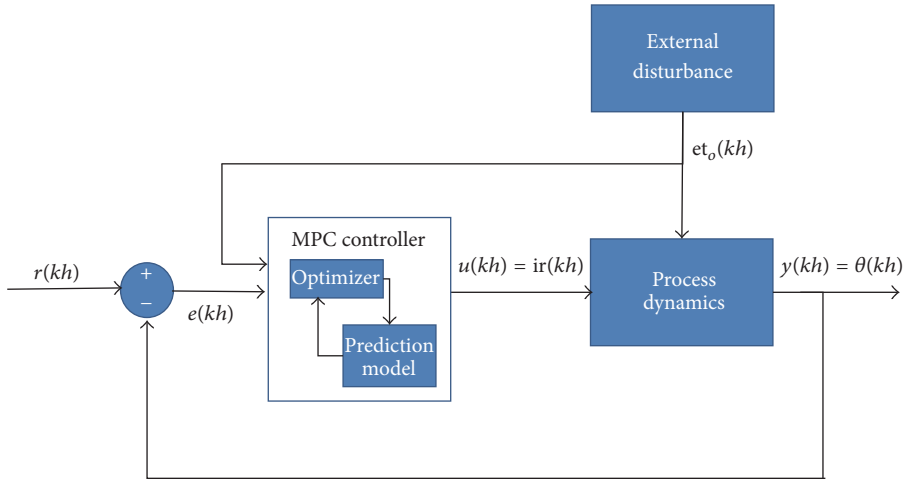


Figure 5: Model predictive control loop.

where irrigation at a specified time $u(kh)$ can only have two values; irrigation is either off (=0) or on ($=I_{max}$). Soil moisture value $\theta(kh)$ is constrained by the measurement range provided by the sensors.

The algorithm determines "when" to start irrigating and "how much" water the crop requires. Therefore, the optimal input sequence $u^*(kh)$ is obtained by adjusting the soil moisture low threshold and the irrigation period to optimal values.

The algorithm starts at kh by reading current values for soil moisture, irrigation, and reference evapotranspiration; these are considered as initial conditions and represented by $\theta(0)$, ir(0), and $et_o(0)$. Then, for each combination of predefined values for the soil moisture low threshold vector θ_{LT} and the irrigation period vector ir_P, the algorithm estimates the sequence of next state predictions over the finite horizon T_{FH}, while updating inputs, applying the control rules, and calculating the total cost. Finally, the algorithm finds the optimal combination of θ_x and ir_y that provides the best control rules to determine the optimal input sequence.

4. Experiments and Results

4.1. Experimental Platform Set-Up. The experimental platform consists of a data acquisition and control system described in [25], where a modular and scalable design approach is considered in order to provide different levels of access with different data contents. At the lower level, raw data from sensors is available, and at higher levels more refined and consolidated information can be obtained from the system. The experimental platform is divided in three access levels (field, data, and user) and it is capable of controlling four irrigation areas, as depicted in Figure 6.

At the field level, there are four irrigation areas that can be controlled, where each area includes two sensor nodes and one actuator node. Notice that only one weather node can be used for the four irrigation areas, since environmental variables have practically the same values for the complete field area. The actuator node controls an irrigation valve and measures the water flow; meanwhile each sensor node contains three soil moisture sensors to measure the volumetric water content at the crop root level. The sensor, weather and actuator nodes are implemented with low cost boards Arduino Mega based in the microcontroller ATmega328 (https://www.arduino.cc/). Soil measurements are conducted by using a Decagon Devices (http://www.decagon.com/) 10HS volumetric water content sensor, as shown in Figure 7. The sensors are located at the crop root level, with a measurement range from 0% to 60% of volumetric water content and a resolution of 0.1% when calibrated. In the actuator node (see Figure 7) a Rain Bird irrigation valve (Rain Bird Corporation, http://www.rainbird.com/) is used to activate

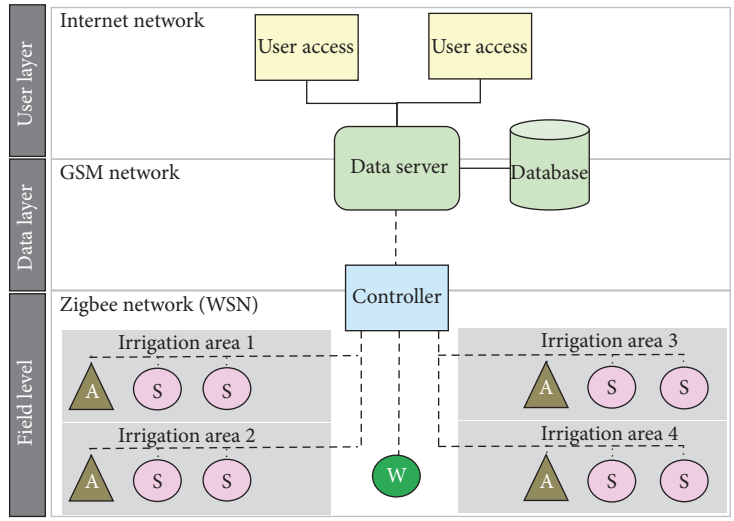


Figure 6: Data acquisition and control system architecture.

$\theta_{LT} = [\theta_1, \theta_2, \theta_3, \ldots, \theta_n]$
$ir_P = [ir_1, ir_2, ir_3, \ldots, ir_m]$
$[\theta(0), ir(0), et_o(0)] = read\_current\_values\ ()$
for $x = 1;\ x \leq n;\ x++$ **do**
  for $y = 1;\ y \leq m;\ y++$ **do**
    $J_{sum} = 0$
    for $i = 0;\ i < T_{FH};\ i++$ **do**
      $\theta(ih + h) = next\_state(\theta(ih), ir(ih), et_o(ih))$
      $et_o(ih + h) = next\_et_o(et_o(ih))$
      $ir(ih + h) = ir(ih)$
      if $\theta(ih) \leq \theta_x$ **then**
        $ir(ih + h) = I_{max}$
        $irrigation\_time = 0$
      end
      if $irrigation\_time > ir_y$ **then**
      $ir(ih + h) = 0$
      end
      $irrigation\_time = irrigation\_time + h$
      $u(ih) = ir(ih)$
      $J_{current} = current\_cost(e(ih), u(ih))$
      $J_{sum} = J_{sum} + J_{current}$
  end
  if $J_{sum} < J_{min}$ **then**
    $J_{min} = J_{sum}$
    $optimal\_\theta_{LT} = \theta_x$
    $optimal\_ir_P = ir_y$
  end
 end
end

Algorithm 2: Model predictive control.

or deactivate the field irrigation, and Hunter Flow Sync (http://www.hunterindustries.com/) sensor is used to measure the water flow. The weather node is used to calculate the reference evapotranspiration et_o from the environmental variables according to Penman-Monteith, a Decagon PYR Sensor measures the solar radiation, from a Decagon Davis Cup anemometer the wind speed is measured, and a Decagon VP-4 sensor is used to obtain the air temperature, the air relative humidity, and the barometric pressure, as depicted in Figure 8.

At the data level, a wireless sensor network (WSN) is implemented where the control node produces aggregated information from different sensor raw data. The WSN is implemented over the IEEE 802.15.4 standard which is the basis for the Zigbee communication protocol (http://www.zigbee.org/). Zigbee has become the de-facto standard for wireless sensor networks due to low cost, low power consumption, and small communication packet size. The wireless communication element is implemented by a radio-frequency Digi International Xbee (https://www.digi.com/) transceiver that operates at 2.4 Ghz with a data rate of 9600 bps and an open-field coverage of 1.6 kms. The controller node main element is a high performance microcontroller dsPIC33F within the Microchip Explorer 16 board (http://www.microchip.com/); see Figure 8. Control tasks are executed on the Erika real-time kernel (Erika Enterprise, http://erika.tuxfamily.org/); the real-time kernel provides to the microcontroller the capability to schedule several periodical tasks. The module has a dual network access, since it communicates with the wireless sensor network and also includes a long range communication access with the data server through the GPRS standard. The GPRS element consists of a SIMComm SIM900 (http://www.sim.com/) integrated circuit which implements the modem functionality.

At the user level, a data server module is implemented by a multicore Dell PowerEdge server (http://www.dell.com/), which includes web services, internet access, and a database in order to store historical information from the central

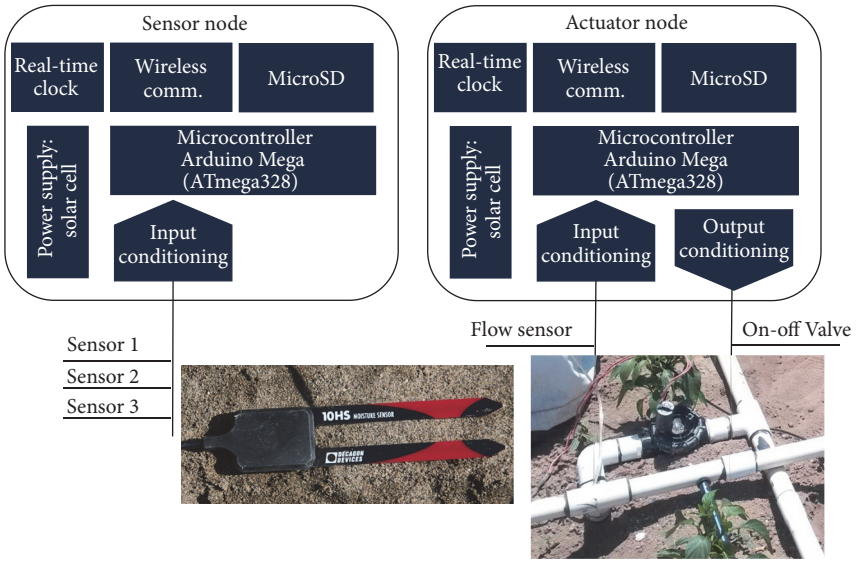


Figure 7: Sensor and actuator nodes hardware implementation.

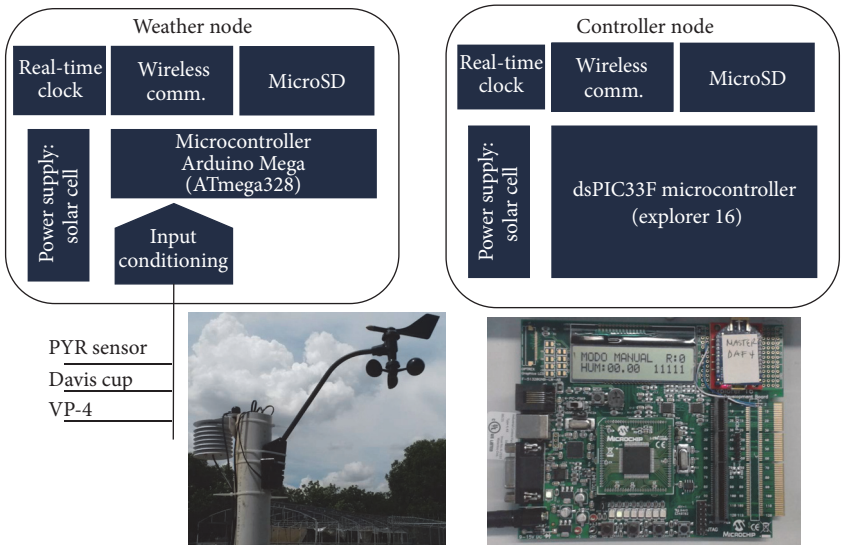


Figure 8: Weather and controller nodes hardware implementation.

modules that can be located in remote areas. The database is implemented through the open-source platform Django (https://www.djangoproject.com/). The user can access the data through web pages.

4.2. Plant Dynamics Validation. Process dynamic data (soil moisture, reference evapotranspiration, and irrigation) was captured by the data acquisition system for 21 days. The first experimental field corresponds to one irrigation area of approximate 20×10 meters; six sensor nodes were distributed along the field. Drip irrigation was used in order to water the area covered by grass; the irrigation process was manually activated at different days with different durations in order to have values from below permanent wilting point until saturation. The sensors were located in a depth of 20 cm in order to measure the soil moisture at the grass root level. For process identification grass was used instead of a specific crop, since grass $K_c = 1$; that is, $et_c = et_o$.

After data was captured, an offline recursive least square algorithm defined by (17) and (18) was executed on Matlab (http://www.mathworks.com/) in order to obtain the process dynamics coefficients c_1, c_2, and c_3, for each one of the three volumetric water levels:

(i) Gravitational water (above field capacity): the process dynamics depends on c_1 with a high value, while c_3 has no effect; c_2 has a great impact during irrigation instants.

(ii) Available water (below field capacity, above permanent wilting point): the process dynamics depends

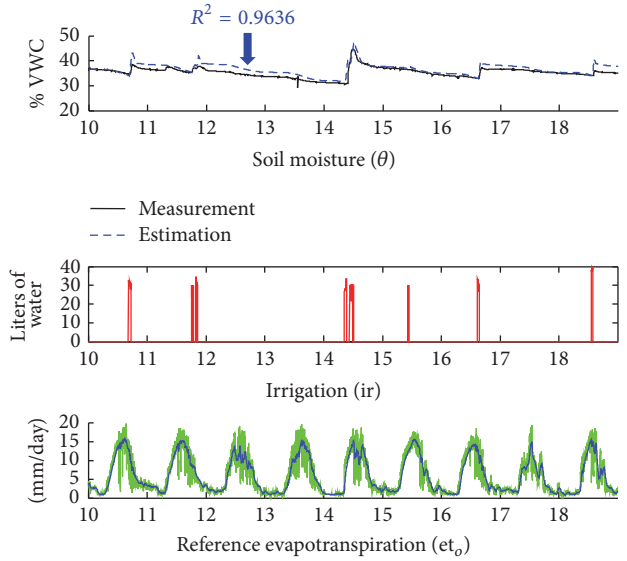


Figure 9: Model validation results.

on mostly c_3, but also c_1 has an effect; c_2 has similar impact as in the gravitational level.

(iii) Unavailable water (below permanent wilting point): the process dynamics depends on c_1 with a low value and c_3 has no effect; c_2 has a lower impact in comparison with the other levels.

Once the model coefficients were obtained, the estimated model was validated with direct measurements. In the same experimental field, a validation run was conducted with random irrigations for 10 days during each one of the months of April, May, June, July, and August. Figure 9 shows the result for the days corresponding to the month of June; as can be seen the process model produces a good estimation of the soil moisture behavior with a correlation coefficient $R^2 = 0.9636$, with an average value of $R^2 = 0.9139$ for the five months, as described in Table 1.

4.3. Evaluated Methods. The objective of the automatic irrigation systems is to make an efficient use of water and energy by applying the right amount of water, at the right time and in the right place, in order to avoid, both, crop water stress and water waste. Many different commercial and research approaches have been proposed; based on the analysis on how these approaches apply control engineering to implement an automated irrigation system, five different irrigations methods were defined for the purpose of this work.

Level 0 (Empirical Open Loop Irrigation). There are no automation elements, irrigation is manually conducted based on the experience and labor from the farmer. This method is still widely used in today's agriculture. This method is not considered for the evaluation.

Level 1 (Time Based Open Loop Irrigation). The automated systems consist of a timer that activates pumps and valves on a predefined basis; no sensing elements are used. Irrigation decision is defined offline and based on farmer empirical knowledge.

Table 1: Model validation correlation coefficient results for each month.

Evaluated month (during 10 days)	Correlation coefficient R^2
April	0.9671
May	0.8798
June	0.9636
July	0.9139
August	0.9412
Average	0.9331

Level 2 (Feed-Forward Open Loop Irrigation). In this type of strategy controller applies irrigation to refill the water consumed by the crop and the environment. The irrigation system must be capable of measuring the crop evapotranspiration by using a sensing system or acquiring the data from near public weather stations. Typically farmers conduct this process on a weekly basis.

Level 3 (Closed-Loop Irrigation). The controller applies irrigation when sensors detect that measurements are below a predefined low threshold and stops irrigation when a high threshold has been reached. Typically soil water content is used as the measured variable.

Level 4 (Model-Based Closed-Loop Irrigation). The control system contains the mathematical model that describes the process dynamics and uses feed-forward and feedback strategies to implement advanced control laws and achieve optimal solutions. A model predictive control algorithm has been implemented in order to look for an optimal irrigation input sequence based on (20) and (21).

4.4. Results and Discussion. The second experimental field corresponds to four contiguous irrigation areas of approximate 20 × 10 meters each, for a total area of 80 × 10 meters, in order to evaluate the four irrigation methods. The type of soil is the same as in the first experimental field, and both fields are in the same physical location. Drip irrigation was used to water a green pepper crop. A 3/4 HP water pump with a maximum flow rate of 170 liters per minute was used to provide water for irrigation; each area had an on-off valve to activate the irrigation. Each irrigation area contains up to 70 drippers in order to provide 560 liters per hour to the area; a total of six soil moisture sensors were located for each area, as seen in Figure 10 indicated by the red-white circles. The evaluation was conducted during the months of September and October. The experimental field is located outside the city of Delicias, Chihuahua, in Mexico (latitude: 28.169149, longitude: −105.502768).

The four evaluated irrigation methods are compared in terms of accumulated error J_{acum} and control effort $J_{control}$. In both cases, the lower the value the better the performance.

The accumulated error indicates how good the system is to maintain the soil moisture levels close to the reference

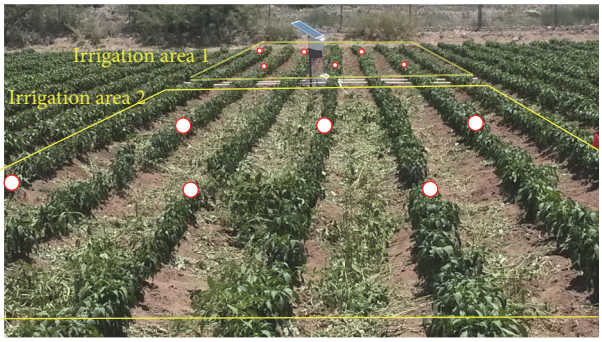


Figure 10: Experimental field with green pepper crop.

Table 2: Water consumption for the evaluated methods.

Evaluated method	Water consumption (in liters)	Level 4 water savings
Level 1	47,940	67.4%
Level 2	26,038	40.0%
Level 3	18,370	14.9%
Level 4	15,622	—

value. The maximum allowable depletion level (MAD) is considered as the process set-point or reference value, since below this level the crop becomes stressed and above it water may be wasted. The error indicates the difference between the current soil moisture value and the set-point, expressed in percentage of volumetric water content (VWC) as defined by (4). The accumulated error represents the sum of errors at the sampling instants during the evaluation period.

The control effort indicates how efficient the system is, in order to minimize the water consumption. The control effort is represented by effective irrigation and it is expressed in liters of water applied to the crop.

Accumulated error and control effort are defined as

$$J_{\text{acum}} = \sum_{i=0}^{T_{\text{eval}}-1} |e(kh)|,$$
$$J_{\text{control}} = \sum_{i=0}^{T_{\text{eval}}-1} |u(kh)|, \quad (22)$$

where T_{eval} is the evaluation time, $e(kh)$ is the difference between the soil moisture reference value and the process output (current soil moisture) at instant kh, and $u(kh)$ is the control signal (effective irrigation) at instant kh. The sampling period h for this evaluation is two minutes and the MPC finite horizon T_{FH} is one week. The time-delay τ from (8) equals 20 minutes, and the matrices Q and R from (19) have values of 0.0001 and 0.1, respectively, in order to give more weight to the control action (irrigation) rather than the error, since the main objective is to save water.

During an evaluation period of 30 days, the water consumption results for each evaluated method are shown in Table 2. The third column expresses the percentage of saved water obtained by Level 4 in comparison with the other methods.

Table 3: Irrigation error for the evaluated methods.

Evaluated method	Irrigation error	Level 4 error reduction
Level 1	100,039	71.3%
Level 2	45,958	37.5%
Level 3	36,741	21.8%
Level 4	28,719	—

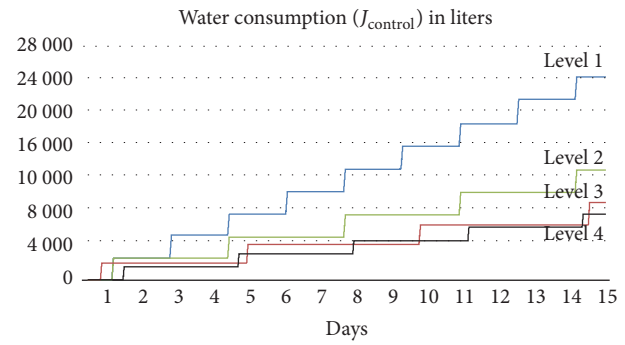


Figure 11: Water consumption evolution for the evaluated methods.

On Figure 11 the evolution of water consumption for each method is observed for the first 15 days of the evaluation period.

The irrigation error results and Level 4 error reduction percentage compared against the other methods are shown in Table 3.

On Figure 12 the evolution of irrigation error for each method is observed for the first 15 days of the evaluation period.

In general Level 4 method (MPC controller) offers the best performance considering the reference error and the control effort parameters. The implementation of the MPC controller requires an intensive computational load; however, high performance embedded devices and real-time kernels support the implementation of complex algorithms such as the required in an MPC controller. Also the relatively slow process dynamics for an irrigation system contribute to the implementation of a real-time predictive control strategy.

5. Conclusions

This paper proposes the use of a model driven control strategy for precision irrigation. Considering that the process dynamics of an irrigation system can be described with the hydrological balance model, evapotranspiration and soil moisture variables can be sensed in order to implement a model predictive control (MPC) to minimize the control signal (effective irrigation) while keeping soil moisture under specific thresholds (avoiding water stress) and considering external disturbances (reference evapotranspiration) to predict the process dynamics.

A recursive least squares algorithm has been used in order to estimate the model coefficients. These coefficients have been validated by using direct measurements from

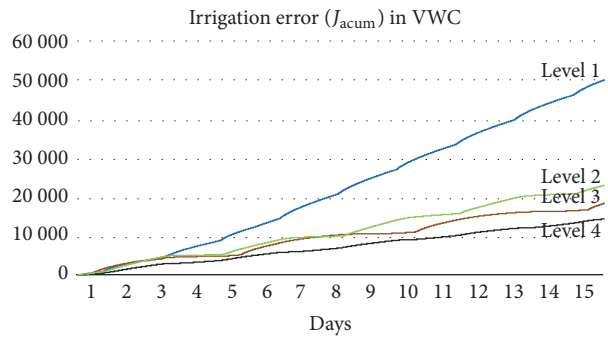


Figure 12: Irrigation error evolution for the evaluated methods.

the irrigation system. Then the proposed predictive control strategy has been implemented over an embedded platform, in order to evaluate the proposed irrigation method against the traditional methods used by the farmers. Experimental results indicate that the use of a model predictive control strategy in an irrigation system achieves a higher control efficiency and significantly reduce the control effort (water consumption).

Future work will focus on conducting the parameter estimation algorithm online and obtaining direct plant measurements by using imaging devices in order to evaluate the crop development and include this element as a variable in the MPC model.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the "Fondo Mixto de Fomento a la Investigación Científica y Tecnológica CONACyT, Gobierno del Estado de Chihuahua, México," for funding this project.

References

[1] UNESCO, *Water in a Changing World: United Nations World Water Development Report 3*, UNESCO Publishing, Paris, France, 2009.

[2] R. Romero, *Hydraulic modelling and control of the soilplantatmosphere continuum in woody crops [Ph.D. thesis]*, System Engineering and Automatic Department, University of Seville, 2011.

[3] R. Romero, J. L. Murielb, I. Garcíab, and D. Muñoz de la Peñac, "Research on automatic irrigation control: state of the art and recent results," *Agricultural Water Management*, vol. 114, pp. 59–66, 2012.

[4] B. Cardenas-Lailhacar, M. D. Dukes, and G. L. Miller, "Sensor-Based automation of irrigation on bermudagrass, during wet weather conditions," *Journal of Irrigation and Drainage Engineering*, vol. 134, no. 2, pp. 120–128, 2008.

[5] B. Cardenas-Lailhacar, M. D. Dukes, and G. L. Miller, "Sensor-based automation of irrigation on bermudagrass during dry weather conditions," *Journal of Irrigation and Drainage Engineering*, vol. 136, no. 3, pp. 184–193, 2010.

[6] Y. Erdem, L. Arin, T. Erdem et al., "Crop water stress index for assessing irrigation scheduling of drip irrigated broccoli (*Brassica oleracea* L. var. *italica*)," *Agricultural Water Management*, vol. 98, no. 1, pp. 148–156, 2010.

[7] S. Ooi, I. Mareels, N. Cooley, G. Dunn, and G. Thoms, "A systems engineering approach to viticulture on-farm irrigation," in *Proceedings of the 17th IFAC World Congress*, Seoul, South Korea, July 2008.

[8] R. Allen, L. Pereira, D. Raes, and M. Smith, *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*, FAO Irrigation and Drainage Paper 56, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, 1998.

[9] S. Saleem, B. Kithsiri, L. Yue, M. Halmamuge, and H. Malano, "Model predictive control for real-time irrigation scheduling," in *Proceedings of the 4th IFAC Conference on Modelling and Control in Agriculture*, Espoo, Finland, 2013.

[10] K. X. Soulis, S. Elmaloglou, and N. Dercas, "Investigating the effects of soil moisture sensors positioning and accuracy on soil moisture based drip irrigation scheduling systems," *Agricultural Water Management*, vol. 148, pp. 258–268, 2015.

[11] J. Gutierrez, J. F. Villa-Medina, A. Nieto-Garibay, and M. A. Porta-Gandara, "Automated irrigation system using a wireless sensor network and GPRS module," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 1, pp. 166–176, 2014.

[12] H. Navarro-Hellín, R. Torres-Sánchez, F. Soto-Valles, C. Albaladejo-Pérez, J. A. López-Riquelme, and R. Domingo-Miguel, "A wireless sensors architecture for efficient irrigation water management," *Agricultural Water Management*, vol. 151, pp. 64–74, 2015.

[13] Z. Li, N. Wang, T. S. Hong, A. Franzen, and J. N. Li, "Closed-loop drip irrigation control using a hybrid wireless sensor and actuator network," *Science China Information Sciences*, vol. 54, no. 3, pp. 577–588, 2011.

[14] Y. Kim, R. G. Evans, and W. M. Iversen, "Evaluation of closed-loop site-specific irrigation with wireless sensor network," *Journal of Irrigation and Drainage Engineering*, vol. 135, no. 1, pp. 25–31, 2009.

[15] C. Lozoya, C. Mendoza, L. Mejía et al., "Model predictive control for closed-loop irrigation," in *Proceedings of the 19th IFAC World Congress*, Cape Town, South Africa, August 2014.

[16] J. Maciejowski, *Predictive Control with Constraints*, Prentice Hall, London, UK, 1st edition, 2000.

[17] V. Puig, C. Ocampo-Martinez, J. Romera et al., "Model predictive control of combined irrigation and water supply systems: application to the Guadiana river," in *Proceedings of the 9th IEEE International Conference on Networking, Sensing and Control (ICNSC '12)*, pp. 85–90, Beijing, China, April 2012.

[18] A. Sadowska, B. De Schutter, and P.-J. Van Overloop, "Delivery-oriented hierarchical predictive control of an irrigation canal: event-driven versus time-driven approaches," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 5, pp. 1701–1716, 2015.

[19] K. Horvath, M. Petreczky, L. Rajaoarisoa, E. Duviella, and K. Chuquet, "MPC control of water level in a navigation canal—the cuinchy-fontinettes case study," in *Proceedings of the 13th European Control Conference (ECC '14)*, pp. 1337–1342, IEEE, Strasbourg, France, June 2014.

[20] J. Guevara-Diaz, "The use of the 1998 Penman-Monteith FAO formula in order to determine referential evapotranspiration," *Terra*, vol. 22, no. 31, pp. 31–72, 2006.

[21] D. Rowell, *Soil Science Methods and Applications*, John Wiley & Son, New York, NY, USA, 1st edition, 1994.

[22] S. Ghosh, D. M. Bell, J. S. Clark, A. E. Gelfand, and P. G. Flikkema, "Process modeling for soil moisture using sensor network data," *Statistical Methodology*, vol. 17, pp. 99–112, 2014.

[23] C. Lozoya, G. Mendoza, C. Mendoza, V. Torres, and M. Grado, "Experimental evaluation of data aggregation methods applied to soil moisture measurements," in *Proceedings of the IEEE SENSORS*, pp. 134–137, IEEE, Valencia, Spain, November 2014.

[24] L. Ljung, *System Identification: Theory for the User*, Prentice Hall, New Jersey, NJ, USA, 2nd edition, 1999.

[25] C. Lozoya, A. Aguilar, and C. Mendoza, "Service oriented design approach for a precision agriculture datalogger," *IEEE Latin America Transactions*, vol. 14, no. 4, pp. 1683–1688, 2016.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 7964727, 11 pages
http://dx.doi.org/10.1155/2016/7964727

Research Article

A Model Reference Adaptive Control/PID Compound Scheme on Disturbance Rejection for an Aerial Inertially Stabilized Platform

Xiangyang Zhou, Chao Yang, and Tongtong Cai

School of Instrumentation Science & Opto-Electronics Engineering, Beihang University, Beijing 100191, China

Correspondence should be addressed to Xiangyang Zhou; xyzhou@buaa.edu.cn

Received 17 June 2016; Revised 14 August 2016; Accepted 4 September 2016

Academic Editor: Rafael Morales

This paper describes a method to suppress the effect of nonlinear and time-varying mass unbalance torque disturbance on the dynamic performances of an aerial inertially stabilized platform (ISP). To improve the tracking accuracy and robustness of the ISP, a compound control scheme based on both of model reference adaptive control (MRAC) and PID control methods is proposed. The dynamic model is first developed which reveals the unbalance torque disturbance with the characteristic of being nonlinear and time-varying. Then, the MRAC/PID compound controller is designed, in which the PID parameters are adaptively adjusted based on the output errors between the reference model and the actual system. In this way, the position errors derived from the prominent unbalance torque disturbance are corrected in real time so that the tracking accuracy is improved. To verify the method, the simulations and experiments are, respectively, carried out. The results show that the compound scheme has good ability in mass unbalance disturbance rejection, by which the system obtains higher stability accuracy compared with the PID method.

1. Introduction

For an aerial remote sensing system, due to the serious effects of internal and external disturbances, the movement of the aircraft is not ideal that makes the sensor's line of sight (LOS) jitter, eventually resulting in the degradation of images quality [1–4]. In order to obtain high-resolution images and satisfy the requirements of high photo overlapping ratio, the sensor's LOS must be strictly controlled. Therefore, inertially stabilized platform (ISP) is a key component for the high-precision aerial remote sense imaging system, which is used to hold and control the LOS of the imaging sensor to keep steady relative to the inertial space or the tracked target [5, 6]. The ISP with high control precision is indispensable for isolating disturbances derived from diverse sources [7, 8], particularly for the case of swings of three angular attitudes of aircraft. It is a principal issue for the control system of ISP of how to minimize the effects of disturbances introduced on the ISP [2].

The most critical performance metric for an ISP is torque disturbance rejection. It is difficult for the conventional PID control method in low speed servo motion to suppress these complex nonlinear disturbances. It is hard for the traditional feedforward control method to further improve the ISP's dynamic performance [9]. Therefore, there is continuous interest for researchers to develop the control methods with higher accuracy and stability by various disturbances rejection. The development of computer technology and advanced intelligent control theory provide a new way for the control of complex dynamic uncertain systems and the disturbance rejection. They have been gradually used in the control of ISP, such as neural network [10], genetic algorithm [11, 12], fuzzy control [13, 14], robust control [15], state compensation control [16], and autodisturbance rejection control [17]. Predictive control and fuzzy control are the effective methods to optimize the control of uncertain systems. The combination of the two methods can enable the system to have the quicker dynamic response and smaller overshot [11–14]. For a system characterized by nonlinear and time-varying behavior, the issue of stability performance becomes very prominent. In [1], the decoupling compensation controller obtains a good result in which the angular velocity

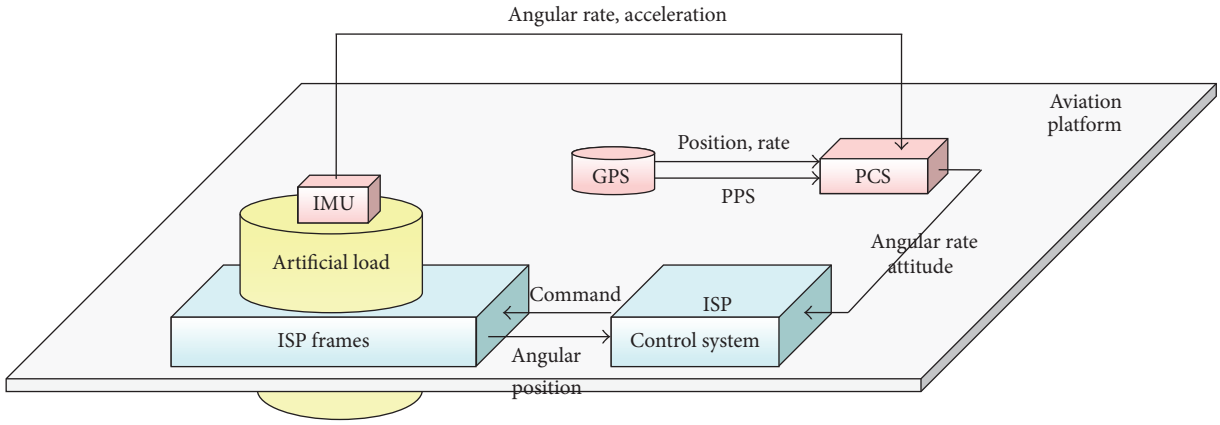


Figure 1: Schematic diagram of an aerial remote sensing system.

coupling, torque coupling, and moment of inertia coupling are considered. In [18], a three-closed loop PID compound control scheme is applied to a two-axis ISP to obtain the desired results. In [19], a feedforward compensation scheme is proposed to achieve vibration rejection of ISP. In [20], an active disturbance rejection control strategy is put forward to improve robustness of electrode regulator system.

Disturbances arise from diverse sources; for example, the angular motion and linear vibration of aircraft platform generate the disturbance torques due to mass unbalance and gimbal geometry [4, 5]. Generally, the influence of unbalance torque is prominent over other disturbances. Since the centroid is not exactly coincident with the center of the rotating axis, the mass unbalance torque will occur when ISP operates, which will severely degrade the system control accuracy. To compensate the mass unbalance torque, the static equilibrium test is required before the operation. However, due to many nonideal cases such as different imaging sensors being installed interchangeably, it is hard to completely eliminate the unbalance torque by static equilibrium mass correction. Moreover, since the mass unbalance torque is a nonlinear and time-varying disturbance, it is difficult for the conventional control to solve it [21]. Therefore, it is necessary to compensate the mass unbalance torque by using the intelligent control methods which have strong adaptive disturbance rejection ability. Previously, some methods have been proposed to compensate the unbalance torques, such as the adaptive control based on disturbance observer [22], the neural network control [10], the feedforward control [23], independent mechanisms [24], fuzzy inference mechanism [25], iterative feedback tuning of fuzzy control [26], adaptive neural network control [27], and so on.

Model reference adaptive control (MRAC) can restrain the influences of external disturbance by effectively revising the model parameter errors. MRAC does not need the online identification of the mathematical model of mass unbalance torque, by which the time of adaptive control is greatly shortened. Therefore, MRAC is appropriate for the occasion of parameters change [21]. In MRAC, an adaptive reference model needs to be designed which can achieve the desired performance index with the same order of the plant. In [28], a MRAC system based on the certainty equivalent (CE) principle for the first-order delay system is proposed. MRAC/PID compound control scheme is a combination of adaptive control and traditional PID control, which can make the PID parameters of the nonlinear time-varying uncertain systems adjusted in real time, so as to improve the system robustness and control accuracy. MRAC/PID controller owns the great robust ability for the nonlinear, hysteresis, and variable parameters systems. Compared with the conventional PID control method, the MRAC/PID controller can tune the PID parameters automatically and make the system stable in the whole working range [29].

In this paper, to improve the tracking accuracy and robustness of an aerial inertially stabilized platform, a MRAC/PID compound control scheme is proposed to weaken the influence of prominent unbalance torque disturbance. The dynamic model is first developed which reveals the unbalance torque disturbance with the characteristic of being nonlinear and time-varying. Then, a MRAC/PID compound controller is designed and simulation analysis is conducted. To verify the method, the experiments are carried out.

2. Background

2.1. Aerial Remote Sensing System. Figure 1 shows the schematic diagram of an aerial remote sensing system. Generally, an aerial remote sensing system consists of four main components, a three-axis ISP, an imaging sensor, a position and orientation system (POS), and the aviation platform. When applied, the three-axis ISP is mounted on the aviation platform, and the imaging sensor and POS are mounted on inner azimuth gimbal of the ISP. When the aviation platform rotates or jitters, the control system of three-axis ISP gets the high-precision attitude reference information measured by POS and then routinely controls the LOS of imaging sensor to achieve accurate pointing and stabilizing relative to ground level and flight track. The POS, which is mainly composed of three main components, that is, inertial measurement unit (IMU), GPS receiving antenna, and data processing system, is used to provide an accurate reference of position and attitude

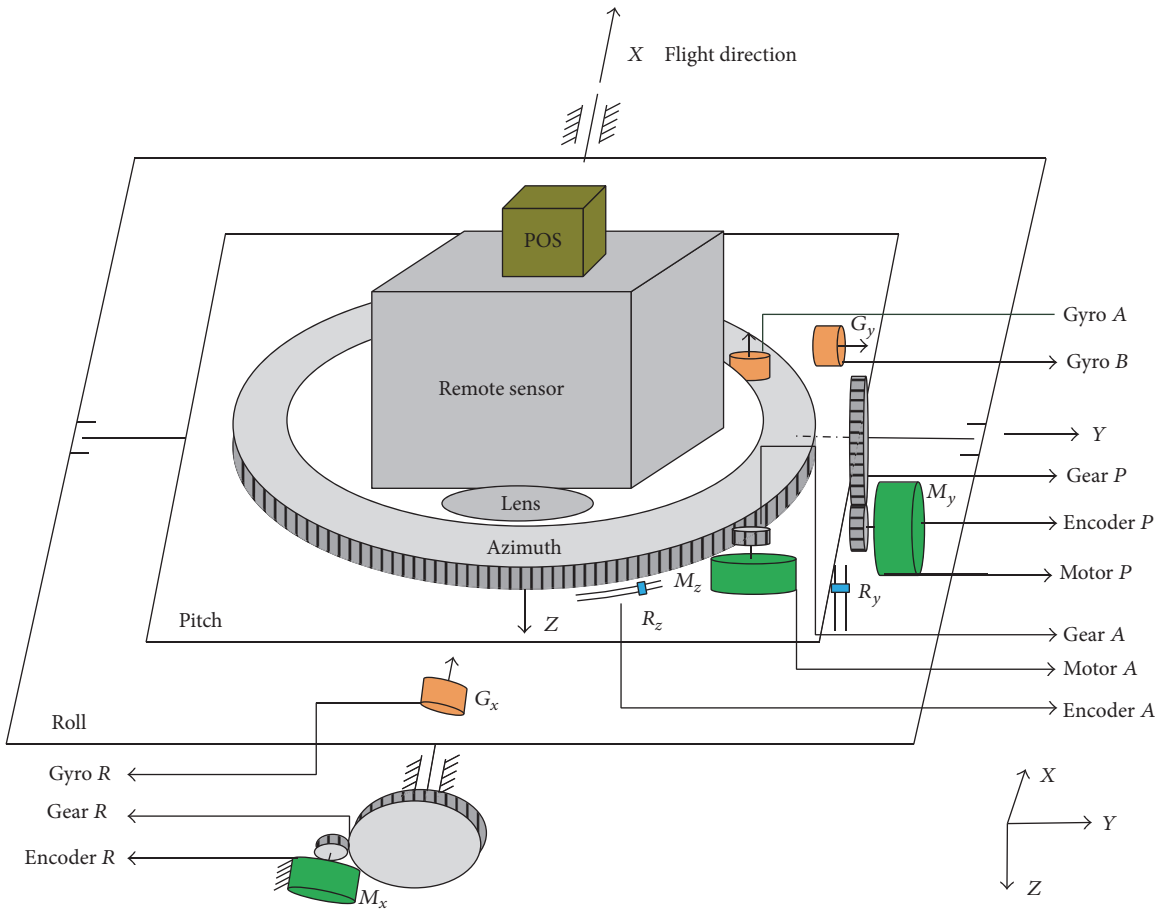


Figure 2: Schematic diagram of the three-axis ISP's working principle.

in inertial space for control system of ISP and imaging sensor through measuring the angular movement of imaging sensor.

2.2. Working Principal of ISP. Figure 2 shows the schematic diagram of the three-axis ISP principle. We can see that the ISP consists of three gimbals, which are azimuth gimbal (A-gimbal), pitch gimbal (P-gimbal), and roll gimbals (R-gimbal). Among them, the A-gimbal is assembled on the P-gimbal and can rotate around Z_a axis. Likewise, the P-gimbal is assembled on the R-gimbal and can rotate around X_p axis. The R-gimbal is assembled on the base of aviation platform and can rotate around Y_r axis. From Figure 2, we can see the relationships between three gimbals: G_p, G_r, and G_a, respectively, stand for rate gyro that measures inertial angular rate of P-gimbals, R-gimbals, and A-gimbals. E_x, E_p, and E_a, respectively, stand for photoelectric encoder which measures relative angular between gimbals. M_r, M_p, and M_a, respectively, stand for gimbals servo motor which drives R-gimbals, P-gimbal, and A-gimbal to keep these three gimbals steady in inertial space. Ax and Ay, respectively, represent accelerometers installed on the R-gimbal and R-gimbal used to measure the gimbals' rotary angular acceleration. E_x and E_y represent encoders installed on two leveling gimbals to detect the gimbals' rotary angular position.

2.3. Three-Closed Loop Compound Control Scheme. Conventional stabilization techniques employ rate gyros, rate integrating gyros, or rate sensors to sense rate disturbances about the LOS. Figure 3 shows the block diagram of traditional three-loop control system for ISP. In Figure 3, the blocks of G-pos, G-spe, and G-cur separately represent the controllers in the position loop, speed loop, and current loop; the PWM block represents the power amplification used for the current amplification to drive the torque motor; L represents the inductance of a torque motor and R represents the resistance; K_t represents the torque coefficient of the motor and N is the transition ratio from the torque motor to the gimbals; J_m represents the moment of inertia of the motor and J_l represents the moment of inertia of the gimbals along the rotation axis.

Figure 4 shows the process model of the ISP control scheme block diagram. G_{AP}, G_{PP}, and G_{RP} are the PID controllers for position loop of azimuth gimbal, pitch gimbal, and roll gimbal, respectively. G_{AR}, G_{PR}, and G_{RR} are the PID controllers for rate loop of azimuth gimbal, pitch gimbal,

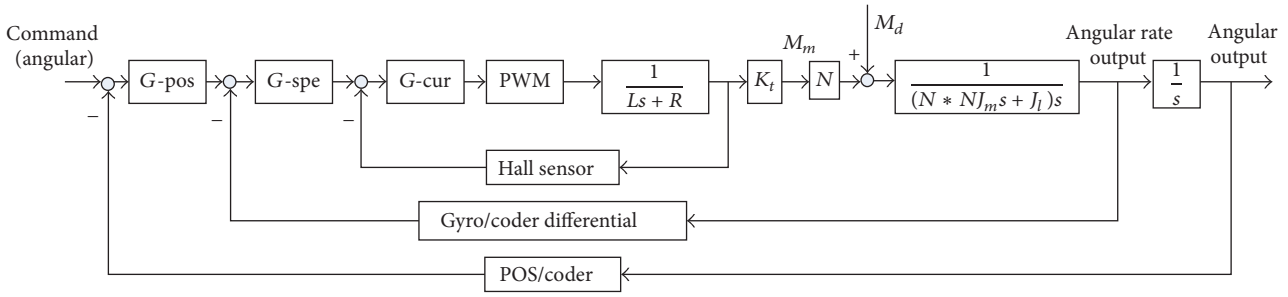


Figure 3: A block diagram of traditional three-loop control system for ISP.

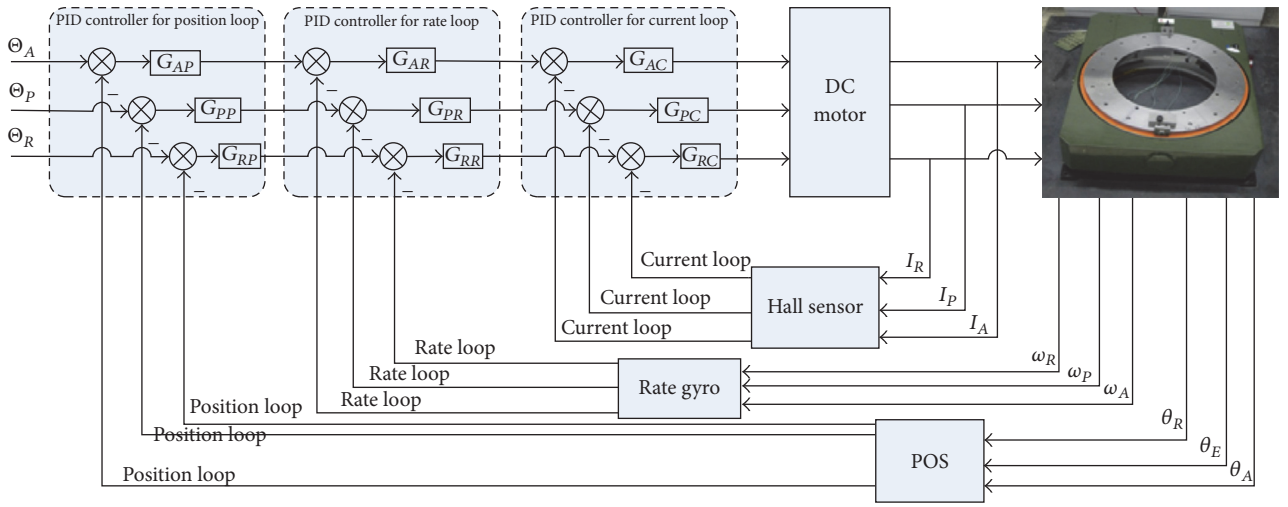


Figure 4: The process model of the ISP control scheme block diagram.

and roll gimbal, respectively. G_{AC}, G_{PC}, and G_{RC} are the PID controllers for current loop of azimuth gimbal, pitch gimbal, and roll gimbal, respectively.

3. Dynamic Modeling of the Mass Unbalance Torque

Due to the influence of mechanical structure design and machining accuracy, three-axe ISP's center of mass is not coincident with the center of rotating shaft completely. So, when there is acceleration that acted on the ISP gimbals, the mass unbalance torque disturbance occurs [23], which will severely degrade the system control accuracy. Mass unbalance torque under the static base is caused by both of the gravity and gimbal acceleration.

The mass imbalance produces LOS jitter when the payload center of gravity is not centered on an axis of rotation for the gimbals. Linear vibration, acting through the lever arm of the center of gravity offset, thus produces torque disturbances. When an ISP is working at a flying aircraft, due to both effects of gravity acceleration and the motion acceleration of ISP's gimbals and imaging sensors, the mass imbalance torque occurs. Figure 5 shows the schematic diagram of the functional mechanism of mass imbalance torque under moving base and static base, respectively.

In Figure 5(a), l_x and l_z, respectively, stand for the eccentric lever arms of mass center of ISP relative to a horizontal rotation axis (x or y) and the vertical axis-z. a_f and a_z, respectively, stand for the horizontal and vertical interference accelerated speed during movement acting on the gimbal and g stands for acceleration of gravity. If we take the counterclockwise direction as a positive direction, imbalance torque can be expressed as

$$T_{\text{im}} = -m\left(a_z + g\right) \cdot l_x - m \cdot a_f \cdot l_z, \tag{1}$$

where T_{im} is the unbalance torque of the moving base and m is the total mass of the frame and the load,

$$T_{\text{motor}} = -\frac{T_{\text{im}}}{N} = -\frac{1}{N}\left[-m\left(a_z + g\right) \cdot l_x - m \cdot a_f \cdot l_z\right], \tag{2}$$

where T_{motor} is the extra unbalance torque produced by motor. N is the transmission ratio.

The unbalance torque caused by gravity can be expressed as

$$T_g = mgl_z \sin\theta - mgl_x \cos\theta. \tag{3}$$

Since the leveling angles of ISP are changed in a small range of about $\theta = \pm 5^{\circ}$, the cosine and sine function values of θ are approximately equal to 1 and 0, respectively, so the

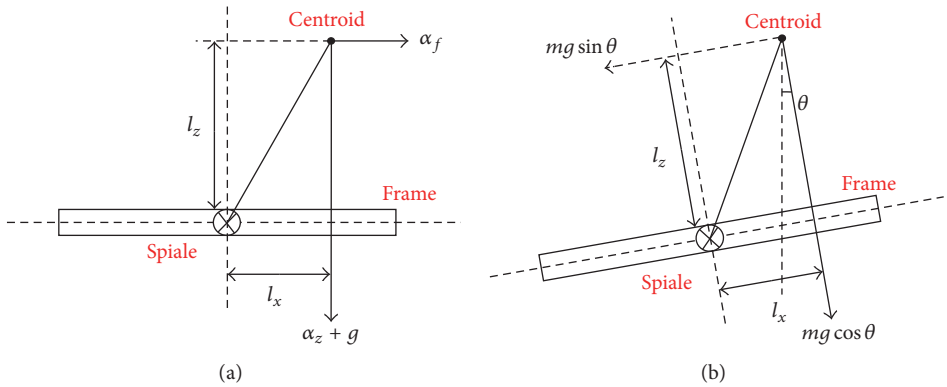


FIGURE 5: The geometrically schematic diagram to show the functional mechanism of mass imbalance torque: (a) under moving base condition and (b) under static base condition.

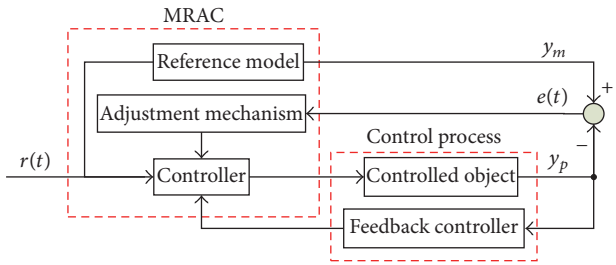


FIGURE 6: The structure of a model reference adaptive control (MRAC) system.

unbalance torque caused by gravity can be further expressed as

$$T_g = -mgl_x \cos\theta. \tag{4}$$

4. MRAC/PID Compound Controller Design

4.1. Adaptive Model Reference Control Principal. Figure 6 shows the structure of an adaptive model reference control (MRAC) system. It is composed of four components, that is, control process, controller, reference model, and adjustment mechanism block [30].

The adaptive model reference control (MRAC) technique is used for devising a controller based on the information y_m, y_p, $r(t)$, and $e(t)$. The adaptive mechanism automatically adjusts controller parameters so that the behavior of the closed loop control plant output y_p closely follows that of y_m of the reference model. Parameters and structure of reference model are specified on the base of requirements of control performance. The adjustment mechanism of MRAC system is constructed by adaptive control rule, which performs the algorithms as follows.

Tracking error is defined as

$$e = y_m - y_p, \tag{5}$$

where e is error of input and output, y_m is the output of the reference model, and y_p is the output of the actual system. Thus, e will eventually be close to a constant or 0. If e is bounded, then y_p is bounded. So the control system is convergent under this controller.

4.2. MRAC/PID Compound Controller. PID control is one of the most commonly used control methods in engineering. It has the advantages of simple algorithm and high reliability, but it needs more accurate system model [31]. When the input signal is applied to a variable structure control system, the system should be stable while altering controller coefficients according to control error signal.

According to Figure 3, the bandwidth of the current loop is much larger than the bandwidth of the stable loop, so the current loop is regarded as a proportional component with coefficient of 1. The transfer functions of the G-pos and G-spe are too small and can be equivalent to 1. $k_c((\tau_c s + 1)/s)$ is the transfer function of G-cur, and $U/(T_{\mathrm{PWM}}S + 1)$ is the transfer function of the PWM [32]. In the simulation, the values of these parameters are too small to be ignored for 0, such as L, N, τ_c, T_{PWM}, and J_m. So the ISP control system is established as the two-order system model.

Based on [29, 30], the controller is designed as follows:

$$J(\theta) = \frac{1}{2}e^2(\theta), \tag{6}$$

where the time rate of change of θ stands for proportional to negative gradient of J,

$$\frac{d\theta}{dt} = -\gamma\frac{\partial J}{\partial \theta} = -\gamma e\frac{\partial e}{\partial \theta}, \tag{7}$$

where θ stands for the controller parameter vector. The components of $\partial e/\partial\theta$ stand for the sensitivity derivatives of the error with respect to θ. The parameter γ stands for known as the adaptation gain.

Considering an aerial ISP system described by second-order model $b/(s^2+\alpha_1 s+\alpha_2)$, the closed loop transfer function is

$$\frac{y_p(s)}{r(s)} = \frac{b\left(K_d s^2 + K_p s + K_i\right)}{s\left(s^2 + \alpha_1 s + \alpha_2\right) + b\left(K_d s^2 + K_p s + K_i\right)}, \quad (8)$$

$$\frac{y_p(s)}{r(s)} = \frac{b\left(K_d s^2 + K_p s + K_i\right)}{s^3 + \left(\alpha_1 + bK_d\right)s^2 + \left(\alpha_2 + bK_p\right)s + bK_i}. \quad (9)$$

From (9) and required performance of system, we obtain a reference model as follows:

$$\frac{y_m(s)}{r_m(s)} = \frac{b_{m1}s^2 + b_{m2}s + b_{m3}}{s^3 + a_{m1}s^2 + a_{m2}s + a_{m3}}, \quad (10)$$

where $a_{m1} = \alpha_1 + bK_d$, $a_{m2} = \alpha_2 + bK_p$, $a_{m3} = bK_i$, $b_{m1} = bK_d$, $b_{m2} = bK_p$, and $b_{m3} = bK_i$.

The values of PID controller parameters are determined, K_p, K_i, and K_d, in

$$\begin{aligned}
\frac{dK_p}{dt} &= -\gamma_p \frac{\partial J}{\partial K_p} = -\gamma_p \left(\frac{\partial J}{\partial e}\right)\left(\frac{\partial e}{\partial y_p}\right)\left(\frac{\partial y_p}{\partial K_p}\right),\\
\frac{dK_i}{dt} &= -\gamma_i \frac{\partial J}{\partial K_i} = -\gamma_i \left(\frac{\partial J}{\partial e}\right)\left(\frac{\partial e}{\partial y_p}\right)\left(\frac{\partial y_p}{\partial K_i}\right),\\
\frac{dK_d}{dt} &= -\gamma_d \frac{\partial J}{\partial K_d} = -\gamma_d \left(\frac{\partial J}{\partial e}\right)\left(\frac{\partial e}{\partial y_p}\right)\left(\frac{\partial y_p}{\partial K_d}\right),
\end{aligned} \quad (11)$$

where $\partial J/\partial e = e$, $\partial e/\partial y = 1$, $D = d/dt$, and

$$\begin{aligned}
\frac{\partial y_p}{\partial K_p} &= \frac{bD}{D^3 + \left(\alpha_1 + bK_d\right)D^2 + \left(\alpha_2 + bK_p\right)D + bK_i} \cdot \left[r - y_p\right],\\
\frac{\partial y_p}{\partial K_i} &= \frac{b}{D^3 + \left(\alpha_1 + bK_d\right)D^2 + \left(\alpha_2 + bK_p\right)D + bK_i} \cdot \left[r - y_p\right],\\
\frac{\partial y_p}{\partial K_d} &= \frac{bD^2}{D^3 + \left(\alpha_1 + bK_d\right)D^2 + \left(\alpha_2 + bK_p\right)D + bK_i} \cdot \left[r - y_p\right].
\end{aligned} \quad (12)$$

Then dK_p/dt, dK_i/dt, and dK_d/dt can be derived by

$$\begin{aligned}
\frac{dK_p}{dt} &= -\gamma_p \frac{\partial J}{\partial K_p}\\
&= -\gamma_p e \frac{bD}{D^3 + \left(a_1 + bK_d\right)D^2 + \left(a_2 + bK_p\right)D + bK_i} \cdot \left[r - y_p\right],\\
\frac{dK_i}{dt} &= -\gamma_i \frac{\partial J}{\partial K_i}\\
&= -\gamma_i e \frac{b}{D^3 + \left(a_1 + bK_d\right)D^2 + \left(a_2 + bK_p\right)D + bK_i} \cdot \left[r - y_p\right],\\
\frac{dK_d}{dt} &= -\gamma_d \frac{\partial J}{\partial K_d}\\
&= -\gamma_d e \frac{bD^2}{D^3 + \left(a_1 + bK_d\right)D^2 + \left(a_2 + bK_p\right)D + bK_i} \cdot \left[r - y_p\right].
\end{aligned} \quad (13)$$

If the adaptive control law is given, the system can guarantee the stability of the system under the action of the model reference adaptive PID controller and can achieve the control goal. Thus, a new PID control based on the reference model is formed. The basic structure of the MRAC/PID compound control system is shown in Figure 7. In Figure 7, the blocks of G-pos, G-spe, and G-cur separately represent the controllers; the PWM block represents the power amplification used for the current amplification to drive the torque motor.

5. Simulation Analysis

5.1. Tracking Performance. According to the adaptive model reference control (MRAC) method, through the parameters adjustment simulation environment, the system can get the ideal output.

The PID parameter values of the reference model are selected by the parameter tuning or the test calculation method, as shown in Figures 8 and 9.

So the result of the PID parameter value is selected and taken out by the reference of the automatic adjustment: $K_p = 10$, $K_i = 0$, and $K_d = 4$. Based on the fixed parameters in the three-loop control system of ISP, the reference model is

$$\frac{y_m(s)}{r_m(s)} = \frac{55.44 + 221.76s + 7.56s^2}{55.44 + 221.76s + 7.56s^2 + 0.018s^3}. \quad (14)$$

Further, by (14), the control strategy of the adaptive PID controller is adjusted by the MRAC/PID compound controller, and the stability and accuracy of the system are improved continuously.

First, the responses for the step input are analyzed. Figure 10 shows the system response curve and its partial

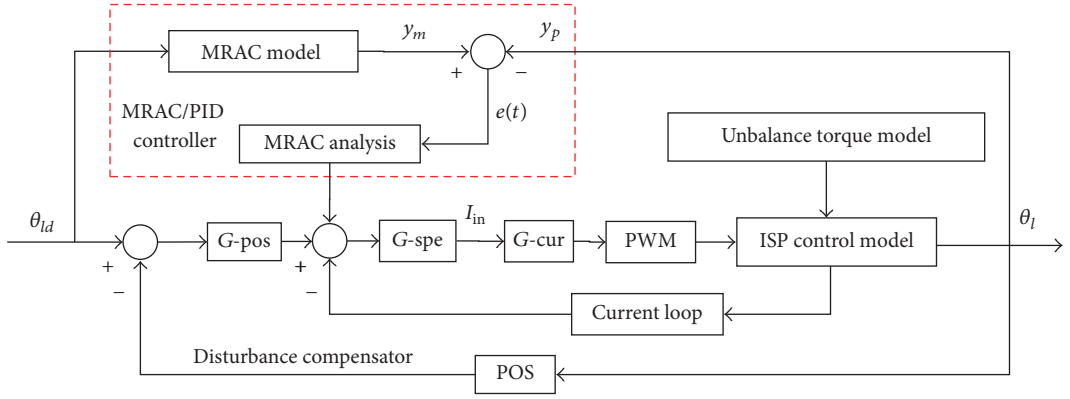


Figure 7: Block diagram of MRAC/PID compound controller.

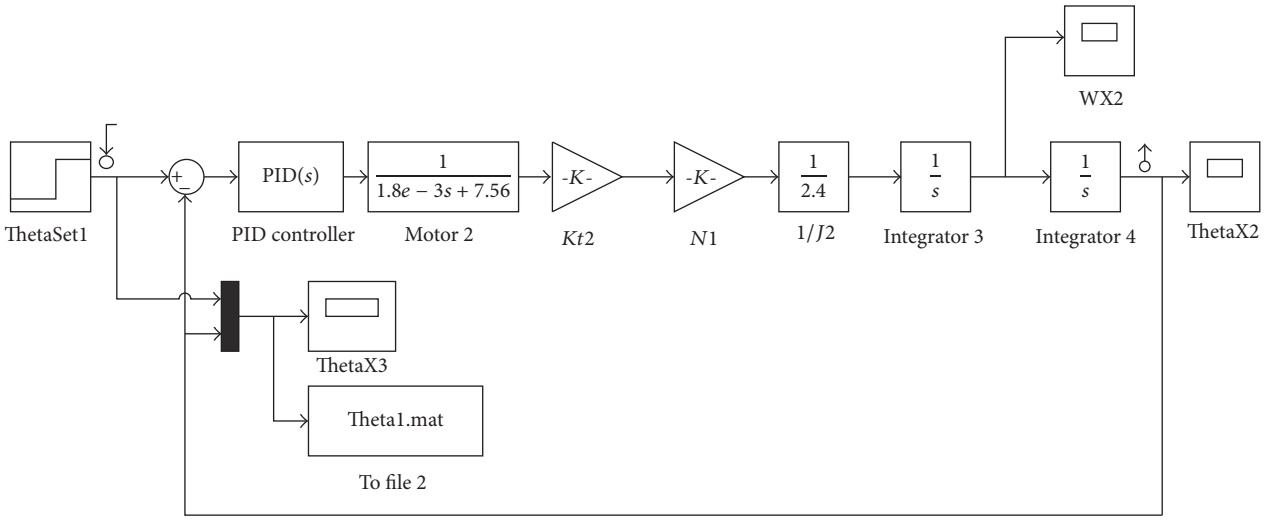


Figure 8: Automatic adjustment of PID parameters for reference model design.

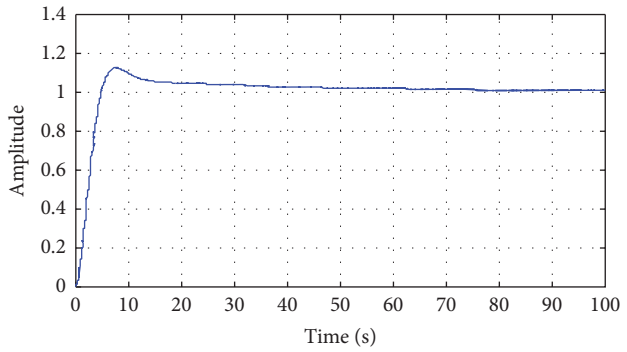


Figure 9: Automatic adjustment of PID parameters model.

enlarged detail for the traditional PID controller and the MRAC/PID compound controller. As seen in the figures, compared with PID control, the stability time and accuracy of the MRAC/PID compound control are much shorter and higher.

5.2. Adaptive PID Parameters. Figure 11 shows the parameters variation curves of the MRAC/PID compound controller.

From Figures 10 and 11, we see that the MRAC/PID compound control scheme can obviously improve the accuracy of the control system. Under the same unbalance torque disturbance conditions, the position output peak-valley errors of MRAC/PID compound scheme and traditional PID control method are +0.0075° to −0.006° and +0.035° to −0.05°, respectively, meaning that the position accuracy is improved up to 84.1% after MRAC/PID scheme applied. The simulation results illustrate the compound scheme has high disturbance rejection ability compared with the PID controller.

6. Experimental Verification

In order to validate the proposed MRAC/PID compound scheme, the experiments are then performed, which are conducted on a three-axis ISP.

According to the specific functional requirements of the system, the system hardware circuit is designed. Figure 12

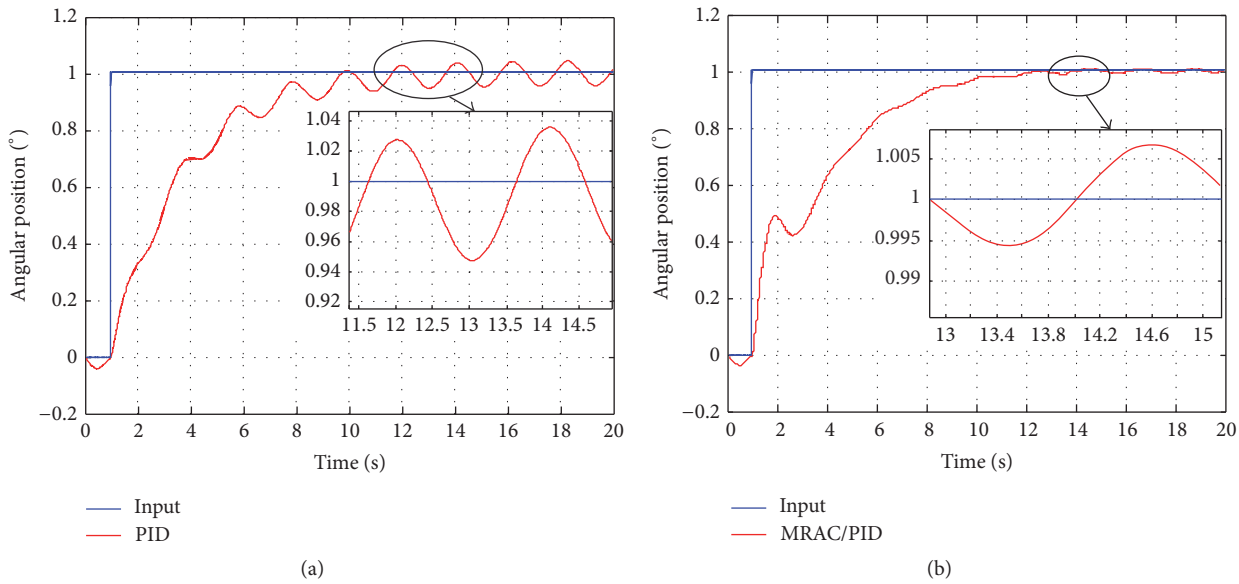


FIGURE 10: System response curves and partial enlarged detail for the (a) traditional PID controller and (b) MRAC/PID compound controller.

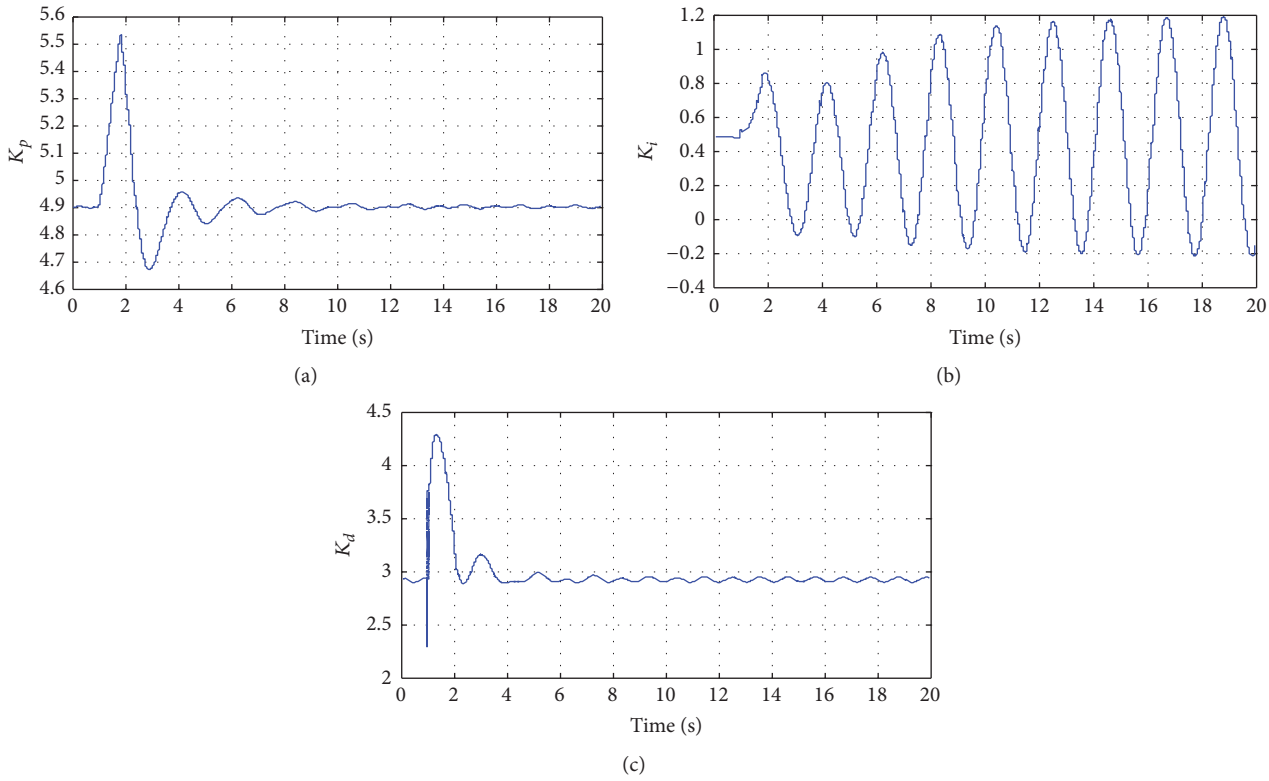


FIGURE 11: The parameter variation curves of the MRAC/PID compound controller: (a) K_p parameter; (b) K_i parameter; (c) K_d parameter.

shows the hardware control system circuit connection diagram of three-axis ISP. The main functional devices are DC torque motor, POS, gyroscope, encoder, and so forth.

The main program flowchart is shown in Figure 13. After the system is reset and the external device is initialized, the external device interrupt is opened. The six interrupt management procedures include ADC sequencer ADCSEQ interrupt, interrupt timer CPU-Timer 0, SPI receiving interrupt, SCI A sending/receiving interrupt, SCI B sending/receiving interrupt, and SCI C receiving interrupt using the software

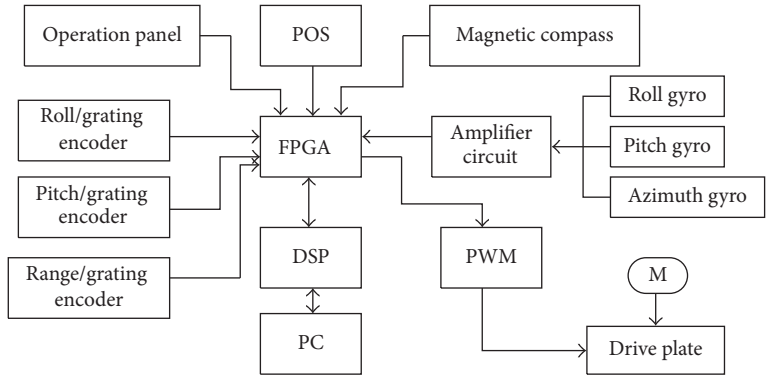


Figure 12: Hardware control system circuit connection diagram.

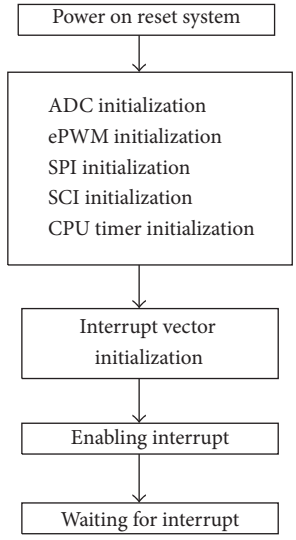


Figure 13: Main program flowchart.

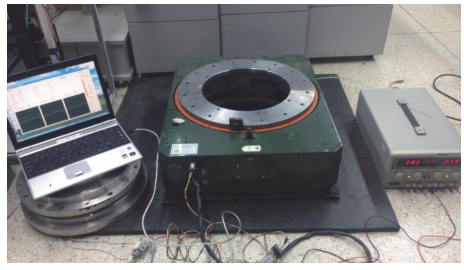


Figure 14: The three-axis ISP experimental system.

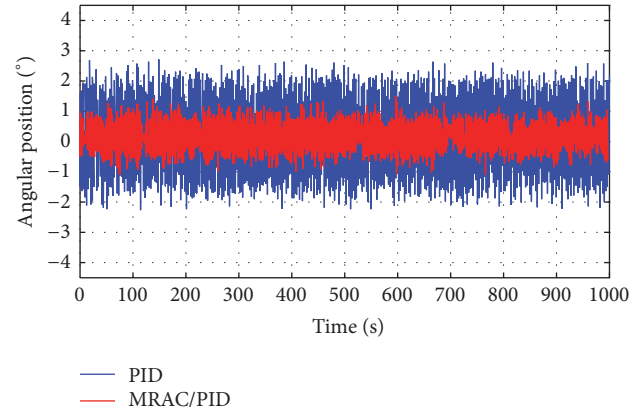


Figure 15: Experiment comparison between the PID and adaptive MRAC/PID controllers under unbalance disturbance.

system of the platform. Platform interrupt priority order is ADCSEQ > CPU-Timer 0 > SPI > SCI A > SCI B > SCI C.

Figure 14 shows the picture of the experimental system. The main parameters of the ISP are as follows: maximum load and self-weight are, respectively, 80 kg and 40 kg, the maximum leveling rotation angle range is ±5°, and the maximum heading rotation angle range is ±25°. In the experiments, the artificial load is 20 kg and the power supply voltage is 28 V. As a comparison, the results obtained by PID controller are also displayed.

The experiments are designed aiming at rejecting unbalanced torque disturbance. From (4), we know that the unbalance torque is nonlinear and time-varying with characteristic of cosine function. Therefore, the cosine interference component is artificially added into the current loop of the gimbal system to represent the unbalance torque disturbance. Thus, we can evaluate the disturbance rejection ability of the compound scheme.

Figure 15 shows the experiment comparison between the PID and adaptive MRAC/PID controllers when the interference current $I = 0.5\cos(0.4\pi t) + \cos(0.9\pi t)$. We can see that, for the MRAC/PID and PID controllers, the output error ranges of the attitude angle are about ±1° and ±2°, respectively. Obviously, compared with the PID controller, the position accuracy is improved up to nearly 50% by the MRAC/PID compound controller.

7. Conclusion

In this paper, to suppress the nonlinear and time-varying mass unbalance torque disturbance of the aerial-axis inertially stabilized platform (ISP), a compound control scheme

based on both of model reference adaptive control (MRAC) and PID control methods is proposed. In this way, the tracking accuracy and stability of the ISP are improved significantly. To verify the method, the simulations and experiments are, respectively, carried out. The results show that MRAC/PID compound scheme is good at the disturbance rejection. Compared with the PID, the tracking accuracy of the MRAC/PID compound controller is improved by about 50%.

Competing Interests

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

Acknowledgments

This work was funded by the National Natural Science Foundation of China (nos. 51375036, 51205019, and 61573040) and the China Scholarship Council (CSC).

References

[1] X. Zhou, H. Zhang, and R. Yu, "Decoupling control for two-axis inertially stabilized platform based on an inverse system and internal model control," *Mechatronics*, vol. 24, no. 8, pp. 1203–1213, 2014.

[2] M. K. Masten, "Inertially stabilized platforms for optical imaging systems," *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 47–64, 2008.

[3] J. M. Hilkert, "Inertially stabilized platform technology: concepts and principles," *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 26–46, 2008.

[4] P. J. Kennedy and R. L. Kennedy, "Direct versus indirect line of sight (LOS) stabilization," *IEEE Transactions on Control Systems Technology*, vol. 11, no. 1, pp. 3–15, 2003.

[5] X. Zhou, Y. Jia, Q. Zhao, and T. Cai, "Dual-rate-loop control based on disturbance observer of angular acceleration for a three-axis aerial inertially stabilized platform," *ISA Transactions*, vol. 63, pp. 288–298, 2016.

[6] M. M. Abdo, A. R. Vali, A. R. Toloei, and M. R. Arvan, "Stabilization loop of a two axes gimbal system using self-tuning PID type fuzzy controller," *ISA Transactions*, vol. 53, no. 2, pp. 591–602, 2014.

[7] J. L. Miller, S. Way, B. Ellison, and C. Archer, "Design challenges regarding high-definition electro-optic/infrared stabilized imaging systems," *Optical Engineering*, vol. 52, no. 6, Article ID 061310, 2013.

[8] W. W. Zhang, *System Identification and Modeling on the Combustion System of Industry Boiler*, Shanghai Jiao Tong University, Shanghai, China, 2007.

[9] J. G. Fu, *Design and Research of Inertial Platform Stabilization Loop*, Harbin Engineering University, Harbin, China, 2005.

[10] X. Song, H. Chen, and Y. Xue, "Stabilization precision control methods of photoelectric aim-stabilized system," *Optics Communications*, vol. 351, Article ID 20107, pp. 115–120, 2015.

[11] B. Yang and J. Wang, "Hybrid control based on improved CMAC for motor-driven loading system," *Hangkong Xuebao/Acta Aeronautica et Astronautica Sinica*, vol. 29, no. 5, pp. 1314–1318, 2008.

[12] W. Ji, Q. Li, and B. Xu, "Design study of adaptive fuzzy PID controller for LOS stabilized system," in *Proceedings of the 6th International Conference on Intelligent Systems Design and Applications (ISDA '06)*, pp. 336–341, Washington, DC, USA, October 2006.

[13] Y. Han, Y. Lu, and H. Qiu, "An improved control scheme of gyro stabilization electro-optical platform," in *Proceedings of the IEEE International Conference on Control and Automation (ICCA '07)*, pp. 346–351, Guangzhou, China, June 2007.

[14] J. X. Zhao, G. Su, and L. Y. Hao, "Modeling and simulation of strapdown stabilization platform," *Fire and Command Control*, vol. 33, no. 4, pp. 101–106, 2008.

[15] G. Abitova, M. Beisenbi, and V. Nikulin, "Design of a control system with high robust stability characteristics," in *Proceedings of the IEEE 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT '11)*, pp. 1–5, Budapest, Hungary, 2011.

[16] Z. Y. Tang, Z. C. Pei, and B. L. Wu, "Research on image-stabilizing system based on gyro-stabilized platform with reflector," in *Proceedings of the 7th International Symposium on Instrumentation and Control Technology: Measurement Theory and Systems and Aeronautical Equipment*, pp. 1156–1168, October 2008.

[17] G. N. Zhang, Z. G. Liu, S. L. Yao, Y. C. Liao, and C. Xiang, "Suppression of low-frequency oscillation in traction network of high-speed railway based on auto-disturbance rejection control," *IEEE Transactions on Transportation Electrification*, vol. 2, no. 2, pp. 244–255, 2016.

[18] X. Zhou, Y. Jia, Q. Zhao, and R. Yu, "Experimental validation of a compound control scheme for a two-axis inertially stabilized platform with multi-sensors in an unmanned helicopter-based airborne power line inspection system," *Sensors*, vol. 16, no. 3, pp. 366–381, 2016.

[19] M.-C. Zhu, H. Liu, X. Zhang, and H.-G. Jia, "Adaptive feedforward control for inertially stabilized platform," *Optics and Precision Engineering*, vol. 23, no. 1, pp. 141–148, 2015.

[20] Y. Wang, X. Lu, and S. Wang, "Double close loop electrode regulator system based on active disturbance rejection control technology," in *Proceedings of the 2nd International Symposium on Information Science and Engineering (ISISE '09)*, pp. 565–569, December 2009.

[21] L. Wang, *Research on Tension Control System Based on Model Reference Adaptive Algorithm*, Central South University, Changsha, China, 2008.

[22] J. M. Hilkert and D. A. Hullender, "Adaptive control system techniques applied to inertial stabilization systems," in *Acquisition, Tracking, and Pointing IV*, vol. 1304 of *Proceedings of SPIE*, pp. 190–206, International Society for Optical Engineering, Orlando, Fla, USA, September 1990.

[23] C.-L. Lin and Y.-H. Hsiao, "Adaptive feedforward control for disturbance torque rejection in seeker stabilizing loop," *IEEE Transactions on Control Systems Technology*, vol. 9, no. 1, pp. 108–121, 2001.

[24] A. G. Rodriâguez, R. M. Herrera, V. F. Batlle, and P. P. Sanjuan, "Improving the mechanical design of new staircase wheelchair," *Industrial Robot*, vol. 34, no. 2, pp. 110–115, 2007.

[25] S.-K. Oh, W.-D. Kim, W. Pedrycz, and B.-J. Park, "Polynomial-based radial basis function neural networks (P-RBF NNs) realized with the aid of particle swarm optimization," *Fuzzy Sets and Systems*, vol. 163, no. 1, pp. 54–77, 2011.

[26] R.-E. Precup, M.-B. Rădac, M. L. Tomescu, E. M. Petriu, and S. Preitl, "Stable and convergent iterative feedback tuning of fuzzy

controllers for discrete-time SISO systems," *Expert Systems with Applications*, vol. 40, no. 1, pp. 188–199, 2013.

[27] X. Lei, S. S. Ge, and J. Fang, "Adaptive neural network control of small unmanned aerial rotorcraft," *Journal of Intelligent & Robotic Systems: Theory and Applications*, vol. 75, no. 2, pp. 331–341, 2014.

[28] M. Sawada and K. Itamiya, "A design scheme of model reference adaptive control system with using a smooth parameter projection adaptive law," in *Proceedings of the SICE Annual Conference (SICE '11)*, pp. 1704–1709, Tokyo, Japan, September 2011.

[29] S. L. Xiao, Y. M. Li, and J. G. Liu, "A model reference adaptive PID control for electromagnetic actuated micro-positioning stage," in *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE '12)*, pp. 97–102, Seoul, South Korea, August 2012.

[30] P. Kungwalrut, M. Thumma, V. Tipsuwanporn, A. Numsomran, and P. Boonsrimuang, "Design MRAC PID control for fan and plate process," in *Proceedings of the 50th Annual Conference on Society of Instrument and Control Engineers (SICE '11)*, pp. 2944–2948, September 2011.

[31] H. Ouyang, J. Yue, and Y. Su, "Design and application of PID controllers based on interval computing theory," in *Proceedings of the 2nd Annual Conference on Electrical and Control Engineering (ICECE '11)*, pp. 1505–1510, Yichang, China, September 2011.

[32] J.-C. Fang, Z.-H. Qi, and M.-Y. Zhong, "Feedforward compensation method for three axes inertially stabilized platform imbalance torque," *Journal of Chinese Inertial Technology*, vol. 18, no. 1, pp. 38–43, 2010.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 8567429, 10 pages
http://dx.doi.org/10.1155/2016/8567429

Research Article

Sliding Mode Control for Bearingless Induction Motor Based on a Novel Load Torque Observer

Zebin Yang,[1] Ling Wan,[1] Xiaodong Sun,[2] Lin Chen,[1] and Zheng Chen[1]

[1]*School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China*
[2]*Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China*

Correspondence should be addressed to Xiaodong Sun; xdsun@ujs.edu.cn

Received 11 June 2016; Revised 26 August 2016; Accepted 6 September 2016

Academic Editor: Rafael Morales

For the problem of low control performance of Bearingless Induction Motor (BIM) control system in the presence of large load disturbance, a novel load torque sliding mode observer is proposed on the basis of establishing sliding mode speed control system. The load observer chooses the speed and load torque of the BIM control system as the observed objects, uses the speed error to design the integral sliding mode surface, and adds the low-pass filter to reduce the torque observation error. Meanwhile, the output of the load torque is used as the feedforward compensation for the control system, which can provide the required current for load changes and reduce the adverse influence of disturbance on system performance. Besides, considering that the load changes lead to the varying rotational inertia, the integral identification method is adopted to identify the rotational inertia of BIM, and the rotational inertia can be updated to the load observer in real time. The simulation and experiment results all show that the proposed method can track load torque accurately, improve the ability to resist disturbances, and ameliorate the operation quality of BIM control system. The chattering of sliding mode also is suppressed effectively.

1. Introduction

Based on the similarity principles of magnetic bearing and alternating current (AC) motor stator structure, BIM is formed. Two sets of windings are embedded in the stator slot of BIM, which can separately produce electromagnetic torque and radial levitation force. BIM achieves the integration of rapid rotation and stable suspension of rotor by changing the currents in the windings and avoids the mechanical bearing friction, wear and tear, and lubrication. It breaks the bottleneck of traditional asynchronous motor developing towards the higher precision and higher speed direction [1–5]. BIM has many better advantages than the traditional asynchronous motor, such as simple structure, uniform air gap, high mechanical strength, high speed, and ultrahigh speed running in the corrosion or other special environments. Therefore, it shows broad development prospect in medical equipment, transportation, national defense, and so forth [6–9]. However, BIM has the characteristics of nonlinearity, multivariability, and strong coupling. The traditional PI controller cannot acquire high-performance control for BIM when the control system is disturbed by load torque [10].

Sliding mode variable structure control, as a kind of special nonlinear control, can operate in accordance with the trajectory designed by people and purposefully adjust operation according to the system status, which can gain excellent control performance. Due to the fact that the sliding mode control not only can be set by people, but also does not need high precision mathematical model and has strong robustness to disturbances, it is becoming a hot research topic [11–15], and it is gradually applied in the AC servo system. In [16], a new reaching law was designed to improve the operation quality of sliding mode. At the same time, it was applied in the speed control, which effectively enhanced the robustness of permanent magnet synchronous motor (PMSM) system. In [17], the sliding mode control combining with model reference adaptive was used to obtain the speed. The results showed that it increased the estimation precision of rotor velocity for PMSM and decreased the chattering. In [18], the sliding mode control was used in a generator based

on the exercise equipment with nonlinear P-V characteristic curves. The amount of generator input current harmonic is greatly reduced. In [19], the conventional sliding mode control was united with the adaptive fuzzy backstepping scheme. The simulation proved that this method improved the performance of mismatched uncertain system. In [20], the sliding mode control dealt with the difficult problem of obtaining the counterelectromotive force, and it finally implemented the direct torque control of brushless direct current motor. In [21], the sliding mode control was used to detect the speed and position for PMSM. The experimental results proved the validity of the proposed sliding mode observer. In [22], based on the nonsingular terminal sliding mode algorithm and backstepping method, the sliding mode observer and position controller were put forward, which can estimate the torque accurately and track the position quickly. In [23], the adaptive sliding mode control for uncertain singularly perturbed nonlinear system was designed. It not only reduced the effects of uncertainty, but also guaranteed the control performance. In [24, 25], the load sliding mode observers were proposed. They diminished the adverse effects of load changes on PMSM and improved the antidisturbance ability of controlled system at some level. However, they all ignored the problem that the load changes result in the different rotational inertia and the controlled system had large chattering. Hence, the system cannot achieve the best dynamic performance.

A novel sliding mode observer of load torque, of which the state variables are the speed and load torque, is proposed to suppress the impacts of the load torque changes on BIM control system. A low-pass filter used in the observer reduces the observation error of torque. Moreover, the observer as feedforward compensation for the given current alleviates the output pressure of sliding mode controller (SMC). In addition, adopting the integral identification method validly identifies the rotational inertia and improves the precision of BIM. The simulation and experimental results show that the proposed method overcomes the disadvantageous effects on the speed regulation system generated by load disturbances and strengthens the antidisturbance ability of the system.

2. The Dynamics Model of BIM

According to the electromagnetic field theory, the radial levitation force of BIM in the d-q coordinates can be established as [6]

$$\begin{aligned} F_x &= K\left(\psi_{1d} i_{s2d} + \psi_{1q} i_{s2q}\right), \\ F_y &= K\left(\psi_{1d} i_{s2q} - \psi_{1q} i_{s2d}\right), \end{aligned} \tag{1}$$

where $K = K_m + K_l$, $K_m = \pi P_1 P_2 L_{m1}/18lr\mu_0 N_1 N_2$, and $K_l = P_1 N_2/2rN_1$; F_x and F_y are the components of the radial levitation force in x and y directions; the subscript "1" represents the torque windings, the subscript "2" represents the radial levitation force windings, "s" represents the stator, and "r" represents the rotor; P_1 and P_2 separately represent the pole pairs of torque windings and suspension windings; i_{s2d} and i_{s2q} are the current components of the stator in levitation force windings under the d-q axis; L_{m1} is mutual inductance of the levitation force windings; l is the effective length of the rotor; r is the stator inner diameter; μ_0 is the permeability of vacuum; N_1 and N_2, respectively, show the effective number of turns of the torque windings and the levitation force windings; and ψ_{1d} and ψ_{1q} are the components of flux linkage for the torque winding in the d-q coordinates, respectively.

With the torque windings and the levitation force windings, BIM is a nonlinear, strongly coupled, and complex system. In order to simplify the analysis of BIM, a hypothesis is given that the levitation force windings only create a rotating magnetic field. The rotor voltage equation can be described as

$$\begin{aligned} u_{r1d} &= R_{r1} i_{r1d} + p\psi_{r1d} - \left(\psi_{1q} + L_{r1l} i_{r1q}\right)(\omega_1 - \omega_r) \\ &= 0, \\ u_{r1q} &= R_{r1} i_{r1q} + p\psi_{r1q} - \left(\psi_{1d} + L_{r1l} i_{r1d}\right)(\omega_1 - \omega_r) \\ &= 0, \end{aligned} \tag{2}$$

where u_{r1d} and u_{r1q} are the rotor voltages of torque windings in d-q coordinates; R_{r1} is the rotor resistance; ω_1 and ω_r are separately the air gap field speed and rotor speed; and p is the differential operator.

The flux linkage can be expressed as

$$\begin{aligned} \psi_{1d} &= \left(i_{s1d} + i_{r1d}\right) L_{m1}, \\ \psi_{1q} &= \left(i_{s1q} + i_{r1q}\right) L_{m1}, \\ \psi_{s1d} &= \psi_{1d} + i_{s1d} L_{s1l}, \\ \psi_{s1q} &= \psi_{1q} + i_{s1q} L_{s1l}, \\ \psi_{r1d} &= \psi_{1d} + i_{r1d} L_{r1l}, \\ \psi_{r1q} &= \psi_{1q} + i_{r1q} L_{r1l}, \end{aligned} \tag{3}$$

where L_{s1l} and L_{r1l} are the stator leakage inductance and rotor leakage inductance of torque windings, respectively.

The electromagnetic torque equation is set up as

$$T_e = P_1\left(\psi_{1d} i_{s1q} - \psi_{1q} i_{s1d}\right). \tag{4}$$

The equation of motion is written as

$$T_e = T_L + \frac{J}{P_1} p\omega_r, \tag{5}$$

where T_L is the load torque and J is the rotational inertia.

After coordinate transforming, the rotor flux in d-q axis can be expressed as

$$\begin{aligned} \psi_{r1d} &= L_{m1} i_{s1d} + L_{r1} i_{r1d}, \\ \psi_{r1q} &= L_{m1} i_{s1q} + L_{r1} i_{r1q}. \end{aligned} \tag{6}$$

Making the axis of the rotating coordinates d coincide with the rotor flux linkage of torque windings, it is written as $\psi_{r1d} = \psi_{r1}$. Formula (6) can be simplified as

$$\begin{aligned} i_{r1d} &= \frac{\psi_{r1} - L_{m1} i_{s1d}}{L_{r1}}, \\ i_{r1q} &= -\frac{L_{m1}}{L_{r1}} i_{s1q}. \end{aligned} \tag{7}$$

Putting Formula (7) into Formula (2), the excitation current i_{s1d} and slip speed ω_s can be obtained as follows:

$$\begin{aligned} i_{s1d} &= \frac{T_{r1} p + 1}{L_{m1}} \psi_{r1}, \\ \omega_s &= \frac{L_{m1}}{T_{r1} \psi_{r1}} i_{s1q}, \end{aligned} \tag{8}$$

where $\omega_s = \omega_1 - \omega_r$ and $T_{r1} = L_{r1}/R_{r1}$ is the time constant of rotor.

The electromagnetic torque equation turns into

$$T_e = P_1 \frac{L_{m1}}{L_{r1}} i_{s1q} \psi_{r1}, \tag{9}$$

where L_{r1} is the rotor self-induction. Figure 1 is the block diagram of rotor field-oriented decoupling control.

3. The Speed SMC of BIM

The system state variables are defined as

$$\begin{aligned} e_{\omega 1} &= \omega^* - \omega, \\ e_{\omega 2} &= e'_{\omega 1} = -\omega', \end{aligned} \tag{10}$$

where ω^* is the given speed and ω is the actual speed.

After combining with (5), Formula (10) is described as

$$\begin{aligned} e'_{\omega 1} = e_{\omega 2} &= -\frac{P_1^2 \psi_{r1}}{J} \frac{L_{m1}}{L_{r1}} i_{s1q} + \frac{P_1}{J} T_L, \\ e'_{\omega 2} = e''_{\omega 1} &= -\frac{P_1^2 \psi_{r1}}{J} \frac{L_{m1}}{L_{r1}} i'_{s1q}. \end{aligned} \tag{11}$$

The sliding mode surface is chosen as

$$s = e_{\omega 1} + c_1 e_{\omega 2}. \tag{12}$$

This paper chooses the reaching law [16] to weaken the inherent chattering:

$$\begin{aligned} \frac{ds}{dt} &= -\varepsilon |X|^2 \operatorname{sgn}(s) - k |X|^2 s, \\ \lim_{t \to \infty} |X| &= 0, \ a \geq 0, \ b \geq 0, \ \varepsilon > 0, \ k > 0. \end{aligned} \tag{13}$$

According to the Lyapunov stability theory, the existence condition of generalized sliding mode is written as [12]

$$V = \frac{1}{2} s^2 < 0. \tag{14}$$

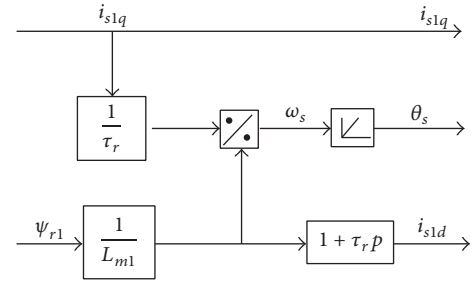


Figure 1: Rotor field-oriented decoupling control.

Differentiating (14) with respect to time, it becomes (15) by substituting (13):

$$\begin{aligned} \dot{V} = s\dot{s} &= s\left(-\varepsilon |X|^2 \operatorname{sgn}(s) - k |X|^2 s\right) \\ &= -\varepsilon |X|^2 |s| - k |X|^2 s^2 < 0. \end{aligned} \tag{15}$$

Formula (15) always stands up. Hence, the system can arrive at sliding mode surface in limited time.

Differentiating $s = e_{\omega 1} + c_1 e_{\omega 2}$ with respect to time, it can be gained as

$$s' = e'_{\omega 1} + c_1 e'_{\omega 2} = e_{\omega 2} - \frac{c_1 P_1^2 \psi_{r1} L_{m1}}{J L_{r1}} i'_{s1q}. \tag{16}$$

Combining Formula (13) with Formula (16) gives the following formula:

$$e_{\omega 2} - \frac{c_1 P_1^2 \psi_{r1} L_{m1}}{J L_{r1}} i'_{s1q} = -\varepsilon |X|^2 \operatorname{sgn}(s) - k |X|^2 s. \tag{17}$$

Choosing $X = e_{\omega 1}$ to avoid the differential interference in $e_{\omega 2}$, the sliding mode controller is designed as

$$\begin{aligned} i_{s1q} = &\frac{J L_{r1}}{P_1^2 \psi_{r1} L_{m1} c_1} \\ &\cdot \int \left(\varepsilon |e_{\omega 1}|^2 \operatorname{sat}(s) + k |e_{\omega 1}|^2 s + e_{\omega 2}\right) dt. \end{aligned} \tag{18}$$

From (18), it can be seen that the current can eliminate steady-state error and improve the accuracy of system.

4. The Design of Novel Load Torque Sliding Mode Observer

4.1. The Design of Load Torque Sliding Mode Observer. Considering the high switch frequency of the controller, the load torque can be deemed to be a constant value in a control cycle. Considering the load torque as an extension, the state equation of BIM can be expressed as

$$\begin{aligned} \frac{d\omega}{dt} &= \frac{P_1^2 \psi_{r1}}{J} \frac{L_{m1}}{L_{r1}} i_{s1q} - \frac{P_1}{J} T_L, \\ \frac{dT_L}{dt} &= 0. \end{aligned} \tag{19}$$

Based on the equation above, the extended load torque observer is written as

$$\frac{d\hat{\omega}}{dt} = \frac{P_1^2\psi_{r1}}{J}\frac{L_{m1}}{L_{r1}}i_{s1q} - \frac{P_1}{J}\hat{T}_L + V,$$
$$\frac{d\hat{T}_L}{dt} = \eta V, \tag{20}$$

where $V = \gamma \operatorname{sgn}(\omega - \hat{\omega})$; γ is the sliding mode gain; η is the feedback gain; and $\hat{\omega}$ and $\hat{T}_L$ are the estimations of electrical angular velocity and load torque, respectively.

The estimation errors of speed and the load are defined as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \omega - \hat{\omega} \\ T_L - \hat{T}_L \end{bmatrix}. \tag{21}$$

After Formula (19) subtracts Formula (20), the observation errors of sliding mode are obtained as

$$\frac{dx_1}{dt} = -\frac{P_1}{J}x_2 - V,$$
$$\frac{dx_2}{dt} = -\eta V. \tag{22}$$

Because the torque change is expressed in the form of speed finally, the designed sliding mode surface consists of the state variable $x_1 = \omega - \hat{\omega}$. The sliding mode surface is established as (23) to reduce the system overshoot:

$$s = x_1 + c\int x_1 dt. \tag{23}$$

Differentiating the sliding mode surface and combining with (13) and (5), it can be acquired as

$$\hat{T}_L = \frac{J}{P_1}\left[\gamma \operatorname{sgn}(\omega - \hat{\omega}) - c(\omega - \hat{\omega}) - k|\omega - \hat{\omega}|^2 s - \varepsilon|\omega - \hat{\omega}|^2 \operatorname{sgn}(s)\right] + T_L(0). \tag{24}$$

Because the system load torque T_L is an unknown variable, $T_L(0)$ is recorded as the estimation of load torque at time zero.

In order to decrease the chattering in the sliding mode, the sign function sgn (s) is replaced by saturation function sat (s, Δ) [25]:

$$\operatorname{sat}(s,\Delta) = \begin{cases} 1 & s_i(x) > -\Delta \\ \dfrac{s_i(x)}{\Delta} & -\Delta < s_i(x) < \Delta \\ -1 & s_i(x) < -\Delta. \end{cases} \tag{25}$$

Formula (24) can be expressed as

$$\hat{T}_L = \frac{J}{P_1}\left[\gamma \operatorname{sat}(\omega - \hat{\omega}) - c(\omega - \hat{\omega}) - k|\omega - \hat{\omega}|^2 s - \varepsilon|\omega - \hat{\omega}|^2 \operatorname{sat}(s,\Delta)\right] + T_L(0). \tag{26}$$

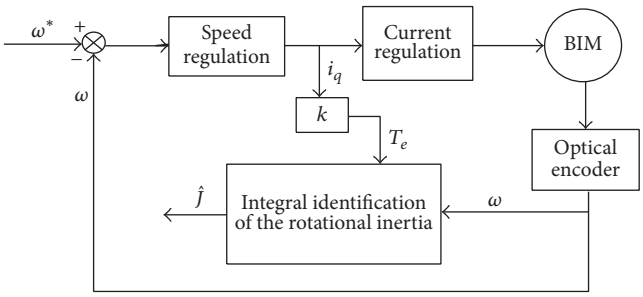


Figure 2: Diagram of inertia identification.

For the purpose of improving the observation precision of sliding mode observer, a low-pass filter [7] shown in (27) is added to the observer:

$$\tilde{T}_L = \frac{\omega_c}{s + \omega_c}\hat{T}_L. \tag{27}$$

The output of load torque observer is as the feedforward disturbance compensation i_{q2} and the given current i_q^* is described as

$$i_q^* = i_{s1q} + i_{q2} = \frac{JL_{r1}}{P_1^2\psi_{r1}L_{m1}c_1}\int\left(\varepsilon|e_{\omega1}|^2 \operatorname{sat}(s) + k|e_{\omega1}|^2 s + e_{\omega2}\right)dt + \frac{\tilde{T}_L}{\partial}, \tag{28}$$

where $\partial > 0$ is the feedforward gain of torque observation.

From (18) and (28), it can be known that ε and k in Formula (18) need to be large enough for meeting the load disturbance, while large ε and k increase the amplitude of discrete magnitude and result in big chattering. However, in (28), the observed disturbance is used to provide the required current for disturbance changes and needs no large ε and k. As a consequence, the feedforward compensation scheme of load torque reduces the gain amplitude of sliding mode and lowers the negative impacts on the control system caused by the disturbances.

4.2. The Load Torque Sliding Mode Observer with the Rotational Inertia Online Identification. When a sliding mode observer is designed, the rotational inertia of system is usually regarded as a known quantity. However, in practice application, the load changes will lead to the inertia changes. Therefore, identifying the inertia online and timely updating it to the observer will greatly improve the overall control performance of BIM system. With the characteristics of high precision and strong robustness, the integral identification algorithm is used to identify the rotational inertia online. The structure diagram of BIM's speed loop based on the integral identification is shown in Figure 2.

Equation (5) is rewritten as

$$T_e = \frac{\hat{J}}{P_1}\frac{d\omega}{dt} + \hat{T}_r, \tag{29}$$

where $\widehat{J}$ is the estimation of rotational inertia and $\widehat{T}_r$ is the collection of disturbances and its specific expression is as follows:

$$\widehat{T}_r(t) = \Delta J \frac{d\omega}{dt} + T_L, \tag{30}$$

where ΔJ is the inertia error, $J - \widehat{J} = \Delta J$.

By the formula above, it can be found that the output of disturbance is in the form of torque. At the same time, observing torque can obtain ΔJ. Using ΔJ and constant recursion correction based on selecting initial value of inertia, the recursion equation can be expressed as

$$\widehat{J}(k) = \widehat{J}(k-1) + \Delta J. \tag{31}$$

Thus, the identification accuracy of inertia depends on ΔJ. In order to obtain high accuracy ΔJ, this paper chooses a periodic speed signal and uses the integral to eliminate the influences of torque disturbances on rotational inertia.

Because of the high sample frequency, the load torque T_L is regarded as $T_L(t) = T_L(t+T)$ and the speed signal ω meets $\omega(t) = \omega(t+T)$. After on both sides of Formula (30) multiplying by $\dot{\omega}(t)$ and integrating it, the equation is written as

$$\begin{aligned}\int_{(k-1)T}^{kT} \widehat{T}_r(t)\,\dot{\omega}(t)\,dt &= \Delta J \int_{(k-1)T}^{kT} \dot{\omega}^2(t)\,dt \\ &\quad + \int_{(k-1)T}^{kT} T_L(t)\,\dot{\omega}(t)\,dt.\end{aligned} \tag{32}$$

Due to the fact that the load torque is a constant within a cycle, the definite integral of the load torque in (32) can be calculated as follows:

$$\begin{aligned}\int_{(k-1)T}^{kT} T_L(t)\,\dot{\omega}(t)\,dt &= T_L\omega(t)\Big|_{(k-1)T}^{kT} \\ &= T_L\omega(kT) - T_L\omega((k-1)T) \\ &= 0.\end{aligned} \tag{33}$$

Hence, (32) can be simplified as

$$\int_{(k-1)T}^{kT} \widehat{T}_r(t)\,\dot{\omega}(t)\,dt = \Delta J \int_{(k-1)T}^{kT} \dot{\omega}^2(t)\,dt. \tag{34}$$

From (34), it can be found that the effects of load disturbance on the inertia have been solved.

Equation (34) can be expressed as

$$\Delta J = \frac{\int_{(k-1)T}^{kT} \widehat{T}_r(t)\,\dot{\omega}(t)\,dt}{\int_{(k-1)T}^{kT} \dot{\omega}^2(t)\,dt}. \tag{35}$$

Combining with (35), the recursive equation of inertia identification is available as

$$\widehat{J}(k) = \widehat{J}(k-1) + \frac{\int_{(k-1)T}^{kT} \widehat{T}_r(t)\,\dot{\omega}(t)\,dt}{\int_{(k-1)T}^{kT} \dot{\omega}^2(t)\,dt}. \tag{36}$$

Table 1: Parameters of the Bearingless Induction Motor (BIM).

Parameters	Torque winding	Suspension winding
Rated power (Kw)	1	0.5
Rated current (A)	2.86	2.86
Stator resistance (Ω)	2.01	1.03
Rotor resistance (Ω)	11.48	0.075
Mutual inductance of stator and rotor (H)	0.15856	0.00932
Stator leakage inductance (H)	0.16310	0.01199
Rotor leakage inductance (H)	0.16778	0.01474
Rotational inertia (kg·m^2)	0.00769	0.00769
Rotor mass (kg)	2.85	2.85
Stator inner diameter (mm)	98	98
Core length (mm)	105	105
Pole pairs	1	2

Inertia identification is realized and can be updated to the load observer automatically.

5. Simulation and Experimental Research

5.1. Results and Analysis of the Simulation. To validate the effectiveness of novel load torque observer with the function of inertia identification online and feedforward compensation scheme for BIM speed regulation system, a simulation mode of control system is constructed. The control block diagram of BIM includes two parts: the rotation part and the suspension part. In rotation part, the SMC outputs the given current i_q^* by inputting the speed error. Combining with the given air gap flux ψ_1^*, the excitation component is received. With the coordinate transformation method, the two-phase excitation current is transformed into the three-phase given current. With the current regulation, the three-phase current is obtained which is used to control the rotation of rotor. In suspension part, the radial levitation force of BIM is output by PID controllers. With ψ_1^*, the current in the levitation windings can be calculated. By the coordinate transformation and current regulation, the required three-phase current is gained. The steady suspension and rapid rotation of rotor are realized finally. The whole control block diagram of BIM is shown in Figure 3 and the specific parameters of BIM are shown in Table 1.

In the simulation, the rotational inertia is set to 0.00769/kg·m^2 and the given speed is $n = 10000$ r/min. Based on the tracking characteristics of integral identification algorithm, a step signal is selected as the given speed whose amplitude is 10000 r/min and the sampling cycle is defined as $T = 0.02$ s. Figure 4 shows the identification waveform of rotational inertia when the system suffers the load disturbances. It can be seen that the rotational inertia J converges to the given value within a sampling period at first. When

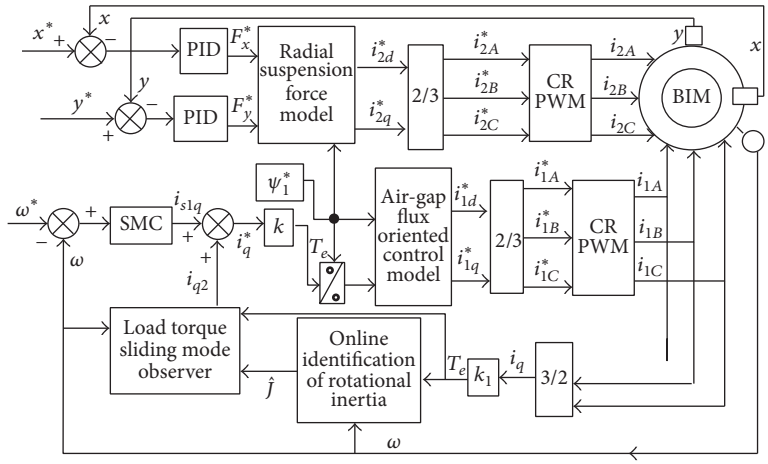


Figure 3: Control system diagram of Bearingless Induction Motor (BIM).

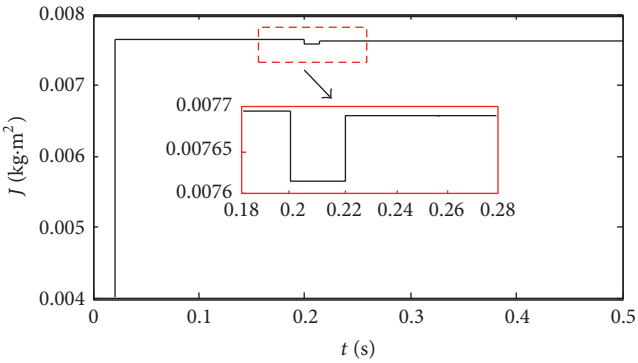


Figure 4: The identification result of rotational inertia.

the system is attacked by the load disturbance (8 N·m), J just has small fluctuation and restores the stabilization quickly. Therefore, the integral identification algorithm has better tracking ability and shows good robustness for disturbance.

Figures 5(a) and 5(b) show the estimation load $\widehat{T}_L$ and the actual load T_L under load mutation. As shown in (a) and (b), a sudden load (8 N·m) is added to the system at $t = 0.4$ s. Then, the load drops to 0 N·m at $t = 0.6$ s. It can be seen from Figure 5 that the sliding mode observer can accurately track the load torque and has strong robustness.

Figure 6 shows the simulations of BIM in the presence of sudden load (8 N·m). Figures 6(a) and 6(b) present the output currents from the SMC and the load torque observer, respectively. Figure 6(c) shows the speed response of BIM under the proposed control strategy in this paper and the conventional SMC. Figure 6(a) shows that the current from SMC just slightly increases and it returns back to the stable value rapidly. Figure 6(b) indicates that the compensation current from load torque observer rises quickly, which provides enough current for disturbance. Figure 6(c) demonstrates that the speed of BIM controlled by the conventional SMC has larger fluctuation and needs more time to operate at the original speed value than the proposed method when it is attacked by load mutation. Based on the above analysis, the following two conclusions can be obtained:

(1) The feedforward compensation scheme of load torque observer can provide the required current for load changes. It can reduce the output pressure of SMC and make the output of SMC almost invariant.

(2) Based on the novel load observer and feedforward compensation strategy, the speed of BIM under big disturbance has no fluctuation and can quickly converge to the original value. The method weakens the system chattering effectively and enhances the stability of system.

Figures 7(a) and 7(b) show the rotor radial displacement at the speed of $n = 10000$ r/min. It can be known that the rotor can arrive at the steady point rapidly with the proposed control strategy in this paper. It achieves the integration of rapid rotation and stable suspension. The system has excellent control performance.

5.2. Results and Analysis of the Experiment. In order to further verify the effectiveness of the proposed control method, an experimental prototype with two degrees of freedom is used to build experimental platform. Due to the limits of photoelectrical encoder measuring speed, the speed is set to 2000 r/min in the experiment. The air gap of motor auxiliary bearing is 0.4 mm. Moreover, the load mutation is carried out to detect antijamming performance of BIM. The experimental results are shown in Figure 8.

Figure 8(a) shows the radial displacement when the system is controlled by the proposed method in this paper. The rotor is running around the equilibrium point and the maximum offset value is far less than the air gap of auxiliary bearing. It indicates that the rotor is suspended steadily under the proposed control strategy. Figures 8(b)–8(e) show the responses of BIM with the load mutation. Figure 8(b) gives the identification results of rotational inertia. In view of the excellent robustness of integral identification algorithm, the

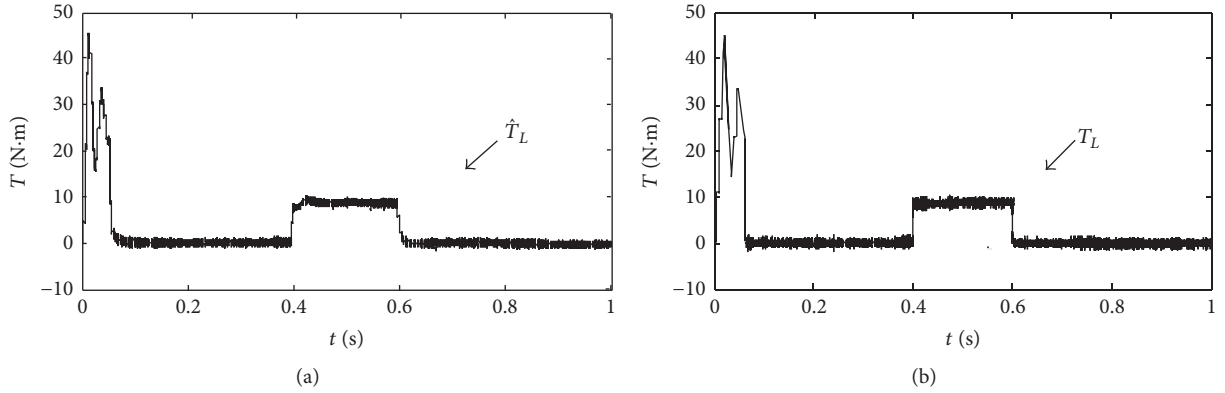


Figure 5: The torque waveforms under load mutation: (a) the waveform of the estimation torque; (b) the waveform of the actual torque.

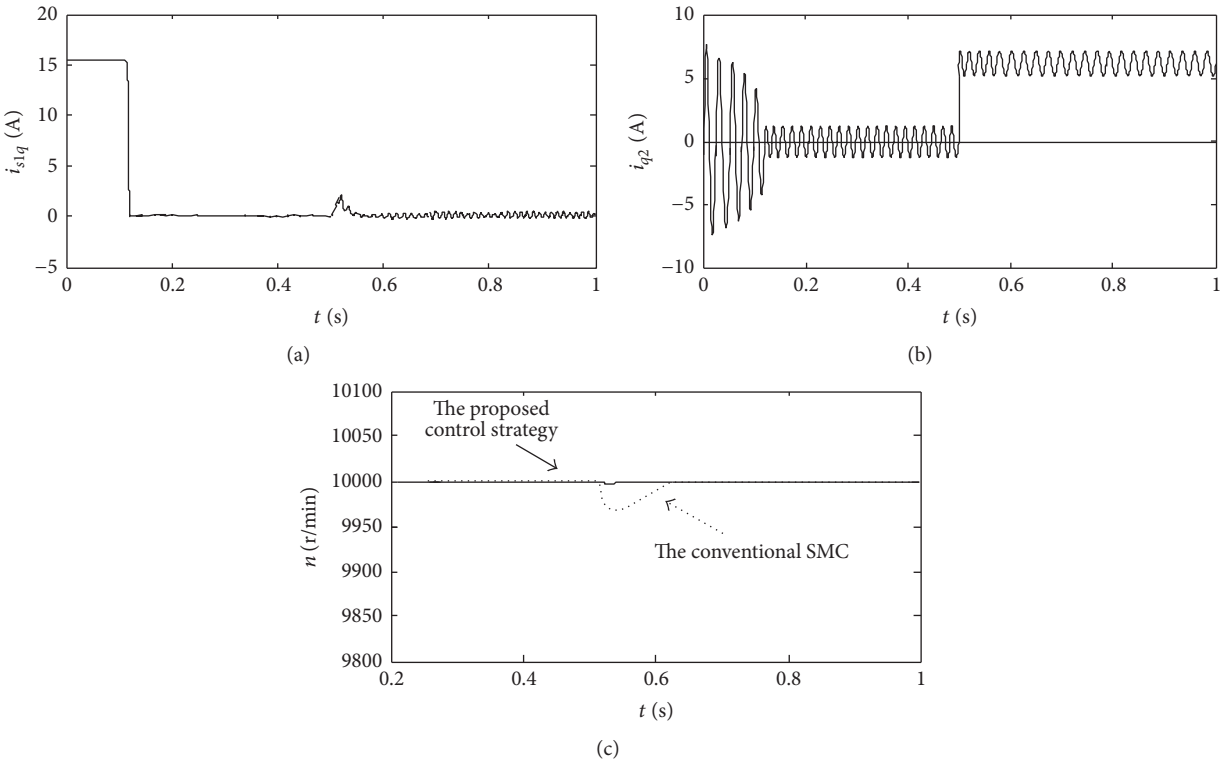


Figure 6: The responses of current and speed under load mutation: (a) the output current of SMC; (b) the compensation current of load torque; and (c) the speed response of BIM under sudden load.

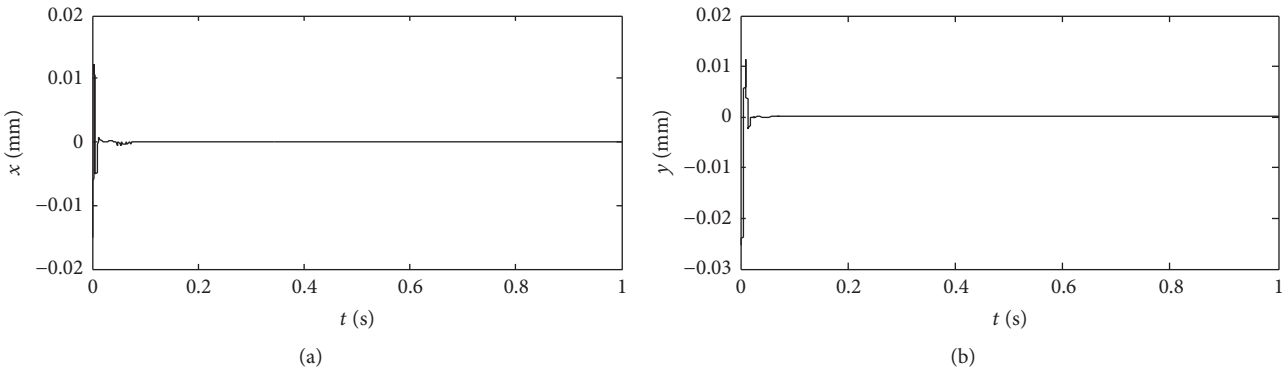


Figure 7: Waveforms of rotor radial displacement: (a) the radial displacement on x-axis; (b) the radial displacement on y-axis.

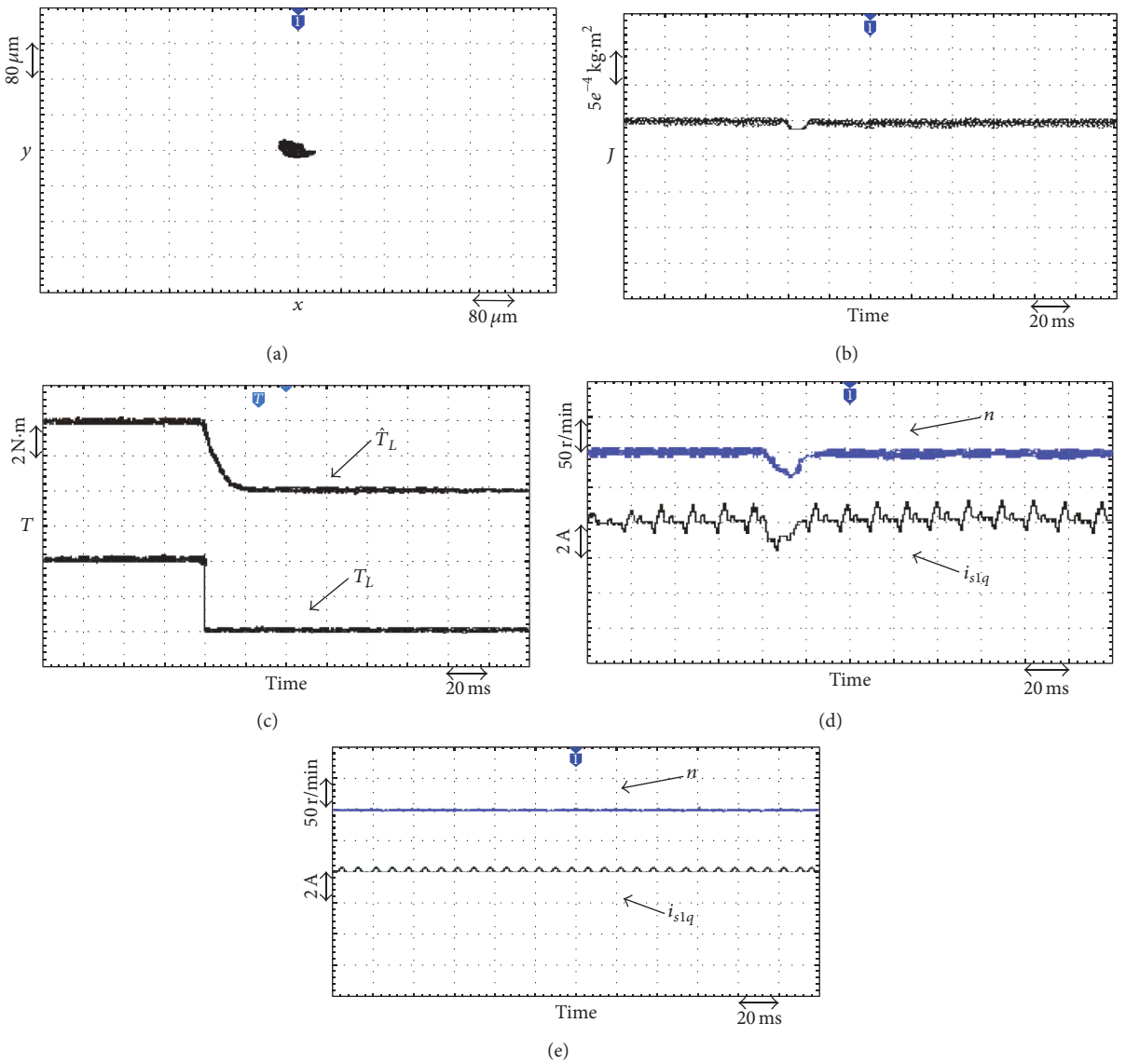


Figure 8: The experimental waveforms underload mutation: (a) the radial displacement of rotor; (b) the identification results of rotational inertia under load mutation; (c) the waveforms of torque; (d) the responses of speed and current under the conventional SMC; and (e) the responses of speed and current under the proposed method.

rotational inertia returns back to the stable value after a slight fluctuation. The observing waveform of load torque is given in Figure 8(c). It can be seen from the waveform that the sliding mode observer tracks the load precisely. Figures 8(d) and 8(e) present the experimental results based on the control strategy of ordinary SMC and the proposed method in this paper separately. From Figure 8(d), it can be found that the current has big undulation. The speed decreases by 40 r/min and needs 20 ms to restore stability. From Figure 8(e), the load mutation does not affect the output of SMC and the speed is smooth. By comparing (d) with (e), it can be found that using the new load torque sliding mode observer and the feedforward compensation scheme can make BIM exactly and stably operate. In addition, it has low sensitivity with respect to disturbances.

6. Conclusions

A novel load torque sliding mode observer was proposed to eliminate the adverse impacts caused by load disturbance in BIM control system. Owning to the application of a low-pass filter in the sliding mode observer, the observation error of load torque is effectively reduced. At the same time, the output of load torque observer as disturbance compensation greatly diminishes the amplitude of discrete quantity and weakens chattering. With the integral identification method, the proposed sliding mode observer can identify the rotational inertia accurately and improve the robustness of rotational inertia for disturbances. The simulation and experimental results all show that the proposed control scheme in this paper effectively improves the dynamic and

static performance of BIM control system, suppresses the system chattering, and enhances the robustness of system.

Competing Interests

The authors declare no conflict of interests.

Authors' Contributions

Zebin Yang and Ling Wan proposed the new idea of sliding mode control for Bearingless Induction Motor based on a novel load torque observer. Ling Wan derived the equations. Xiaodong Sun established the simulation model. Lin Chen was in charge of analyzing the data. Zheng Chen checked the language. All authors were involved in preparing the manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Projects 51475214, 61104016, and 51305170, the China Postdoctoral Science Foundation Funded Project 2015T80508, the Natural Science Foundation of Jiangsu Province of China under Projects BK20130515, BK20141301, and BK20150524, the Professional Research Foundation for Advanced Talents of Jiangsu University under Projects 12JDG057 and 14JDG076, Six Talent Peaks of Jiangsu Province under Projects ZBZZ-017 and 2015-XNYQC-003, and the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions.

References

[1] T. Hiromi, T. Katou, A. Chiba, M. A. Rahman, and T. Fukao, "A novel magnetic suspension-force compensation in bearingless induction-motor drive with squirrel-cage rotor," *IEEE Transactions on Industry Applications*, vol. 43, no. 1, pp. 66–76, 2007.

[2] X. Sun, L. Chen, and Z. Yang, "Overview of bearingless permanent-magnet synchronous motors," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 12, pp. 5528–5538, 2013.

[3] E. F. Rodriguez and J. A. Santisteban, "An improved control system for a split winding bearingless induction motor," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 8, pp. 3401–3408, 2011.

[4] X. Sun, L. Chen, H. Jiang, Z. Yang, J. Chen, and W. Zhang, "High-performance control for a bearingless permanent magnet synchronous motor using neural network inverse scheme plus internal model controllers," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3479–3488, 2016.

[5] A. T. De Almeida, F. J. T. E. Ferreira, and A. Q. Duarte, "Technical and economical considerations on super high-efficiency three-phase motors," *IEEE Transactions on Industry Applications*, vol. 50, no. 2, pp. 1274–1285, 2014.

[6] X. Sun, L. Chen, Z. Yang, and H. Zhu, "Speed-sensorless vector control of a bearingless induction motor with artificial neural network inverse speed observer," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 4, pp. 1357–1366, 2013.

[7] Z. Yang, D. Dong, R. Fan, X. Sun, and R. Jin, "Radial displacement-sensorless control for bearingless induction motor," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 41, no. 8, pp. 1388–1395, 2015.

[8] T. Schuhmann, W. Hofmann, and R. Werner, "Improving operational performance of active magnetic bearings using Kalman filter and state feedback control," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 2, pp. 821–829, 2012.

[9] Z. Yang, D. Dong, H. Gao, X. Sun, R. Fan, and H. Zhu, "Rotor mass eccentricity vibration compensation control in bearingless induction motor," *Advances in Mechanical Engineering*, vol. 7, no. 1, Article ID 168428, 2015.

[10] B.-J. Hou, J.-S. Gao, X.-Q. Li, and Y.-F. Zhou, "Study on repetitive PID control of linear motor in wafer stage of lithography," *Procedia Engineering*, vol. 29, pp. 3863–3867, 2012.

[11] C.-K. Lai and K.-K. Shyu, "A novel motor drive design for incremental motion system via sliding-mode control method," *IEEE Transactions on Industrial Electronics*, vol. 52, no. 2, pp. 499–507, 2005.

[12] Z. Yang, L. Wan, X. Sun, F. Li, and L. Chen, "Sliding mode variable structure control of a bearingless induction motor based on a novel reaching law," *Energies*, vol. 9, no. 6, p. 452, 2016.

[13] K. J. Lin, "Sliding mode control design for uncertain singular systems," *Applied Mechanics & Materials*, vol. 145, no. 8, pp. 16–20, 2011.

[14] S. J. Zhu, Y. P. He, and J. Ren, "Design of vehicle active suspension system using discrete-time sliding mode control with parallel genetic algorithm," in *Proceedings of the ASME International Mechanical Engineering Congress and Exposition (IMECE '13)*, 8 pages, November 2013.

[15] Y. Chu and J. Fei, "Adaptive global sliding mode control for MEMS gyroscope using RBF neural network," *Mathematical Problems in Engineering*, vol. 2015, Article ID 403180, 9 pages, 2015.

[16] X. G. Zhang, K. Zhao, L. Sun, and Q. An, "Sliding mode control of permanent magnet synchronous motor based on a novel exponential reaching law," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 31, no. 15, pp. 47–52, 2011.

[17] R. Li, G. Y. Zhao, and S. J. Xu, "Sensorless control of permanent magnet synchronous motor based on extended sliding mode observer," *Transactions of China Electrotechnical Society*, vol. 3, pp. 79–85, 2012.

[18] M.-H. Wang, M.-L. Huang, and W.-J. Jiang, "Maximum power point tracking and harmonic reducing control method for generator-based exercise equipment," *Energies*, vol. 9, no. 2, article 103, pp. 1–13, 2016.

[19] H. Q. Hou, Q. Miao, Q. H. Gao, and B. T. Aorue, "Fuzzy backstepping sliding mode control for mismatched uncertain system," *Journal of Engineering Science and Technology Review*, vol. 7, no. 2, pp. 175–181, 2014.

[20] J. J. Zhu, M. Su, X. Z. Wang, and L. Ma, "Direct-torque-control of brushless DC motors based on segmented sliding-mode-variable-structure," *Chinese Journal of Scientific Instrument*, vol. 34, no. 11, pp. 2634–2640, 2013.

[21] Z.-C. Qiu, J.-L. Guo, B. Wang, and J. Xiao, "Deadbeat predictive current control based on a sliding mode observer with Kalman filter for PMSM speed and rotor position," *Electric Machines and Control*, vol. 18, no. 4, pp. 60–65, 2014 (Chinese).

[22] Y.-M. Fang, Z. Li, Y.-Y. Wu, and X. Yu, "Backstepping control of PMSM position systems based on terminal-sliding-mode load observer," *Electric Machines and Control*, vol. 18, no. 9, pp. 105–111, 2014 (Chinese).

[23] K.-J. Lin, "Adaptive sliding mode control design for a class of uncertain singularly perturbed nonlinear systems," *International Journal of Control*, vol. 87, no. 2, pp. 432–439, 2014.

[24] X. G. Zhang, L. Sun, and K. Zhao, "Sliding mode control of PMSM based on a novel load torque sliding mode observer," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 32, no. 3, pp. 111–116, 2012 (Chinese).

[25] Z. Li, G. D. Hu, J. R. Cui, and G. Cui, "Sliding-mode variable structure control with integral action for permanent magnet synchronous motor," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 34, no. 3, pp. 431–437, 2014.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 1572303, 7 pages
http://dx.doi.org/10.1155/2016/1572303

Research Article

Adaptive Fuzzy Sliding Mode Control of MEMS Gyroscope with Finite Time Convergence

Jianxin Ren, Rui Zhang, and Bin Xu

School of Automation, Northwestern Polytechnical University, Xi'an 710129, China

Correspondence should be addressed to Bin Xu; smileface.binxu@gmail.com

Received 25 April 2016; Accepted 18 July 2016

Academic Editor: José A. Somolinos

This paper presents adaptive fuzzy finite time sliding mode control of microelectromechanical system gyroscope with uncertainty and external disturbance. Firstly, fuzzy system is employed to approximate the uncertainty nonlinear dynamics. Secondly, nonlinear sliding mode hypersurface and double exponential reaching law are selected to design the finite time convergent sliding mode controller. Thirdly, based on Lyapunov methods, adaptive laws are presented to adjust the fuzzy weights and the system can be guaranteed to be stable. Finally, the effectiveness of the proposed method is verified with simulation.

1. Introduction

MEMS gyroscopes have become the most growing microsensors in recent years due to the characteristics of compact size, low cost, and high sensitivity. Most MEMS gyroscopes sales in the market are vibrating silicon micromechanical gyroscopes, whose basic principle is to generate and detect Coriolis Effect. As depicted in Figure 1, under assumption that the proof mass m of gyroscope rotates around z-axis at a speed of $\vec{\Omega}$ and makes uniform motion along the x-axis at a speed of $\vec{v}$, a Coriolis force of $\vec{F} = -2m\vec{\Omega} \times \vec{v}$ is produced along y-axis.

In the last few years, numerous advanced control approaches with intelligent design have been studied to realize the trajectory tracking in [1–4] and to handle the system parametric uncertainties and disturbances, and the adaptive control can be found in [5–7]. For control of MEMS gyroscope, Park and Horowitz firstly applied adaptive state feedback control method [8]. Both drive shaft and sensitive axis were subjected to feedback control force in this control method, which administered two axial modal vibration track specified reference trajectories, weakening the boundary between drive mode and test mode as well.

Sliding mode control changes its structure to force the system in accordance with a predetermined trajectory. Batur et al. developed a sliding mode control for MEMS gyroscope system in [9]. Since then, adaptive sliding mode control approach with the advantages of variable structure methods and adaptive control strategies are presented to control MEMS gyroscopes in [10, 11].

Due to the necessity of ideal sliding mode, good dynamic quality, and high robustness, several methods are extended to improve the performance. Yu and Man investigated a nonlinear sliding mode hypersurface to ensure that systems from any point of the sliding mode surface were able to reach the balance point in a limited time in [12–16]. Bartoszewicz [17] examined the reaching laws introduced by Gao and Hung in [18] and proposed an enhanced version of those reaching laws, which was more appropriate for systems subject to constraints. Recently, Fallaha et al. studied a novel approach, which allowed chattering reduction on control input while keeping tracking performance in steady-state regime [19]. This approach consisted of designing a nonlinear reaching law by using an exponential function that dynamically adapted to the variations of the controlled system. Mei and Wang in [20] proposed a nonlinear sliding mode surface which converged to the equilibrium point with a higher speed than both linear sliding mode surface and terminal sliding mode surface. In addition, a new two-power reaching law was proposed to make the system move toward the sliding mode faster.

As a matter of fact, the methods mentioned above are highly dependent on the structure of the nonlinearity, while,

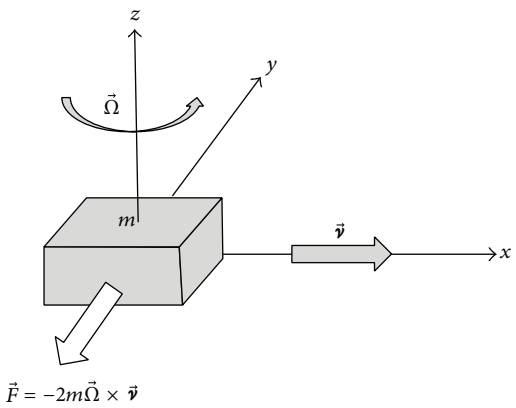


Figure 1: Coriolis Effect.

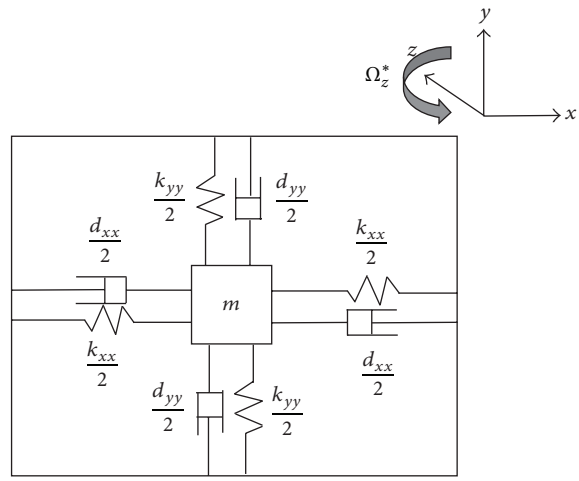


Figure 2: The basic principle diagram of z-axis vibratory MEMS gyroscope.

currently, accurate model is unavailable. Thus, fuzzy model has been widely used to approximate nonlinear objects in [21, 22]. Robust adaptive sliding mode control with on-line identification for the upper bounds of external disturbance and estimator for the nonlinear dynamics of MEMS gyroscope uncertainty parameters was proposed in [23].

In this paper, an adaptive fuzzy sliding mode control strategy with nonlinear sliding mode hypersurface and double exponential reaching law is developed to track MEMS gyroscope. Furthermore, it converges faster compared with strategies using conventional sliding mode surface in [23] and terminal sliding mode surface in [12–16].

The rest of this paper is organized as follows. The dynamics of MEMS gyroscope with parametric uncertainties and disturbances are given in Section 2. Controller design and stability analysis are discussed in Section 3. Numerical simulations are conducted to verify the superiority of the proposed approach in Section 4, compared with conventional adaptive fuzzy sliding mode control. Conclusions are drawn in Section 5.

2. Dynamics of MEMS Gyroscope

The basic principle of z-axis vibratory MEMS gyroscope is shown in Figure 2, which can be described as a quality-stiffness-damping system. Owing to mechanical coupling caused by fabrication imperfections, the dynamics can be derived as

$$\begin{aligned} &m\ddot{x} + d_{xx}\dot{x} + \left(d_{xy} - 2m\Omega_z^*\right)\dot{y} + \left(k_{xx} - m\Omega_z^{*^2}\right)x \\ &\quad + k_{xy}y = u_x^*, \\ &m\ddot{y} + d_{xx}\dot{y} + \left(d_{xy} + 2m\Omega_z^*\right)\dot{x} + \left(k_{yy} - m\Omega_z^{*^2}\right)y \\ &\quad + k_{xy}x = u_y^*, \end{aligned} \tag{1}$$

where m is the mass of proof mass; Ω_z^* is the input angular velocity; x, y represent the system generalized coordinates; d_{xx}, d_{yy} represent damping terms; d_{xy} represents asymmetric damping term; k_{xx}, k_{yy} represent spring terms; k_{xy} represents asymmetric spring terms; and u_x^*, u_y^* represent the control forces.

On issues related to the study of mechanism, the law described by model is required to be independent of dimensions. So, it is necessary to establish nondimensional vector dynamics. Because of the nondimensional time $t^* = \omega_o t$, both sides of (1) should be divided by reference frequency ω_o^2, reference length q_o, and reference mass m. Then the dynamics can be rewritten in vector forms:

$$\begin{aligned} &\frac{q^*}{q_o} + \frac{D^*}{m\omega_o}\frac{q^*}{q_o} + 2\frac{S^*}{\omega_o}\frac{q^*}{q_o} - \frac{\Omega_z^{*^2}}{\omega_o^2}\frac{q^*}{q_o} + \frac{K_1^*}{m\omega_o^2}\frac{q^*}{q_o} \\ &\quad = \frac{u^*}{m\omega_o^2 q_o}, \end{aligned} \tag{2}$$

where $q^* = \begin{bmatrix} x \\ y \end{bmatrix}$, $u^* = \begin{bmatrix} u_x^* \\ u_y^* \end{bmatrix}$, $D^* = \begin{bmatrix} d_{xx} & d_{xy} \\ d_{xy} & d_{yy} \end{bmatrix}$, $S^* = \begin{bmatrix} 0 & -\Omega_z^* \\ \Omega_z^* & 0 \end{bmatrix}$, $K_1^* = \begin{bmatrix} k_{xx} & k_{xy} \\ k_{xy} & k_{yy} \end{bmatrix}$.

New parameters are defined as follows:

$$\begin{aligned} q &= \frac{q^*}{q_o}, \\ u &= \frac{u^*}{m\omega_o^2 q_o}, \\ \Omega_z &= \frac{\Omega_z^*}{\omega_o}, \\ D &= \frac{D^*}{m\omega_o}, \\ K_1 &= \frac{K_1^*}{m\omega_o^2}, \\ S &= -\frac{S^*}{\omega_o}. \end{aligned} \tag{3}$$

Thus, the final form of the nondimensional vector dynamics is

$$\ddot{q} = (2S - D)\,\dot{q} + \left(\Omega_z^2 - K_1\right) q + u. \tag{4}$$

In presence of parametric uncertainties and external disturbance, based on (4), state equation of dynamics is established as

$$\ddot{q} = (A + \Delta A)\,\dot{q} + (B + \Delta B)\,q + Cu + d\,(t)\,, \tag{5}$$

where $A \in R^{2\times2}$, $B \in R^{2\times2}$, $C \in R^{2\times2}$ are system known matrices; ΔA, ΔB are parametric uncertainties; and $d(t)$ is an external disturbance. Besides, $A = 2S - D$, $B = \Omega_z^2 - K_1$.

If the system total interference (consisting of parametric uncertainties and external disturbance) is represented by $P(t)$, we know

$$\ddot{q} = A\dot{q} + Bq + Cu + P\left(\dot{q}, q, t\right), \tag{6}$$

where $P(\dot{q}, q, t) = \Delta A\dot{q} + \Delta Bq + d(t)$.

It is vital that (6) must meet the following assumptions.

Assumption 1. The total interference $\|P(\dot{q}, q, t)\| \le P_c$, where P_c is an unknown positive vector.

Assumption 2. The total interference $P(\dot{q}, q, t)$ meets sliding mode matching conditions; namely, $\Delta A = CH_1$, $\Delta B = CH_2$, $d(t) = CH_3$, where H_1, H_2, H_3 are unknown matrices with appropriate dimensions.

Assumption 3. A, B are observability matrices.

Based on the above assumptions, the controller can be designed to compensate the total interference.

3. Adaptive Fuzzy Finite Time Sliding Mode Control

The fuzzy model of $P(t)$ could be composed of M IF-THEN rules, and the ith rule has the form

Rule i: *IF* $\dot{x}_i$ *is* A_{1i} *and* $\dot{y}_i$ *is* A_{2i} *and* x_i *is* A_{3i} *and* y_i *is* A_{4i}

$$\textit{THEN } \widehat{P}(\dot{q}, q \mid \boldsymbol{\theta}_P)_i \textit{ is } B_i,\ i = 1, 2, \ldots, M.$$

Based on singleton fuzzifier, product inference, and center-average defuzzifier, its output can be expressed as

$$\widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right) = \boldsymbol{\theta}_P^T \boldsymbol{\mu}\left(\dot{q}, q\right), \tag{7}$$

where $\boldsymbol{\mu}(\dot{q}, q) = (\boldsymbol{\eta}_{A_{1i}} \times \boldsymbol{\eta}_{A_{2i}} \times \boldsymbol{\eta}_{A_{3i}} \times \boldsymbol{\eta}_{A_{4i}})/(\sum_{j=1}^{M} \boldsymbol{\eta}_{A_{1j}} \times \boldsymbol{\eta}_{A_{2j}} \times \boldsymbol{\eta}_{A_{3j}} \times \boldsymbol{\eta}_{A_{4j}})$, $\boldsymbol{\eta}_{A_{1i}}$, $\boldsymbol{\eta}_{A_{2i}}$, $\boldsymbol{\eta}_{A_{3i}}$, $\boldsymbol{\eta}_{A_{4i}}$ are membership function values of the fuzzy variables $\dot{x}$, $\dot{y}$, x, y with respect to fuzzy sets A_1, A_2, A_3, A_4, respectively.

The fuzzy sets of input variables are defined as $\{N, Z, P\}$, where N is negative, Z is zero, and P is positive. Then

Table 1: Fuzzy control rules.

$\dot{x}_i$	N	Z	P
$\dot{y}_i$	N	Z	P
x_i	N	Z	P
y_i	N	Z	P

N: negative; Z: zero; P: positive.

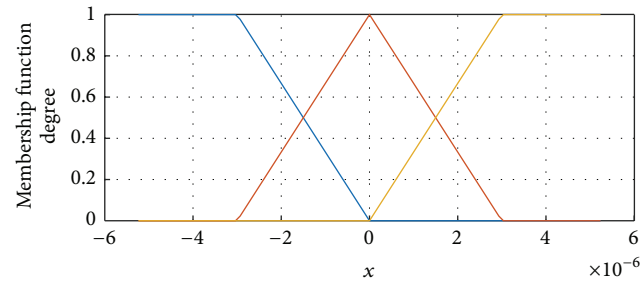


Figure 3: The membership functions for $\dot{x}_i$.

membership functions of $\dot{x}$ are selected as the following triangular functions:

$$\boldsymbol{\eta}_N\left(\dot{x}_i\right) = \begin{cases} 1 & x \le -3 \\ -\dfrac{1}{3}x & -3 \le x \le 0 \end{cases}$$

$$\boldsymbol{\eta}_Z\left(\dot{x}_i\right) = \begin{cases} \dfrac{1}{3}x + 1 & -3 \le x \le 0 \\ -\dfrac{1}{3}x + 1 & 0 \le x \le 3 \end{cases} \tag{8}$$

$$\boldsymbol{\eta}_P\left(\dot{x}_i\right) = \begin{cases} 1 & x \ge 3 \\ \dfrac{1}{3}x & 0 \le x \le 3. \end{cases}$$

The corresponding membership functions of these fuzzy sets labels are depicted in Figure 3. In addition, the membership functions of $\dot{y}_i$, x_i, and y_i are the same with $\dot{x}_i$.

Based on the aforementioned fuzzy sets and membership functions, the fuzzy rules are described in Table 1. Therefore, 81 fuzzy rules are chosen.

The control target for MEMS gyroscope is to maintain the proof mass oscillation at given frequency and amplitude, such as $x_d = A_x \sin(\omega_x t)$, $y_d = A_y \sin(\omega_y t)$ in the x and y directions, respectively. So, reference model can be designed as

$$\ddot{q}_d = A_d q_d, \tag{9}$$

where $q_d = \begin{bmatrix} x_d \\ y_d \end{bmatrix}$, $A_d = \begin{bmatrix} -\omega_x^2 & 0 \\ 0 & -\omega_y^2 \end{bmatrix}$.

And the tracking error is defined as

$$e = q - q_d. \tag{10}$$

Nonlinear sliding mode hypersurface is chosen as

$$s = \dot{e} + \boldsymbol{\alpha} e^{n_1/m_1} + \boldsymbol{\beta} e^{m_2/n_2}, \tag{11}$$

where $\boldsymbol{\alpha} > 0$, $\boldsymbol{\beta} > 0$; $m_1 > n_1 > 0$, $m_2 > n_2 > 0$; what is more, m_1, n_1, m_2, n_2 are odd.

Then the reaching law is designed as the following double exponential function:

$$\dot{s} = -k_1 |s|^a \operatorname{sgn}(s) - k_2 |s|^b \operatorname{sgn}(s), \tag{12}$$

where $k_1 > 0, k_2 > 0, 0 < a < 1, b > 1$.

It should be noted that the converge speed depends on parameters such as $\boldsymbol{\alpha}, \boldsymbol{\beta}, m_1, n_1, m_2, n_2$ and k_1, k_2, a, b.

According to (6), equivalent control law is obtained as

$$\begin{aligned} u_{\text{eq}} &= C^{-1}\left(\ddot{q} - A\dot{q} - Bq - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right)\right) \\ &= C^{-1}\left[\left(\ddot{q}_d + \dot{e}\right) - A\dot{q} - Bq - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right)\right]. \end{aligned} \tag{13}$$

And the derivative of sliding surface (11) is

$$\dot{s} = \ddot{e} + \boldsymbol{\alpha}\left(\frac{n_1}{m_1}\right) e^{n_1/m_1 - 1} + \boldsymbol{\beta}\left(\frac{m_2}{n_2}\right) e^{m_2/n_2 - 1}. \tag{14}$$

Then substituting (12) into (14),

$$\begin{aligned} \ddot{e} = {} & -k_1 |s|^a \operatorname{sgn}(s) - k_2 |s|^b \operatorname{sgn}(s) - \boldsymbol{\alpha}\left(\frac{n_1}{m_1}\right) e^{n_1/m_1 - 1} \\ & - \boldsymbol{\beta}\left(\frac{m_2}{n_2}\right) e^{m_2/n_2 - 1}. \end{aligned} \tag{15}$$

And substituting (15) into (13),

$$\begin{aligned} u_{\text{eq}} = C^{-1}\Bigg[& \ddot{q}_d - k_1 |s|^a \operatorname{sgn}(s) - k_2 |s|^b \operatorname{sgn}(s) \\ & - \boldsymbol{\alpha}\left(\frac{n_1}{m_1}\right) e^{n_1/m_1 - 1} - \boldsymbol{\beta}\left(\frac{m_2}{n_2}\right) e^{m_2/n_2 - 1} - A\dot{q} - Bq \\ & - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right)\Bigg]. \end{aligned} \tag{16}$$

Besides, a robust item is designed to guarantee that the system is asymptotically stable:

$$u_s = -C^{-1} K s. \tag{17}$$

Thus, the adaptive fuzzy finite time sliding mode controller is obtained as

$$u = u_{\text{eq}} + u_s. \tag{18}$$

According to (10), we have

$$\ddot{e} = \ddot{q} - \ddot{q}_d = A\dot{q} + Bq + Cu + P\left(\dot{q}, q, t\right) - \ddot{q}_d. \tag{19}$$

Substituting (18) into (19),

$$\begin{aligned} \ddot{e} = {} & A\dot{q} + Bq + \Bigg[\ddot{q}_d - k_1 |s|^a \operatorname{sgn}(s) - k_2 |s|^b \operatorname{sgn}(s) \\ & - \boldsymbol{\alpha}\left(\frac{n_1}{m_1}\right) e^{n_1/m_1 - 1} - \boldsymbol{\beta}\left(\frac{m_2}{n_2}\right) e^{m_2/n_2 - 1}\Bigg] - A\dot{q} \\ & - Bq - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right) - Ks + P\left(\dot{q}, q, t\right) - \ddot{q}_d \\ = {} & P\left(\dot{q}, q, t\right) - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right) - Ks - k_1 |s|^a \operatorname{sgn}(s) \\ & - k_2 |s|^b \operatorname{sgn}(s) - \boldsymbol{\alpha}\left(\frac{n_1}{m_1}\right) e^{n_1/m_1 - 1} - \boldsymbol{\beta}\left(\frac{m_2}{n_2}\right) \\ & \cdot e^{m_2/n_2 - 1}. \end{aligned} \tag{20}$$

Substituting (20) into (14),

$$\begin{aligned} \dot{s} = {} & P\left(\dot{q}, q, t\right) - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right) - Ks - k_1 |s|^a \operatorname{sgn}(s) \\ & - k_2 |s|^b \operatorname{sgn}(s). \end{aligned} \tag{21}$$

The optimal parameters are set as

$$\begin{aligned} \boldsymbol{\theta}_P^* = \arg\min \left[\sup \left|\widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right) - P\left(\dot{q}, q, t\right)\right|\right] \\ \boldsymbol{\theta}_P \in \boldsymbol{\Omega}_P,\ \dot{q}, q \in R^{2\times 2}, \end{aligned} \tag{22}$$

where $\boldsymbol{\Omega}_P$ is a set of $\boldsymbol{\theta}_P$.

And the minimum approximation errors are defined as

$$w = P\left(\dot{q}, q, t\right) - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right). \tag{23}$$

Substituting (23) into (21), we derive

$$\begin{aligned} \dot{s} = {} & \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P^*\right) - \widehat{P}\left(\dot{q}, q \mid \boldsymbol{\theta}_P\right) + w - Ks \\ & - k_1 |s|^a \operatorname{sgn}(s) - k_2 |s|^b \operatorname{sgn}(s). \end{aligned} \tag{24}$$

Considering (7), (24) can be expressed as

$$\begin{aligned} \dot{s} = {} & \boldsymbol{\varphi}_P^T \boldsymbol{\mu}\left(\dot{q}, q\right) + w - Ks - k_1 |s|^a \operatorname{sgn}(s) \\ & - k_2 |s|^b \operatorname{sgn}(s), \end{aligned} \tag{25}$$

where $\boldsymbol{\varphi}_P = \boldsymbol{\theta}_P^* - \boldsymbol{\theta}_P$.

So adaptive law can be selected as

$$\dot{\boldsymbol{\varphi}}_P = -r s^T \boldsymbol{\mu}\left(\dot{q}, q\right). \tag{26}$$

Namely,

$$\begin{aligned} \dot{\boldsymbol{\theta}}_{Px} &= r s(1)\, \boldsymbol{\mu}_x\left(\dot{q}, q\right), \\ \dot{\boldsymbol{\theta}}_{Py} &= r s(2)\, \boldsymbol{\mu}_y\left(\dot{q}, q\right), \end{aligned} \tag{27}$$

where $\dot{\boldsymbol{\varphi}}_P = -\dot{\boldsymbol{\theta}}_P$.

Lyapunov function is defined as

$$V = \frac{1}{2}\left(s^T s + \frac{1}{r} \boldsymbol{\varphi}_P^T \boldsymbol{\varphi}_P\right). \tag{28}$$

Differentiate V with respect to time yields, and substitute (26) as

$$\dot{V} = s^T w - s^T K s - k_1 s^T |s|^a \operatorname{sgn}(s) - k_2 s^T |s|^b \operatorname{sgn}(s). \tag{29}$$

Owing to the fuzzy approximation theory, adaptive fuzzy system can approximate nonlinear system closely. Therefore, $\dot{V} \leq 0$; namely, the system is asymptotically stable.

4. Simulation Study

In this section, numerical simulations are investigated to track the position and speed trajectories of MEMS gyroscope, compensate parametric uncertainties and external disturbances, and verify the superiority of the proposed approach compared with conventional adaptive fuzzy sliding mode control strategy using linear sliding mode surface. Those two methods are defined as follows.

Method 1. Define the adaptive fuzzy sliding mode control proposed in this paper as Method 1, whose sliding mode surface is shown in (11), and the reaching law is expressed in (12).

Method 2. Define the conventional adaptive fuzzy sliding mode control as Method 2, whose sliding mode surface is $= \dot{e} + \boldsymbol{\beta} e$, and the reaching law is $\dot{s} = 0$.

Parameters of the MEMS gyroscope are as follows:

$$\begin{aligned} m &= 0.57 \times 10^{-8}\,\text{kg}, \\ d_{xx} &= 0.429 \times 10^{-6}\,\text{Ns/m}, \\ d_{yy} &= 0.0429 \times 10^{-6}\,\text{Ns/m}, \\ d_{xy} &= 0.0429 \times 10^{-6}\,\text{Ns/m}, \\ k_{xx} &= 80.98\,\text{N/m}, \\ k_{yy} &= 71.62\,\text{N/m}, \\ k_{xy} &= 5\,\text{N/m}, \\ \Omega_z &= 5.0\,\text{rad/s}. \end{aligned} \tag{30}$$

Since the position of proof mass ranges within the scope of submillimeter and the natural frequency is generally in the range of kilohertz, assume that reference length is $q_o = 10 \times 10^{-6}$ m, reference frequency is $\boldsymbol{\omega}_o = 1$ kHz, and the reference trajectories are $x_d = \sin(6.71t)$, $y_d = 1.2\sin(5.11t)$, respectively.

Then set other simulation parameters as

$$A = \begin{bmatrix} -0.075 & 0.0025 \\ -0.0175 & -0.0075 \end{bmatrix},$$

$$B = \begin{bmatrix} -14207 & -877 \\ -877 & -12564 \end{bmatrix},$$

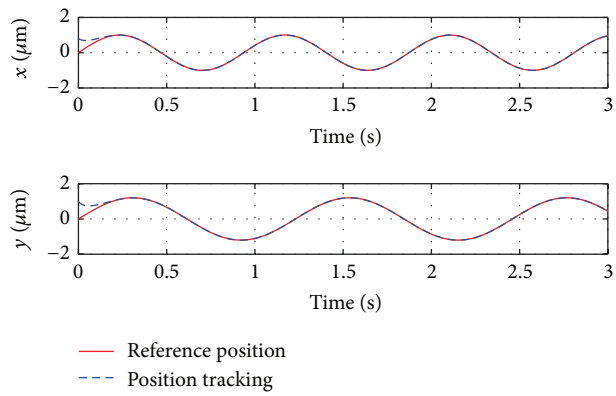


Figure 4: Position tracking of Method 1.

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$K = \begin{bmatrix} 1000 & 0 \\ 0 & 1000 \end{bmatrix},$$

$$\boldsymbol{\alpha} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix},$$

$$\begin{aligned} m_1 &= 3, \\ n_1 &= 2, \\ m_2 &= 3, \\ n_2 &= 1, \\ P(t) &= \begin{bmatrix} 3.2 \times 10^{-6} \\ 5 \times 10^{-6} + 5 \times 10^{-6} \sin(5.11\,(t + 0.3)) \end{bmatrix}, \\ r &= 0.01, \\ a &= 0.5, \\ b &= 10, \\ k_1 &= 1000, \\ k_2 &= 1000. \end{aligned} \tag{31}$$

And select the initial state values of the system as $[0.8 \;\; 0 \;\; 1 \;\; 0]^T$.

Then the position and speed trajectories of Method 1 are shown in Figures 4 and 5 and those of Method 2 are depicted in Figures 6 and 7.

The position tracking error and speed tracking error of Methods 1 and 2 are shown in Figures 8–11, respectively.

Through the tracking simulation of MEMS gyroscope, the proposed approach is with satisfying performance; in

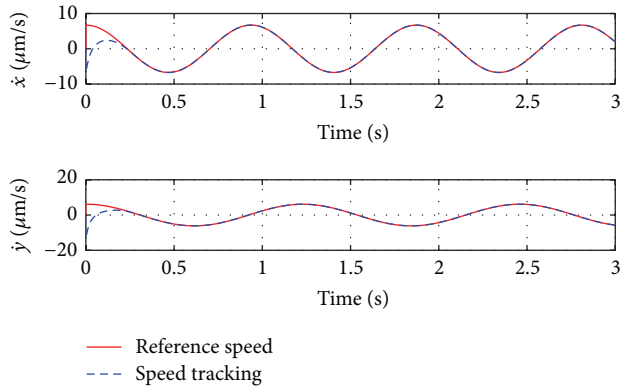


Figure 5: Speed tracking of Method 1.

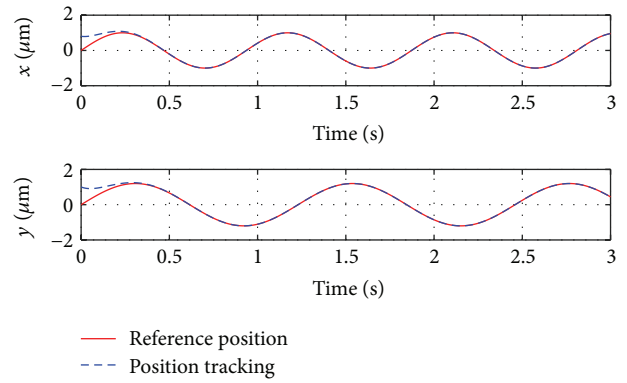


Figure 6: Position tracking of Method 2.

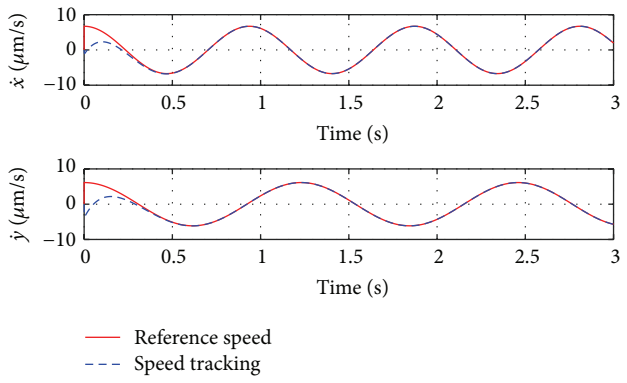


Figure 7: Speed tracking of Method 2.

addition, in comparison to Method 2, the convergence time is shortened to 0.3″ from 0.6″.

5. Conclusion and Future Work

An adaptive fuzzy finite time sliding mode control strategy using nonlinear sliding mode hypersurface and double exponential reaching law is proposed to compensate parametric uncertainties and external disturbance of MEMS gyroscope in this paper. Based on Lyapunov methods, the stability of system can be guaranteed. Simulations verify that, compared

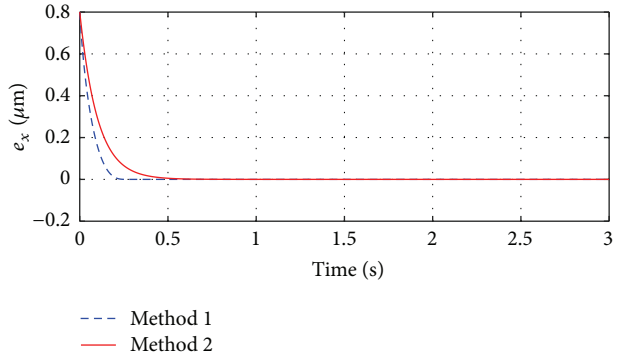


Figure 8: Position tracking error of gyroscope x.

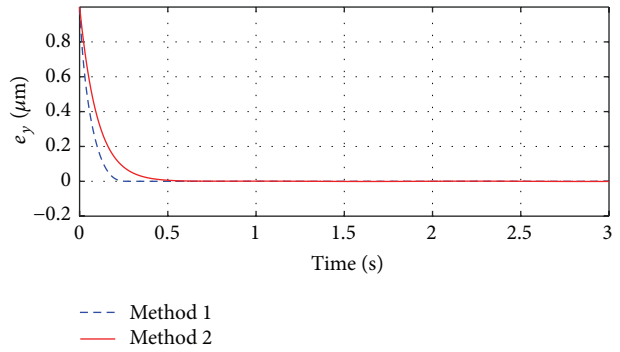


Figure 9: Position tracking error of gyroscope y.

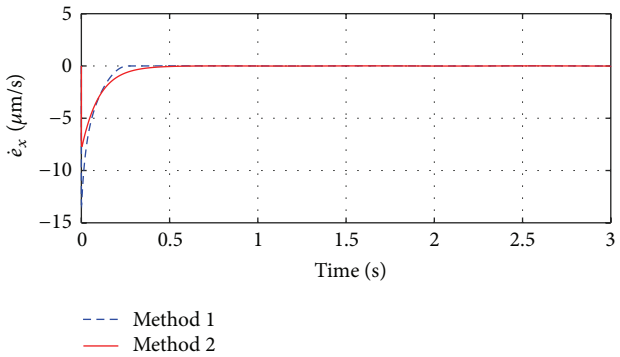


Figure 10: Speed tracking error of gyroscope x.

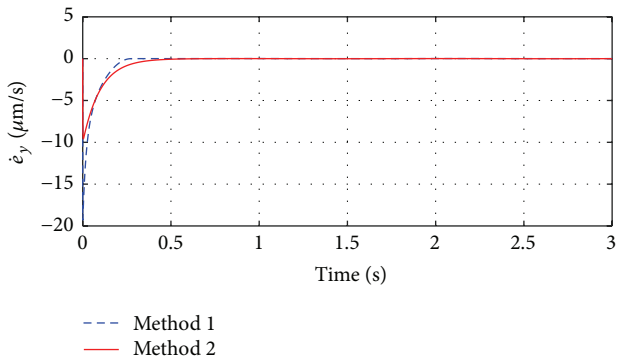


Figure 11: Speed tracking error of gyroscope y.

with conventional adaptive fuzzy linear sliding mode control strategy, the convergence time of finite time convergent control strategy proposed in this paper is shortened to $0.3''$ from $0.6''$; namely, convergence has been significantly improved. For future work, the novel adaptive online constructing fuzzy algorithm [24] can be employed for more efficient learning while disturbance observer based design [25, 26] can be considered to improve system performance.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61304098, 60204005, and 60974109), Aeronautical Science Foundation of China (2015ZA53003), Natural Science Basic Research Plan in Shaanxi Province (2014JQ8326, 2015JM6272, and 2016KJXX-86), Fundamental Research Funds for the Central Universities (3102015AX001, 3102015BJ(II)CG017), Fundamental Research Funds of Shenzhen Science and Technology Project (JCYJ20160229172341417), and International Science and Technology Cooperation Program of China under Grant no. 2014DFA11580.

References

[1] M. Chen and S. S. Ge, "Adaptive neural output feedback control of uncertain nonlinear systems with unknown hysteresis using disturbance observer," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 12, pp. 7706–7716, 2015.

[2] B. Xu, C. Yang, and Y. Pan, "Global neural dynamic surface tracking control of strict-feedback systems with application to hypersonic flight vehicle," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2563–2575, 2015.

[3] B. Xu, Y. Fan, and S. Zhang, "Minimal-learning-parameter technique based adaptive neural control of hypersonic flight dynamics without back-stepping," *Neurocomputing*, vol. 164, pp. 201–209, 2015.

[4] Q. Yang, S. Jagannathan, and Y. Sun, "Robust integral of neural network and error sign control of MIMO nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3278–3286, 2015.

[5] B. Xu, Q. Zhang, and Y. Pan, "Neural network based dynamic surface control of hypersonic flight dynamics using small-gain theorem," *Neurocomputing*, vol. 173, pp. 690–699, 2016.

[6] Y.-J. Liu and S. Tong, "Adaptive NN tracking control of uncertain nonlinear discrete-time systems with nonaffine dead-zone input," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 497–505, 2015.

[7] W. He, Y. Chen, and Z. Yin, "Adaptive neural network control of an uncertain robot with full-state constraints," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 620–629, 2016.

[8] S. Park and R. Horowitz, "New adaptive mode of operation for MEMS gyroscopes," *ASME Journal of Dynamic Systems, Measurement and Control*, vol. 126, no. 4, pp. 800–810, 2004.

[9] C. Batur, T. Sreeramreddy, and Q. Khasawneh, "Sliding mode control of a simulated MEMS gyroscope," *ISA Transactions*, vol. 45, no. 1, pp. 99–108, 2006.

[10] A. Ebrahimi, "Regulated model-based and non-model-based sliding mode control of a MEMS vibratory gyroscope," *Journal of Mechanical Science and Technology*, vol. 28, no. 6, pp. 2343–2349, 2014.

[11] J. Fei and C. Batur, "A novel adaptive sliding mode control with application to MEMS gyroscope," *ISA Transactions*, vol. 48, no. 1, pp. 73–78, 2009.

[12] Z. A. K. Michail, "Terminal attractors in neural networks," *Neural Networks*, vol. 2, no. 2, pp. 259–274, 1989.

[13] X. Yu and M. Zhihong, "Fast terminal sliding-mode control design for nonlinear dynamical systems," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, no. 2, pp. 261–264, 2002.

[14] X. Yu and Z. Man, "Terminal sliding mode control of MIMO systems," *IEEE Trans on Circuits Systems I*, vol. 44, no. 11, pp. 1065–1070, 1997.

[15] A. Ghanbari and M. R. Moghanni-Bavil-Olyaei, "Adaptive fuzzy terminal sliding-mode control of MEMS z-axis gyroscope with extended Kalman filter observer," *Systems Science & Control Engineering*, vol. 2, pp. 183–191, 2014.

[16] W. F. Yan, S. X. Hou, Y. M. Fang, and J. Fei, "Robust adaptive nonsingular terminal sliding mode control of MEMS gyroscope using fuzzy-neural-network compensator," *International Journal of Machine Learning and Cybernetics*, 2016.

[17] A. Bartoszewicz, "A new reaching law for sliding mode control of continuous time systems with constraints," *Transactions of the Institute of Measurement and Control*, vol. 37, no. 4, pp. 515–521, 2015.

[18] W. Gao and J. C. Hung, "Variable structure control of nonlinear systems: a new approach," *IEEE Transactions on Industrial Electronics*, vol. 40, no. 1, pp. 45–55, 1993.

[19] C. J. Fallaha, M. Saad, H. Y. Kanaan, and K. Al-Haddad, "Sliding-mode robot control with exponential reaching law," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 2, pp. 600–610, 2011.

[20] H. Mei and Y. Wang, "Fast convergent sliding mode variable structure control of robot," *Information and Control*, vol. 38, no. 5, pp. 552–557, 2009.

[21] L.-X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least-squares learning," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 807–814, 1992.

[22] X. Zeng and M. G. Singh, "Singh approximation theory of fuzzy systems-MIMO case," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 2, pp. 219–235, 1995.

[23] J. Fei and S. Wang, "Robust adaptive sliding mode control of MEMS gyroscope using T-S fuzzy model," *Nonlinear Dynamics*, vol. 77, no. 1-2, pp. 361–371, 2014.

[24] N. Wang and M. J. Er, "Direct adaptive fuzzy tracking control of marine vehicles with fully unknown parametric dynamics and uncertainties," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 5, pp. 1845–1852, 2016.

[25] B. Xu, F. Sun, Y. Pan, and B. Chen, "Disturbance observer based composite learning fuzzy control of nonlinear systems with unknown dead zone," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2016.

[26] B. Xu, "Disturbance observer-based dynamic surface control of transport aircraft with continuous heavy cargo airdrop," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2016.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 1963450, 8 pages
http://dx.doi.org/10.1155/2016/1963450

Research Article

Detection and Tracking of Road Barrier Based on Radar and Vision Sensor Fusion

Taeryun Kim and Bongsob Song

Department of Mechanical Engineering, Ajou University, Suwon 16499, Republic of Korea

Correspondence should be addressed to Bongsob Song; bsong@ajou.ac.kr

Received 19 June 2016; Accepted 16 August 2016

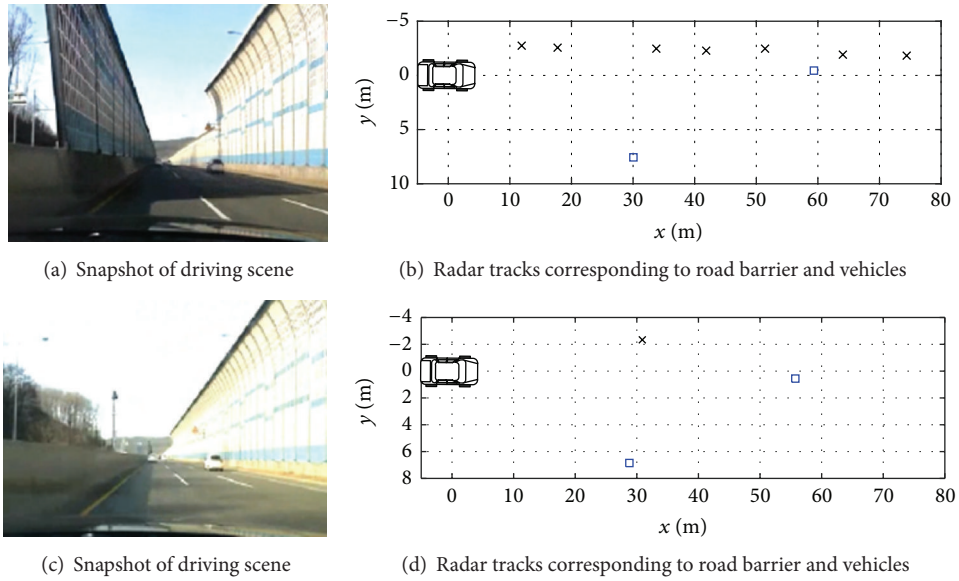
Academic Editor: Antonio Fernández-Caballero

The detection and tracking algorithms of road barrier including tunnel and guardrail are proposed to enhance performance and reliability for driver assistance systems. Although the road barrier is one of the key features to determine a safe drivable area, it may be recognized incorrectly due to performance degradation of commercial sensors such as radar and monocular camera. Two frequent cases among many challenging problems are considered with the commercial sensors. The first case is that few tracks of radar to road barrier are detected due to material type of road barrier. The second one is inaccuracy of relative lateral position by radar, thus resulting in large variance of distance between a vehicle and road barrier. To overcome the problems, the detection and estimation algorithms of tracks corresponding to road barrier are proposed. Then, the tracking algorithm based on a probabilistic data association filter (PDAF) is used to reduce variation of lateral distance between vehicle and road barrier. Finally, the proposed algorithms are validated via field test data and their performance is compared with that of road barrier measured by lidar.

1. Introduction

The driver assistance systems (DAS) such as adaptive cruise control (ACC), forward collision warning, and lane departure warning systems have been commercialized on the market [1]. They have evolved to more intelligent DAS such as automatic emergency braking (AEB), lane change assistance (LCA), and lane keeping assistance (LKA) systems [2]. As prototypes of a highly automated vehicle have been introduced on the media recently, reliability of the performance becomes more important. That is, once false decision is made by a computer or vehicle, it makes the driver have low reliability of the system and may thus result in no use of the system. The reliability of the decision mainly counts on accurate detection and recognition of multiple obstacles and vehicles. For instance, Honda Motor Company had to recall certain model year 2014-2015 Acura vehicles with AEB in the United States. The reason was that a collision mitigation braking system (CMBS) may inappropriately interpret certain roadside infrastructure such as iron fences or metal guardrails as obstacles and unexpectedly apply the brakes [3]. Furthermore, NHTSA in the United States investigated complaints alleging unexpected braking incidents of the autonomous braking system in Jeep Grand Cherokee vehicles with no visible objects on the road [3].

The detection and tracking algorithms of road barrier, which may be called either road border or boundary in the literature, depend on sensor configuration and their models for the road barrier. Most of the sensor configurations are single or a combination of radar [1, 2, 4, 5], camera [6], and lidar (or laser scanners) [7, 8] to recognize the drivable area via reflections from guardrail and curb. Next, extended objects such as road and road barrier are described as clothoid, circle, and elliptical model and their tracking algorithm is based on Kalman filter, probabilistic data association filter (PDAF), and interacting multiple model (IMM) PDAF [1, 2, 4, 9]. In this study, it is assumed that a front radar and a monocular camera are only used for detection and tracking of road barrier. Although additional sensors can be implemented for better performance or the lidar may be used as in the literature, the sensor configuration is limited in the viewpoint of commercialization in the near future. For



(a) Snapshot of driving scene

(b) Radar tracks corresponding to road barrier and vehicles

(c) Snapshot of driving scene

(d) Radar tracks corresponding to road barrier and vehicles

Figure 1: Detection characteristics to road barrier by radar.

instance, cost of sensors, robustness to weather, installation inside bumper, and popularity in automotive market are considered to choose the sensor configuration.

The contribution of this paper is to enhance reliability of road border detection when only few tracks are generated by radar with respect to road border and to improve lateral position accuracy of road border. Since the performance of a commercial radar relies on material and geometry of tunnel and guardrail, different number of tracks are made depending on driving environments. Thus, the estimation for stationary tracks out of detection range is proposed for better performance of road barrier detection. Furthermore, the tracking algorithm of road barrier based on a probabilistic data association filter (PDAF) is proposed to reduce variation of lateral offset, which is a lateral distance between an ego vehicle and road barrier.

2. Problem Statement

When commercial radars for driver assistance systems such as ACC and AEB are used, two challenging problems will be considered in this paper. A normal detection scenario is shown in Figure 1(a) and the corresponding tracks are shown in Figure 1(b). Two tracks to front vehicles are marked as a square and the others marked as × correspond to left guardrail in Figure 1(a). However, as shown in Figures 1(c) and 1(d), few tracks of radar are generated for the road barrier. Its detection performance may rely on the material type and/or shape of road barrier. This problem may lead to difficulty in determining whether there is road barrier in either left or right side.

Next, radar tracks of road barrier are compared with cloud points with magenta color measured by a front lidar sensor in Figures 2(b) and 2(d). As shown in Figures 2(a) and 2(c), inaccuracy of lateral position measured by radar with respect to lidar measurements can be larger in the same driving scenario. This is expected to have large variance of lateral position of road barrier when only radar is used to recognize it.

3. Road Barrier Detection

The proposed road barrier detection based on sensor fusion of radar and monocular camera consists in four steps: selection of region of interest (ROI), estimation, clustering, and representation. First, the selection of ROI is roughly described in Figure 3. That is, based on the assumption that road barrier is placed on either left or right side, zone ② is defined with respect to a body fixed coordinate. If the road in Figure 3 is modeled as [4]

$$y = \frac{1}{2}\kappa x^2 + \varphi x, \tag{1}$$

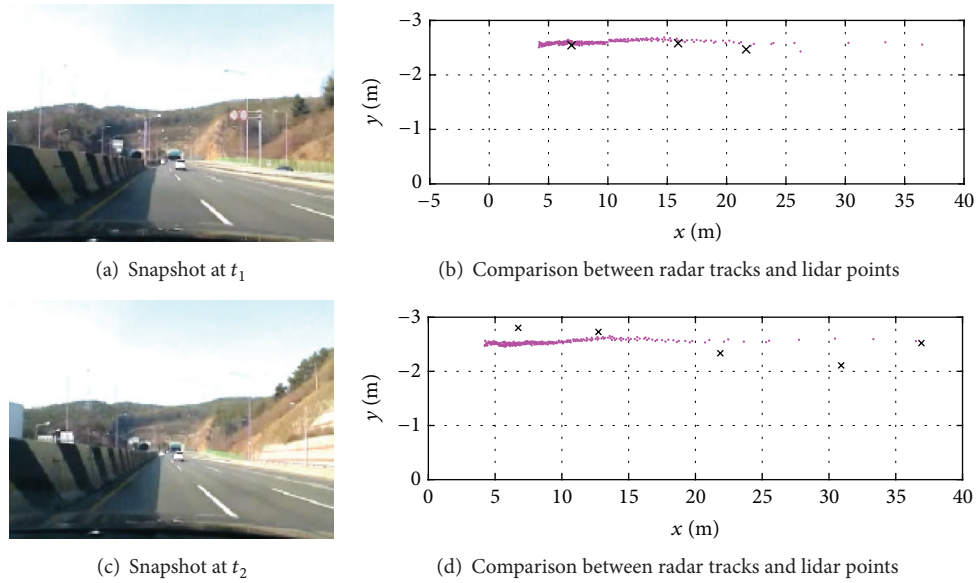
where x and y are longitudinal and lateral position, respectively, in a body fixed coordinate in Figure 4, κ is curvature, and φ is the angle between the longitudinal axis of the vehicle and the road lane from a monocular camera as shown Figure 4. Then, zone ② is written as

$$f_{\text{barrier}}(x) = \frac{1}{2}\kappa x^2 + \varphi x + \varepsilon_1, \tag{2}$$

where ε_1 is a lateral offset which determines a width of zone ②; that is, $l_{\min} \le \varepsilon_1 \le l_{\max}$ and $-l_{\max} \le \varepsilon_1 \le -l_{\min}$.

Next, based on detection range of radar, zone ③ in Figure 3 is defined for estimation of stationary tracks. Before the estimation, a motion attribute to determine whether tracks in zone ② are either stationary or dynamic is decided by

$$p_i = \begin{cases} \text{dynamic} & \text{if } \left|v + \left(\dot{R}_i + y_i\dot{\psi}\right)\right| > \varepsilon_2 \\ \text{stationary} & \text{otherwise,} \end{cases} \tag{3}$$



(a) Snapshot at t_1

(b) Comparison between radar tracks and lidar points

(c) Snapshot at t_2

(d) Comparison between radar tracks and lidar points

FIGURE 2: Road barrier detection by radar and lidar.

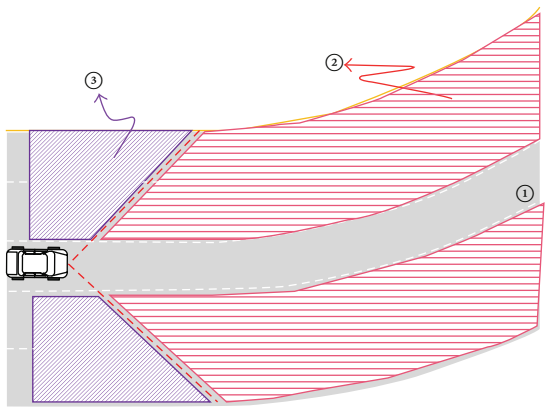


FIGURE 3: Region of interest (ROI) for road barrier detection.

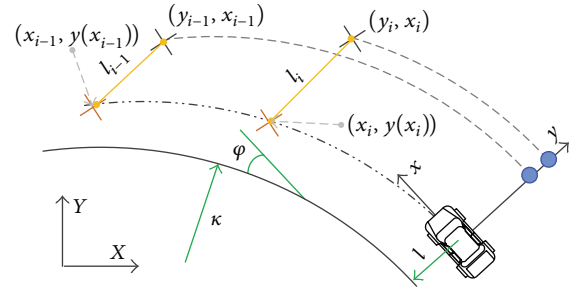


FIGURE 4: Definitions of lateral offset (l_i) and projection point.

where v is the vehicle velocity, $\dot{\psi}$ is the yaw rate, $\dot{R}$ is the range rate, and subscript i stands for the ith radar track. It is remarked that there is uncertainty of detection range of radar. So, it is essential to divide regions either to track or to estimate. Once a stationary track in zone ② enters zone ③, it is estimated based on discrete Kalman filters as follows [10]:

$$\begin{aligned} \hat{x}[k+1 \mid k] &= A\hat{x}[k \mid k], \\ P[k+1 \mid k] &= AP[k-1 \mid k-1]A^T + Q[k], \end{aligned} \tag{4}$$

where $x = \begin{bmatrix} y & v_y & x & v_x \end{bmatrix}^T$ and x and y denote the relative longitudinal and lateral position, respectively, and v_x and v_y are the relative longitudinal and lateral velocity.

The measurement update equations are given by

$$\begin{aligned} &\hat{x}[k+1 \mid k+1] \\ &\quad = \hat{x}[k+1 \mid k] \\ &\qquad + K[k+1](z[k+1] - H\hat{x}[k+1 \mid k]), \end{aligned}$$

$$\begin{aligned} &P[k+1 \mid k+1] \\ &\quad = P[k+1 \mid k] - K[k+1]HP[k+1 \mid k], \end{aligned} \tag{5}$$

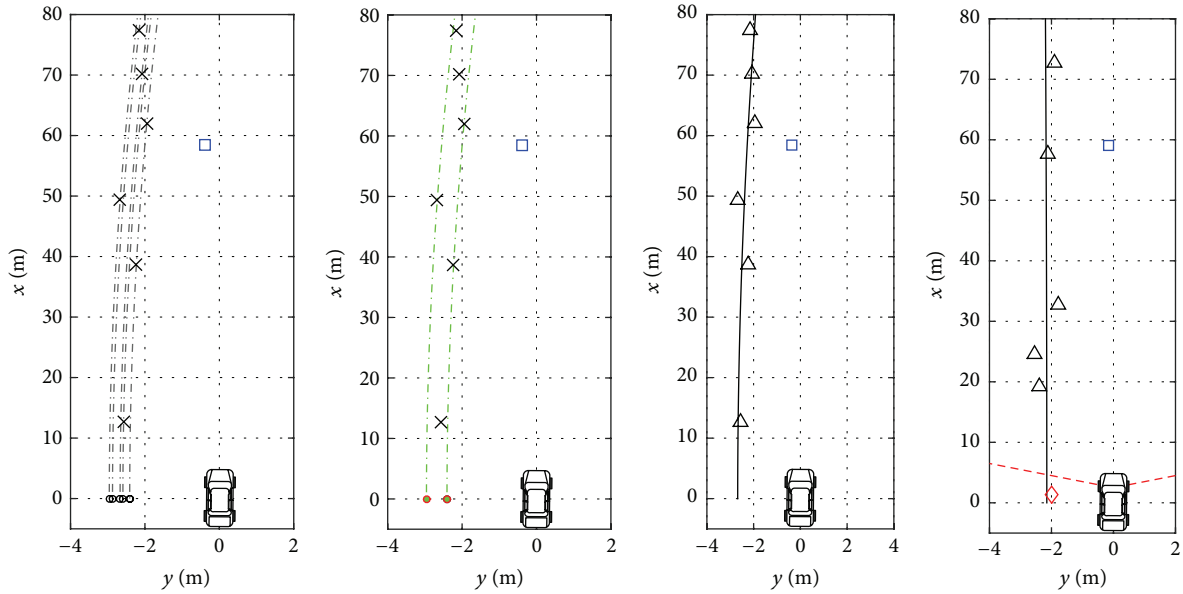
where $K[k+1] = P[k+1 \mid k]H^T(HP[k+1 \mid k]H^T + R[k+1])^{-1}$.

Since a constant velocity (CV) is considered, the system matrix A and measurement matrix H are written as [10, 11]

$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6}$$

where T is sampling time.

In the third step or clustering, in order to group tracks corresponding to road barrier among stationary tracks



(a) Projection (b) Clustering (c) Road barrier detection (d) Estimation

Figure 5: Procedure of road barrier detection.

including estimated track, projection points are calculated as follows (see circles in Figure 4):

$$l_i = y_i - \left(\frac{1}{2}\kappa x_i^2 + \varphi x_i\right), \tag{7}$$

where the subscript i stands for the ith track of radar. After that, the projection points are classified by left or right if the projection points are positioned in zone ②. If the distance between jth and $(j-1)$th projection points is less than ε_2, jth k is increased and $(j-1)$th k is defined as a breakpoint. If the distance between $(m-1)$th and mth projection points is greater than ε_2, the breakpoint is generated at $(m-1)$th k as follows:

$$k[j] = \begin{cases} k[j-1]+1 & \text{if } \left|l_j - l_{j-1}\right| < \varepsilon_3 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$\text{for } j = 2, \ldots, m.$$

Finally, if there are two projection points in two breakpoints or $k[m-1] \geq 1$, all ith tracks satisfying the following inequality are regarded as road barrier [2]:

$$x_p \leq x_i \leq x_{p+1},$$

$$f_{\text{barrier}}\left(x_p\right) - \varepsilon_4 \leq y_i \leq f_{\text{barrier}}\left(x_p\right) + \varepsilon_5 \tag{9}$$

$$\text{for } p = 1, \ldots q,\ i = 1, \ldots, n.$$

Considering the y-coordinates as erroneous in comparison with x-coordinates, the clothoid model which can be approximated by a two-order polynomial is calculated in

$$f_{rb}(x[0]) = \frac{1}{2}\kappa x^2 + \varphi x + l[0], \tag{10}$$

where $l[0] = \{l_1 + l_m\}/2$. It is noted that the calculation of clothoid model corresponds to creation of road barrier.

The procedure to detect road barrier is shown in Figure 5. First, six tracks among seven ones coming from radar are classified as stationary objects. Then, they are projected to y-axis along the road model and the corresponding projection points are shown as circles in Figure 5(a). Next, based on distance between two projection points in (7), two breakpoints are determined as shown in Figure 5(b) and six tracks thus are classified as road barrier based on (9). Then the clothoid model in (10) is determined and shown as a solid line in Figure 5(c). After a few seconds, the closest stationary track in zone ② goes into zone ③ (also refer to Figure 3) and it becomes out of detection range and estimated based on discrete Kalman filter (see also a diamond mark in Figure 5(d)).

4. Tracking Road Barrier

The tracking road barrier is composed of creation, maintenance, and deletion. Creation step uses the result of road barrier detection in (10). Maintenance step based on PDAF tracks lateral offset of road barrier. So, it is rewritten as

$$f_{rb}(x) = \frac{1}{2}\kappa x^2 + \varphi x + \widehat{l}, \tag{11}$$

where $\widehat{l}$ is a tracked value via PDAF and will be derived later.

While the lane information detected by a monocular camera sensor is useful to model roads, its performance depends on light and road conditions. For example, if the lane mark is worn out or covered by soil or snow, it may result in false detection of lane mark. Thus, considering the condition of

lane marks, whether measurements by a monocular camera or estimate value is used is decided as follows [12]:

$$\kappa = \begin{cases} \kappa_v & \text{if } \sigma_L = \text{HIGH } \& \ \sigma_R = \text{HIGH} \\ \dfrac{\dot{\psi}}{v} & \text{otherwise} \end{cases}$$
$$\varphi = \begin{cases} \varphi_v & \text{if } \sigma_L = \text{HIGH } \& \ \sigma_R = \text{HIGH} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where σ is confidence of right or left lanes, $\dot{\psi}$ is yaw rate, and v is vehicle speed.

The PDAF is based on discrete Kalman filter and a state variable is defined as

$$x = \begin{bmatrix} y & v_y \end{bmatrix}^T, \tag{13}$$

where y and v_y are the relative lateral position and velocity, respectively. For time update of the state variable and error covariance matrix,

$$\begin{aligned} \hat{x}\,[k \mid k-1] &= A_d \hat{x}\,[k-1 \mid k-1], \\ P\,[k \mid k-1] &= A_d P\,[k-1 \mid k-1]\, A_d^T + Q\,[k]. \end{aligned} \tag{14}$$

System and measurement metrics are as follows:

$$A_d = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{15}$$

If the number of projection points in two breakpoints is m, the measurement is defined as

$$z\,[k] = \left\{ l_j\,[k] \right\}_{j=1}^{m}, \tag{16}$$

where j means the number of projection points.

In the gating region it has to be decided which measurements (i.e., clustered projection points) are associated with which existing projection points. The corresponding residual and residual covariance matrix S are calculated as

$$\begin{aligned} v_i\,[k] &= z_i\,[k] - H\hat{x}\,[k \mid k-1], \\ S\,[k \mid k-1] &= HP\,[k \mid k-1]\, H^T + R\,[k]. \end{aligned} \tag{17}$$

All measurements are checked whether the normalized residual satisfies the following thresholding condition.

$$v_i^T\,[k]\, S^{-1}\,[k \mid k-1]\, v_i\,[k] \le \varepsilon_5^2. \tag{18}$$

After that, valid measurements which are in gating region are combined to single residual. The weighting is according to the likelihood values of the corresponding measurements. Finally, the measurement update of the state variable calculates estimated track using the combined residual [13]:

$$\begin{aligned} \hat{x}\,[k \mid k] &= \hat{x}\,[k \mid k-1] + K\,[k]\, v\,[k], \\ \hat{l}\,[k \mid k] &= H\hat{x}\,[k \mid k], \end{aligned} \tag{19}$$

where $K[k] = P[k \mid k-1]H^T[HP[k \mid k-1]H^T + R]^{-1}$, $v[k] = \sum_{i=1}^{m[k]} \beta_i v_i[k]$, and

$$\beta_i = \frac{\exp\left[-(1/2)\left(v_i^T\,[k]\, S^{-1}\,[k \mid k-1]\, v_i\,[k]\right)\right]}{\sum_{j=1}^{m[k]} \exp\left[-(1/2)\left(v_i^T\,[k]\, S^{-1}\,[k \mid k-1]\, v_i\,[k]\right)\right]}. \tag{20}$$

5. Experimental Validation

Table 1: Specification of environment sensors.

Sensor	Detection range	Field of view
Radar		
Long mode	174 m	±10°
Short mode	60 m	±45°
Monocular camera	90 m	±40°
Lidar	200 m	±42.5°

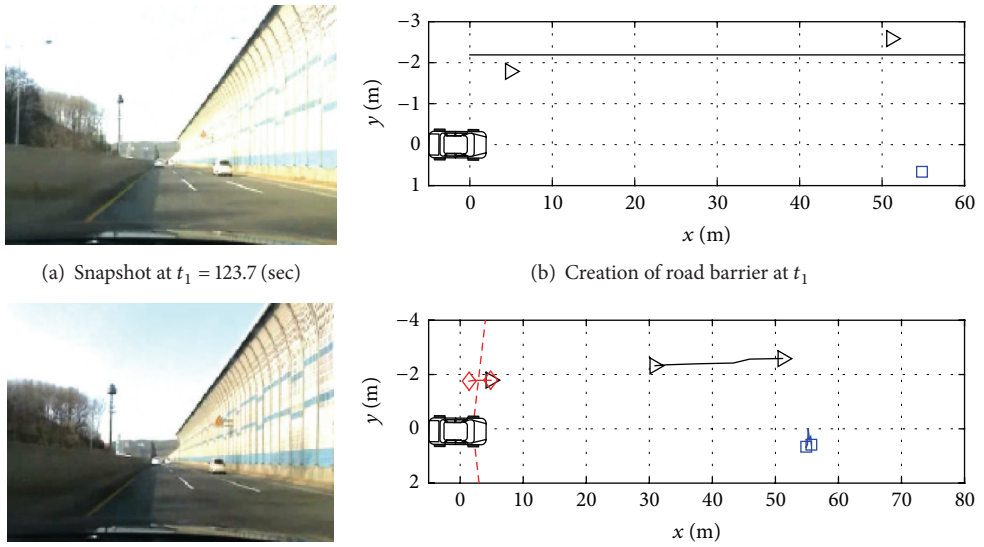
As listed in Table 1, both radar and monocular camera, which are available commercially, are installed on a test vehicle and a front lidar is in addition used for performance comparison. Although a large amount of driving data has to be used for validation, the primitive evaluation of the performance is conducted with driving data of 13 minutes including the driving scenarios in Figures 1 and 2. Most of driving data have been tested on a highway. It is also noted that the driving data corresponding to tollgate, exit zone, and area without any tunnel or guardrail is not considered for performance evaluation.

Based on detection characteristics of radar, five environment scenarios on highway are considered. The first and second scenarios are shown in Figure 1 and called concrete+steel and concrete guardrail, respectively, in the paper. It is interesting to remark that different detection characteristics for the concrete guardrail are shown in Figures 1(c) and 2. In addition, steel guardrail, curb, and tunnel are included for validation as shown in Figure 6. It is thought that most of road barrier on highway in Korea can be described by one of five environment scenarios.

The first case shown in Figures 1(c) and 1(d) is revisited. That is, few tracks for road barrier are generated by radar. As shown in Figure 7, only two tracks with respect to left guardrail are generated at $t_1 = 123.7$ (sec). After 0.8 sec, the nearest stationary track stays out of detection range of radar as shown in Figure 7(d). The solid lines represent trajectory of tracks from $t = 123.7$ to 124.5 (sec). Since the

(a) Steel guardrail (b) Curb (c) Tunnel

Figure 6: Additional environment scenarios for detection and tracking of road barrier.



(a) Snapshot at $t_1 = 123.7$ (sec) (b) Creation of road barrier at t_1

(c) Snapshot at $t_2 = 124.5$ (sec) (d) Trajectory of tracks from t_1 to t_2

Figure 7: Road barrier detection via estimation.

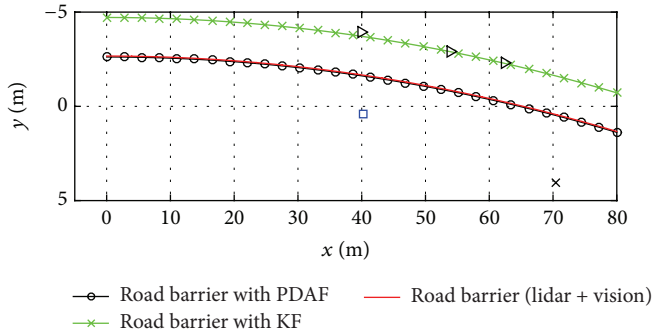
track corresponding to road barrier is located in zone ③ in Figure 3, it becomes a road barrier candidate (see also a diamond mark in the figure). Then there are still two breakpoints for road barrier and the detected road barrier is represented as shown in Figure 7(d).

Two driving scenarios in which lateral position of tracks by radar may be inaccurate are considered when an ego vehicle drives along tunnel or guardrail as shown in Figures 8(a) and 8(c). The proposed tracking algorithm of road barrier is compared with lidar and Kalman filter based approach in [4]. It is shown in Figures 8(b) and 8(d) that the performance of the proposed detection and tracking of road barrier is closer to that of lidar. Furthermore, it is shown in Figure 9 that four different approaches are compared with respect to lateral offset. Then, their relative quantitative performances with respect to lidar are evaluated in terms of root mean square error (RMSE) of lateral offset and recognition accuracy. They are summarized with respect to environment scenarios in Table 2.

Two performance measures are used for validation depending on environment scenarios. The first one is perception of road barrier (RB) and defined as ratio of detection period of RB by the proposed algorithm to detection by image manually. To describe the tracking performance, RMSE of lateral offset between sensor fusion of radar and vision and that of lidar and vision is used for the second performance measure. Finally, the performance comparison is summarized in Table 2. It is validated that the proposed algorithm improves perception of road barrier more when few tracks are often generated by radar in cases of concrete guardrail, tunnel, and curb on highway. Furthermore, it is shown that tracking performance of the proposed algorithm is more robust than others in different environment scenarios.

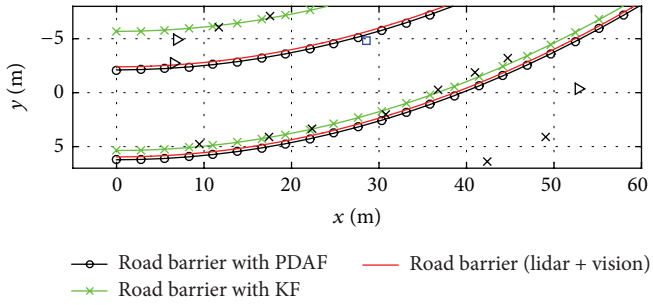
The detection and tracking algorithms are proposed to overcome two problems: one is when few tracks to road barrier are generated by radar and the other is when inaccuracy of lateral position becomes worse in a short time and it happens frequently. Both estimation and clustering

(a) Snapshot at $t = 79.8$ (sec)



(b) Tracking performance in tunnel

(c) Snapshot at $t = 21.9$ (sec)



(d) Tracking performance at curved road

Figure 8: Performance comparison of road barrier tracking.

Table 2: Performance comparison of tracking of road barrier.

Scenario	Detection only	Kalman filter	PDA filter
Perception of RB (%)			
Concrete	65.78	67.53	81.42
Tunnel	62.21	63.49	81.30
Curb	70.96	76.39	89.52
Concrete + iron	98.53	98.87	100.00
Iron	98.76	99.53	100.00
Total	67.75	69.22	84.32
RMSE (m)			
Concrete	0.5396	0.4148	0.3992
Tunnel	0.8010	0.8265	0.7692
Curb	2.6398	2.8001	2.6089
Concrete + iron	0.7354	0.7183	0.6481
Iron	0.7583	0.7438	0.7068
Total	1.1497	1.2231	1.0992

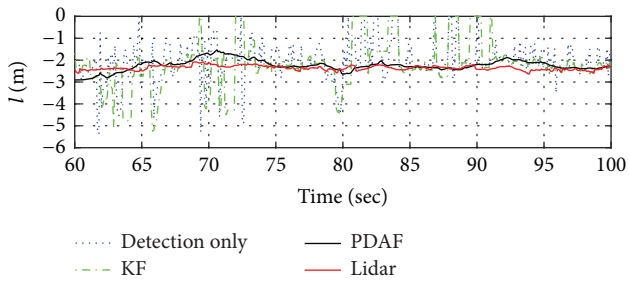


Figure 9: Time response of lateral offset of road barrier.

methods are combined to handle the first problem and the tracking algorithm of lateral offset of road barrier based on PDAF is proposed to deal with the second problem. Its performance is evaluated with comparison of that of lidar and other approaches in the literature. Although it is shown via field test data that the proposed algorithm is good enough to recognize road barrier without lidar, it is quite necessary to validate it with massive field test data in the near future in order to consider various driving scenarios.

Competing Interests

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

Acknowledgments

This work was supported by the Hyundai Motor Company under Grant no. 13RPHMCEL017 and in part by the Formation of Technological Infrastructure Program (N019400027) funded by the Ministry of Trade, Industry and Energy Republic of Korea (MOTIE, Korea).

References

[1] C. Lundquist, L. Hammarstrand, and F. Gustafsson, "Road intensity based mapping using radar measurements with a probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1397–1408, 2011.

[2] F. Breyer, C. Blaschke, B. Färber, J. Freyer, and R. Limbacher, "Negative behavioral adaptation to lane-keeping assistance systems," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 2, pp. 21–32, 2010.

[3] T. Winkle, *Autonomous Driving, Legal and Social Aspects*, Springer, Berlin, Germany, 2016.

[4] C. Lundquist, U. Orguner, and F. Gustafsson, "Extended target tracking using polynomials with applications to road-map estimation," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 15–26, 2011.

[5] A. Polychronopoulos, A. Amditis, N. Floudas, and H. Lind, "Integrated object and road border tracking using 77 GHz automotive radars," *IEE Proceedings—Radar, Sonal and Navigation*, vol. 151, no. 6, pp. 375–381, 2004.

[6] G. Alessandretti, A. Broggi, and P. Cerri, "Vehicle and guard rail detection using radar and vision data fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 95–105, 2007.

[7] J. Han, D. Kim, M. Lee, and M. Sunwoo, "Enhanced road boundary and obstacle detection using a downward-looking LIDAR sensor," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 3, pp. 971–985, 2012.

[8] K. R. S. Kodagoda, S. S. Ge, W. S. Wijesoma, and A. P. Balasuriya, "IMMPDAF approach for road-boundary tracking," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 478–486, 2007.

[9] R. Schubert, K. Schulze, and G. Wanielik, "Situation assessment for automatic lane-change maneuvers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 607–616, 2010.

[10] R. Faragher, "Understanding the basis of the kalman filter via a simple and intuitive derivation," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 128–132, 2012.

[11] H. Kim, B. Song, H. Lee, and H. Jang, "Multiple vehicle tracking and estimation for all-around perception," in *Proceedings of the 12th International Symposium on Advanced Vehicle Control (AVEC '14)*, pp. 480–485, Tokyo, Japan, September 2014.

[12] H.-T. Kim, O. Kwon, B. Song, H. Lee, and H. Jang, "Lane confidence assessment and lane change decision for lane-level localization," in *Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS '14)*, pp. 1448–1451, Seoul, Republic of Korea, October 2014.

[13] R. Möbus and U. Kolbe, "Multi-target multi-object tracking, sensor fusion of radar and infrared," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 732–737, IEEE, June 2004.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 2839372, 9 pages
http://dx.doi.org/10.1155/2016/2839372

Review Article

Validation Techniques for Sensor Data in Mobile Health Applications

Ivan Miguel Pires,[1,2,3] Nuno M. Garcia,[1,3,4] Nuno Pombo,[1,3] Francisco Flórez-Revuelta,[5] and Natalia Díaz Rodríguez[6]

[1]*Instituto de Telecomunicações, Universidade of Beira Interior, Covilhã, Portugal*
[2]*Altranportugal, Lisbon, Portugal*
[3]*Assisted Living Computing and Telecommunications Laboratory (ALLab), Computer Science Department, Universidade of Beira Interior, Covilhã, Portugal*
[4]*ECATI, Universidade Lusófona de Humanidades e Tecnologias, Lisbon, Portugal*
[5]*Department of Computer Technology, Universidad de Alicante, Alicante, Spain*
[6]*Department of Computer Science and Artificial Intelligence, CITIC-UGR, University of Granada, Granada, Spain*

Correspondence should be addressed to Ivan Miguel Pires; impires@it.ubi.pt

Received 27 February 2016; Revised 25 August 2016; Accepted 4 September 2016

Academic Editor: Francesco Dell'Olio

Mobile applications have become a must in every user's smart device, and many of these applications make use of the device sensors' to achieve its goal. Nevertheless, it remains fairly unknown to the user to which extent the data the applications use can be relied upon and, therefore, to which extent the output of a given application is trustworthy or not. To help developers and researchers and to provide a common ground of data validation algorithms and techniques, this paper presents a review of the most commonly used data validation algorithms, along with its usage scenarios, and proposes a classification for these algorithms. This paper also discusses the process of achieving statistical significance and trust for the desired output.

1. Introduction

There has been an increase of the number of mobile applications that make use of sensors to achieve a plethora of goals. Many of these applications are designed and developed by amateur programmers, and that in itself is good as it confirms an increase in the overall set of skills of the developer community. Nevertheless, and even when the applications are developed by professionals or by companies, there are not many applications that publicize or disclose how the sensors' data is processed. This is a problem, in particular when these applications are meant to be used in a scenario where they can influence their users' lives, as for example, when the data is expected to be used to identify Activities of Daily Living (ADLs) or, to an extreme, when the applications are used in medical scenarios.

Due to the nature of the mobile device itself, multiprocessing, with limited computational power and limited battery life, the data that is collected from the sensors is often unusable in its primary form, requiring further processing to allow it to be representative of the event or object that it is supposed to measure. The recording of sensor data and the sequent processing of this data need to include validation subtasks that guarantee that the data are suitable to be fed into the higher-level algorithms.

Moreover, the use of the sensors' data to feed higher-level algorithms needs to guarantee a minimum degree of error, with this error being the difference between the output of these applications, built on limited computational mobile platforms, and the output of a golden standard. To achieve a minimum degree of error, statistical methods need to be applied to ensure that the output of the mobile application is to maximum extent similar to the output given by the relevant golden standard, if and when this is possible.

To mitigate this problem, this paper presents and discusses the most used data validation algorithms and

techniques and their usage in a mobile application that relies on the sensors' data to give meaningful output to its user. The algorithms are listed and their use is discussed. The discussion of the statistical process to ensure maximum reliability of the results is also presented.

The remainder of this paper is organized as follows: this paragraph concludes Section 1, where a short introduction to the problem and a proposal to achieve its mitigation are disclosed; Section 2 presents the most commonly found data validation methods, along with a critical comparison of its usage scenarios; Section 3 deepens the analysis presenting a classification of the data validation methods; Section 4 discusses the applicability of these methods, including the discussion of the degree of trust the data can be expected to provide; finally, Section 5 presents relevant conclusions.

2. Data Validation Methods

Sensor data validation is an important process executed during the data acquisition and data processing modules of the multisensor mobile system. This process consists of the validation of the external conditions of the data and the validity of the data for specific purpose, in order to obtain accurate and reliable results. The sequence of this validation may be applied not only in data acquisition but also in data processing since increase, as these increase the degree of confidence of the systems, with the confidence in the output being of great importance, especially for systems involved in medical diagnosis, but also for the identification of ADLs or sports monitoring.

In addition, data validation methods must be used during the different phases of the conception of a new system, such as design, development, tests, and validation. Therefore, the data validation methods with verified reliability during the conception should be also used to validate the data automatically during the execution time.

One of the causes for the presence of incorrect values during the data acquisition process may be existence of environmental noise. Even when the data is correctly collected, the data may still be incorrect because of noise. Therefore, very often the data captured or processed has to be cleaned, treated, or imputed to obtain better and reliable results. Following the existence of missing values at random instants of time, the causes may be the mechanical problems or power failures of sensors. At this case, data correction methods should be applied, including data imputation and data cleaning. The data validation process may be simplified as presented in Figure 1.

The selection of the best technique for sensor data validation also depends on the type of data collected, the purpose of its application, and the computational platform where the algorithm will be run. Data validation techniques are commonly composed by statistical methods. Due to the characteristics of mobile devices, data validation techniques can be executed locally in the mobile device or at the server-side, depending on the amount of data to validate simultaneously, the frequency of the validation tasks, and the computational, communication, and storage resources needed for the validation. The characteristics of the sensors are also important for the selection of the best techniques, which may be separated in three large groups, which are sensor performance characteristics, pervasive metrics, and environmental characteristics [1].

While data validation is important for improving the reliability of a system, it also depends on other factors, such as power instability, temperature changes, out-of-range data, internal and external noises, and synchronization problems that occur when multiple sensors are integrated into a system [2]. However, the reconstruction of the data and correction for the correct measurement is also important, and several research studies have proposed systems, methods, models, and frameworks to improve the data validation and reconstruction [3, 4].

Sensor data validation methods can be separated in three large groups, such as faulty data detection methods, data correction methods, and other assisting techniques or tools [5].

Firstly, faulty data detection methods may be either simple test based methods or physical or mathematical model based methods, and they are classified in valid data and invalid or missingness data [6, 7]. For the detection of faulty data, the authors in [7] presented an order of methods that should be applied to obtain better results, which are as follows: zero value detection, flat line detection, minimum and maximum values detection, minimum and maximum thresholds based on last values, statistical tests that follow certain distributions, multivariate statistical tests, artificial neural networks (ANNs) [8], one-class support vector machine (SVM) [9], and classification and physical models. On the one hand, simple test based methods include different techniques, such as physical range check, local realistic range detection, detection of gaps in the data, constant value detection, the signals' gradient test, the tolerance band method, and the material redundancy detection [7, 10, 11]. On the other hand, physical or mathematical model based methods include extreme value check using statistics, drift detection by exponentially weighted moving averages, the spatial consistency method, the analytical redundancy method, gross error detection, the multivariate statistical test using Principal Component Analysis (PCA), and data mining methods [7, 12, 13].

Secondly, data correction methods can be carried out by interpolation, smoothing, data mining, and data reconciliation [10, 12, 14]. For the application of the interpolation, the authors of [11] proposed the use of the value measured from the last measurement or the use of the trend from previous sets of measurements. The smoothing methods, for example, moving average and median, may be used to filter out the random noise and convert the data into a smooth curve that is relatively unbiased by outliers [10]. The application of data mining techniques allows the replacement of the faulty values by the measurements performed with several methods, for example, ANNs [14]. The data reconciliation methods, for example, PCA, are used for the calculation of a minimal correction to the measured variables, according to the several constraints of the model [13].

Thirdly, the other assisting techniques or tools are, namely, the checking of the status of the sensors, the checking

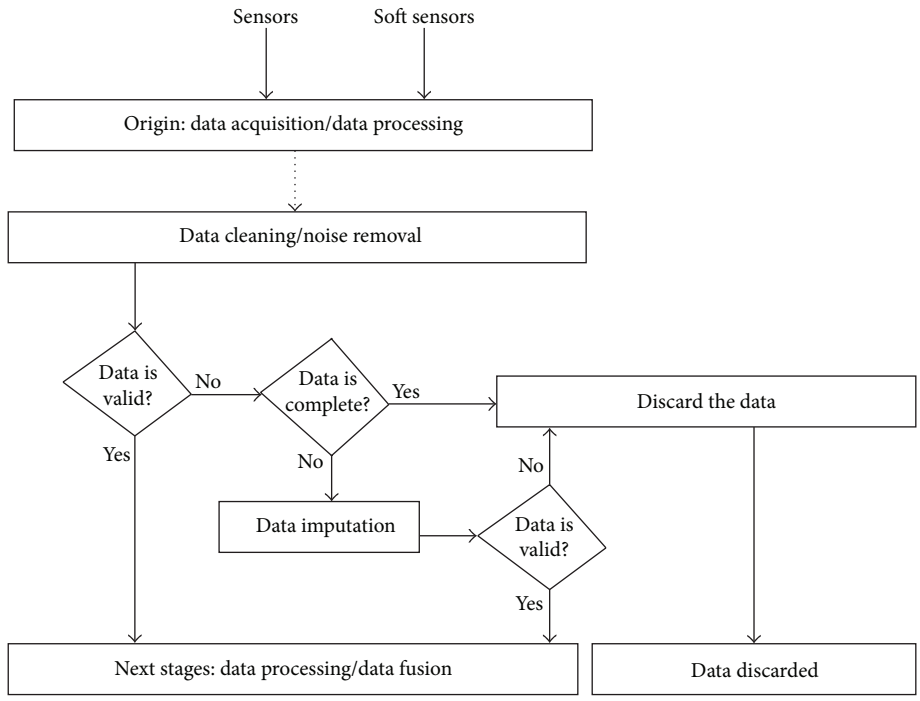


Figure 1: Sequence of activities performed during the data validation process.

of the duration after sensor maintenance, data context classification, the calibration of measuring systems, and the uncertainty consideration [6, 7, 10].

Several research studies have been performed, using data validation techniques. In [15], PCA is used for the compression of linearly correlated data. The authors compared the Auto-Associative Neural Network (AANN) and the Kernel PCA (KPCA) methods for data validation, creating a new approach named as Hybrid AANN-KPCA that uses these two methods. When compared with AANN and KPCA methods, the Hybrid AANN-KPCA achieves better performance results for the prediction or correction of inconsistent data.

In [16], the authors proposed that the data validation may be performed with Kalman filtering and linear predictive coding (LPC), showing that the results using Kalman filtering are better than LPC using several types of data, but the LPC reported a smaller energy consumption.

Several studies proposed the use of ANNs, for example, the Multilayer Perceptron (MLP), that can be trained to perform the identification of faulty sensors using prototype data and used to determine the near optimal subset of sensor data to produce the best results [2, 17–19]. Besides, the sensor data validation may be performed with other probabilistic methods, such as Bayesian Networks, Propagation in Trees, Probabilistic Causal Methods, and Learning Algorithms [20]. The authors of [20] proposed the anytime sensor validation algorithms that combine several probabilistic methods. On the contrary, [21] proposed the validation of data using the Sparse Bayesian Learning and the Relevance Vector Machine (RVM), which are an specialization of SVM.

For the estimation of the values during data validation, the authors of [22] analysed the use of the Kalman filter, which was implemented in two methods: Algorithmic Sensor Validation (ASV), and Heuristic Sensor Validation (HSV). The ASV method implements different statistical methods, for example, mean, standard deviation, and sensor confidence that represent the uncertain nature of sensors. HSV identifies faulty sensor readings as attributable to a sensor or system failure. As an example, the authors of [23] proposed the use of the Kalman filter for the validation of the GPS data.

Other used methods are the grey models, which consists of differential equations describing the behaviour of an accumulated generating operation (AGO) data sequence. As an example, [4] presented a novel self-validating strategy using grey bootstrap method (GBM) for data validation and dynamic uncertainty estimation of self-validating sensor. The GBM can evaluate the measurement uncertainty due to poor information and small sample.

In [2], the Autoregressive Moving Averages (ARMA) transform the process for determining the validity of the acquired data, evaluating the levels of noise and providing a timely warning from the expected signals. The model created for ARMA includes linear regression techniques to predict the invalid values with Autoregressive (AR) and Moving Average (MA) models. Sensor Data Validation in Aeroengine Vibration Tests also implements the Autoregressive (AR) Model, complemented with the Empirical Mode Decomposition (EMD) [24]. Another method presented is the sensor validation and fusion of the Nadaraya-Watson statistical estimator [25], using a Fuzzy Logic model [26]. These methods and others, including the use of Gaussian distributions and error detection methods, may be also used to improve the quality of the measurements [27, 28].

Intelligent sensor systems are able to perform the capture and validation of the sensors' data. Staroswiecki [29] argues that the data validation is important to increase the confidence level of these systems, proposing two types of validation, such as technological and functional. Technological validation consists on the analysis of the conditions of the hardware resources of the sensors, but it does not guarantee that the estimation produced by the sensor is correct, but only that the operating conditions were not against possible correctness. On the contrary, functional validation consists of Fault Detection and Isolation (FDI) procedures, which consists of the use of algorithms to complement the Technological Validation. The authors of [30] also agreed with Staroswiecki in the separation of the data validation in two types, presenting a real time algorithm based on probabilistic methods. Other studies have been researched and developed, including the data validation techniques using intelligent sensor systems [31].

Another powerful technique for data validation consists of the use of self-validating (SEVA) sensors, which provide an estimation of the error bounds during the measurements [32]. SEVA are widely researched in literature. An example, using a Back-Propagation (BP) model, is applied into a system to obtain an estimated value and then a fault detection method called SPRT (sequential probability ratio test), identifying the validity of the system [33]. For the use of SEVA technologies, the authors of [34] also proposed the validated random fuzzy variable (VRFV) based uncertainty evaluation strategy for the online validated uncertainty (VU) estimation. In [35], the authors presented a novel strategy of using polynomial predictive filters coupled with VRFV which is proposed for the online measurements validation and validated uncertainty estimation of multifunctional self-validating sensors. These authors also performed a research about the use of some fuzzy logic rules, comparing the predicted values with the actual measurements to obtain the confidence evaluation [36]. In [37], the authors proposed an approach of sensor data validation using self-reporting, including the measurement based on the data quality, that is, validating the data loss measured by periodic sensors, the timing of data collection, and the accuracy of the detection of changes. ANNs may be used for SEVA with self-organizing maps (SOM) [38], which are trained using unsupervised learning techniques to produce a low-dimensional, discretized representation of the input space of the training samples [39].

The use of valid data is important for the developments of intelligent sensor systems, which may be used for health purposes and, consequently, for the detection of the ADLs [40–45]. The use of mobile devices allows the data acquisition anywhere and at anytime, but these devices have several constraints, such as low memory, processing power, and battery resources, but data validation may help for increasing of the performance of the measurements, reducing the resources needed [46–48]. In general, these systems use probabilistic methods to detect the failures at real-time to obtain better results.

Table 1 presents a summary of the data validation methods included on each category. The methods that are mainly implemented use statistical and artificial intelligence techniques, such as PCA, RVM, ANNs, and others, increasing the reliability of the data acquisition and data processing algorithms. In spite of the SVM and the ANN working in a slightly different manner, their foundations are quite similar. In fact the SVM without kernel is a single neural network neuron with a different cost function. Congruently, when the SVM has a kernel it is comparable with a 2-layer ANN.

Following the methods presented at Table 1, the most studied scenarios for data validation are mainly related to health sciences, laboratory experiments, and other undifferentiated tasks. However, only a minor part of studies is related to the use of mobile devices, smart sensors, and other devices used daily. Besides, the development of healthcare solutions based on the sensors available on the mobile devices increases the requirement of the validation of the data collected by the sensors available on the mobile devices. Depending on the types of the data, for some complex data acquired, such as images, videos, GPS signal, and other complex types of data, the validation of the data should be accomplished by other auxiliary systems working at the same time, validating the data at the server-side, but a constant network connection must be available. Other topologies of systems may be susceptible for the implementation of data validation techniques. The Wireless Sensor Networks (WSN) are an example of systems where the different nodes of the network may perform the validation of the data collected for the neighbourhood nodes, and these nodes may be composed of different types of sensors. However, the main topology for the implementation with mobile devices is the self-validation using only the sensors and the data available on the mobile device.

The data validation may be executed automatically and transparently for the mobile devices' user and, commonly, at least one of the methods for each stage should be implemented in a system to perform the validation of the sensors' data. Firstly, for faulty data detection methods, the ANNs are the most used methods for the training of the data and for the detection of the inconsistent values. Secondly, for data correction methods, the most used method is the Kalman filter. Thirdly, the other assisting techniques that are commonly applied are the data context classification, the checking of the status of sensors, and the uncertainty considerations. Applying the data validation techniques correctly, the reliability and acceptability of the systems may be increased.

3. Classification of Data Validation Methods

Data validation methods may be classified in three large groups [5] as follows: faulty data detection methods, data correction methods, and other assisting techniques or tools.

The faulty data detection methods and the data correction methods may be executed sequentially in a multisensor system in order to obtain the results based on valid data. The other assisting techniques or tools mainly consist of the validation of the working state of the sensors, and this validation may be executed at the same time of the execution of faulty data detection and data correction methods, because these types of failures invalidated the results of the algorithms.

Table 1: Classification of the data validation methods by functionality.

Groups of data validation methods	Methods included	Description
Faulty data detection methods	ANNs (i) MLP; AANN; BP algorithm; SVM; Instance based (i) SOM Gaussian distributions Statistical methods (i) ASV; HSV Probabilistic methods (i) Bayesian Networks; Propagation in Trees; Probabilistic Causal Methods; Learning Algorithms; Sparse Bayesian Learning; RVM; SPRT Dimensionality Reduction (i) Fuzzy logic; PCA; KPCA; others (i) Hybrid AANN-KPCA	Consisting of the detection of faulty or incorrect values discovered during the data acquisition and processing stages
Data correction methods	Kalman filter LPC ARMA (i) AR; MA; EMD Nadaraya-Watson statistical estimator Interpolation Smoothing Data mining techniques Data reconciliation techniques	Consisting of the estimation of faulty or incorrect values obtained during the data acquisition and processing stages
Other assisting techniques or tools	Checking of the status of the sensors Checking of the duration after sensor maintenance Data context classification Calibration of measuring systems Uncertainty consideration Grey models (i) GBM; dynamic uncertainty estimation of self-validating sensor VRFV method	These are different approaches created for the correct validation of the data

These different approaches are based on either mathematical methods, for example, statistical or probabilistic methods, or complex analysis, for example, artificial intelligence methods. According to [49], the data validation methods may be classified in several types of methods, which are presented in Figure 2.

As depicted in Figure 2, the faulty data detection methods, used to detect failures on the sensors' signal, may include ANNs, dimensional reduction methods, instance based methods, probabilistic and statistical methods, and Bayesian methods. On the contrary, the data correction methods include the following methods: filtering, regression, estimation, interpolation, smoothing, data mining, and data reconciliation. These methods work specifically with the sensors' data and the selection of the methods that can be applied by a system should consider the system's usage scenarios.

Finally, the other assisting techniques or tools are mainly related to detection of problems originated by either hardware components or its working environment. In addition, on real-time systems, these problems should be verified constantly to prevent the existence of failures in the data captured.

4. Applicability of the Sensor Data Validation Methods

Mobile devices have a plethora of sensors available for the measurement of several parameters, including the identification of the ADLs. Examples of these sensors include the accelerometer, the gyroscope, the magnetometer, the GPS, and the microphone.

The data acquisition using accelerometers may fail because of several problems, including problems related with the internal electronic amplifier of the Integrated Electronic Piezoelectric (IEPE) device, the exposure to temperatures beyond the accelerometer working range, failure related with electrical components, capture of environmental noise, the multitasking and multithreading capabilities of the mobile devices that may cause irregular sampling rates, the positioning of the accelerometer, the low processing and memory power, and the battery consumption [50]. The causes of failure of an accelerometer are similar to the causes of the failure of a gyroscope, a magnetometer, or a microphone [51]. In addition, the GPS has another failure cause, which consists of the low connectivity of satellites in indoor environments [52].

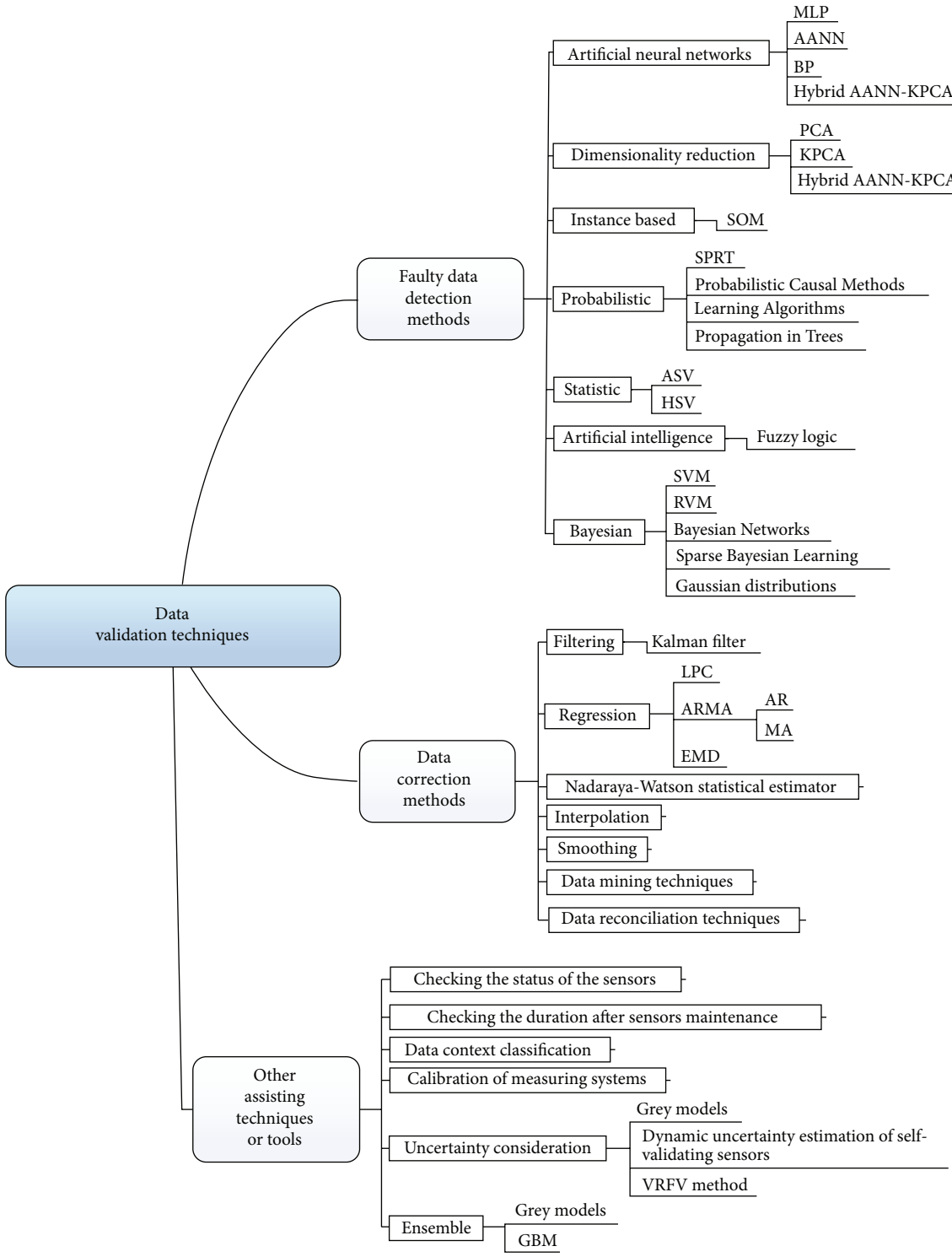


Figure 2: Different categories of the data validation methods.

The validation of the data is important, but, for critical systems, for example, clinical systems, not only should the input data be validated, but also the results should be validated to guarantee the reliability, accuracy and, consequently, acceptance of the system. The validation of the system may consist of the detection of failures and the methods that may be applied are the faulty detection methods. As presented in Section 3, the methods that may be included in this category are probabilistic and statistical methods, among others, which may be used to validate the results of the system. This validation can be performed by comparing the results obtained by an equivalent system which is considered

to be a *gold standard* [53] with the results obtained by the developed methods implemented by different sensors or devices, for example, a mobile device.

Once estimated the initial error of the system, that is, how different the obtained results are from the results obtained by the gold standard system, the validation of the results of the system consist of three steps, such as the definition of the confidence level needed for the acceptance of the system, the determination of the minimum number of experiments needed to validate the application with confidence level defined, and the validation of the results when compared to a golden-standard [54]. The definition of the degree of confidence of the system is a choice of the development team. The system design leader may define what system needs to have a maximum 5% error 95% of the times. Using these parameters, a minimum number of calibration experiments need to be performed to allow the fine tuning of the algorithm. The minimum number of experiments may be measured by several statistical tests, for example, Student's t-test [55].

After the calibration of the algorithms in the system, further tests and comparison with golden-standard systems can be done to insure that the results reported by the system have a 5% maximum error when compared to the golden standard results, for 95% of the time. Note that the 5% and 95% values are merely indicative. Moreover, the data collection stage must hold into consideration the limits for the optimal functioning of the sensors. As these limits are extremely dependent on the task the sensors must perform, we do not discuss them in this paper, for example, if the application is supposed to track the movements of a sportsperson in an open environment, it is possible that a thermal sensor reports an environment temperature of −5°C, yet, for an application that tracks the indoor activity of an elder, such value should raise an alarm. In this extreme case, it is even possible that more robust systems need to contain different types of sensors.

5. Conclusion

The validation of the data collected by sensors in a mobile device is an important issue for two main reasons: the first one is the increasing number of devices and the applications that make use of the devices' sensors; the other is that also increasingly users rely on these devices and applications to collect information and make decisions that may be critical for the user's life and well-being.

Despite the fact that there is a wide array and types of data validation algorithms, there is also a lack of published information on the validity of many mobile applications. Also, it is impossible to present a critical comparison of the discussed methods, even within their respective categories, as their efficiency is extremely dependent on their particular usage; for example, the efficiency of a specific method may be very dependent on the number and type of features the algorithm selects on the signal to be processed, and of course these features are chosen in view of the intended purpose of the application. Additionally, it is possible that even with the same chosen method and the same chosen set of features, different authors report different efficiency ratios; for example, their base population sample varies in size and/or type using different population sizes or using populations that are homogenous in age (elders or youngsters).

This paper has presented a discussion on the different types of data validation methods such as faulty data detection, data correction, and assisting techniques or tools. Furthermore, a classification of these methods in accordance with its functionalities was provided. Finally, the relevance of the data validation methods for critical systems in terms of its reliability, accuracy and acceptance was highlighted. Complementary studies should be addressed aiming at providing an overview on the use of valid data for the identification of the ADLs.

Competing Interests

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

Acknowledgments

This work was supported by FCT project UID/EEA/50008/2013 (Este trabalho foi suportado pelo projecto FCT UID/EEA/50008/2013). The authors would also like to acknowledge the contribution of the COST Action IC1303 Architectures, Algorithms and Protocols for Enhanced Living Environments (AAPELE).

References

[1] A. Braun, R. Wichert, A. Kuijper, and D. W. Fellner, "A benchmarking model for sensors in smart environments," in *Ambient Intelligence: European Conference, (AmI '14), Eindhoven, The Netherlands, November 2014. Revised Selected Papers*, E. Aarts, B. de Ruyter, P. Markopoulos et al., Eds., pp. 242–257, Springer, Cham, Switzerland, 2014.

[2] J. Garza-Ulloa, H. Yu, and T. Sarkodie-Gyan, "A mathematical model for the validation of the ground reaction force sensor in human gait analysis," *Measurement*, vol. 45, no. 4, pp. 755–762, 2012.

[3] M. À. Cugueró, M. Christodoulou, J. Quevedo, V. Puig, D. García, and M. P. Michaelides, "Combining contaminant event diagnosis with data validation/reconstruction: application to smart buildings," in *Proceedings of the 22nd Mediterranean Conference on Control and Automation (MED '14)*, pp. 293–298, June 2014.

[4] Y. Chen, J. Yang, and S. Jiang, "Data validation and dynamic uncertainty estimation of self-validating sensor," in *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC '15)*, pp. 405–410, Pisa, Italy, May 2015.

[5] S. Sun, J. Bertrand-krajewski, A. Lynggaard-Jensen et al., "Literature review for data validation methods," *SciTechnol*, vol. 47, no. 2, pp. 95–102, 2011.

[6] J.-L. Bertrand-Krajewski, S. Winkler, E. Saracevic, A. Torres, and H. Schaar, "Comparison of and uncertainties in raw sewage COD measurements by laboratory techniques and field UV-visible spectrometry," *Water Science and Technology*, vol. 56, no. 11, pp. 17–25, 2007.

[7] N. Branisavljević, Z. Kapelan, and D. Prodanović, "Improved real-time data anomaly detection using context classification," *Journal of Hydroinformatics*, vol. 13, no. 3, pp. 307–323, 2011.

[8] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 1995.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pp. 144–152, ACM Press, Pittsburgh, Pa, USA, 1992.

[10] M. Mourad and J.-L. Bertrand-Krajewski, "A method for automatic validation of long time series of data in urban hydrology," *Water Science and Technology*, vol. 45, no. 4-5, pp. 263–270, 2002.

[11] G. Olsson, M. K. Nielsen, Z. Yuan, and A. Lynggaard-Jensen, *Instrumentation, Control and Automation in Wastewater Systems*, IWA, London, UK, 2005.

[12] F. Edthofer, J. Van den Broeke, J. Ettl, W. Lettl, and A. Weingartner, "Reliable online water quality monitoring as basis for fault tolerant control," in *Proceedings of the 1st Conference on Control and Fault-Tolerant Systems (SysTol '10)*, pp. 57–62, October 2010.

[13] S. J. Qin and W. Li, "Detection, identification, and reconstruction of faulty sensors with maximized sensitivity," *AIChE Journal*, vol. 45, no. 9, pp. 1963–1976, 1999.

[14] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[15] R. Sharifi and R. Langari, "A hybrid AANN-KPCA approach to sensor data validation," in *Proceedings of the 7th Conference on 7th WSEAS International Conference on Applied Informatics and Communications*, vol. 7, pp. 85–91, 2007.

[16] C. C. Castello, J. R. New, and M. K. Smith, "Autonomous correction of sensor data applied to building technologies using filtering methods," in *Proceedings of the 1st IEEE Global Conference on Signal and Information Processing (GlobalSIP '13)*, pp. 121–124, December 2013.

[17] M. Kasinathan, B. S. Rao, N. Murali, and P. Swaminathan, "An artificial neural network approach for the discordance sensor data validation for SCRAM parameters," in *Proceedings of the 1st International Conference on Advancements in Nuclear Instrumentation, Measurement Methods and their Applications (ANIMMA '09)*, pp. 1–5, Marseille, France, June 2009.

[18] E. V. Zabolotskikh, L. M. Mitnik, L. P. Bobylev, and O. M. Johannessen, "Neural networks based algorithms for sea surface wind speed retrieval using SSM/I data and their validation," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '99)*, vol. 2, pp. 1010–1012, Hamburg, Germany, June-July 1999.

[19] E. Gaura, R. Newman, M. Kraft, A. Flewitt, and W. de Lima Monteiro, *Smart MEMS and Sensor Systems*, World Scientific, Singapore, 2006.

[20] P. H. Ibarguengoytia, *Any Time Probabilistic Sensor Validation*, University of Salford, Salford, UK, 1997.

[21] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[22] S. Alag, A. M. Agogino, and M. Morjaria, "A methodology for intelligent sensor measurement, validation, fusion, and fault detection for equipment monitoring and diagnostics," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 15, no. 4, pp. 307–320, 2001.

[23] E. Shi, "An improved real-time adaptive Kalman filter for low-cost integrated GPS/INS navigation," in *Proceedings of the International Conference on Measurement, Information and Control (MIC '12)*, vol. 2, pp. 1093–1098, May 2012.

[24] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[25] S. J. Wellington, J. K. Atkinson, and R. P. Sion, "Sensor validation and fusion using the Nadaraya-Watson statistical estimator," in *Proceedings of the 5th International Conference on Information Fusion*, vol. 1, pp. 321–326, Annapolis, Md, USA, 2002.

[26] K. E. Holbert, A. S. Heger, and N. K. Alang-Rashid, "Redundant sensor validation by using fuzzy logic," *Nuclear Science and Engineering*, vol. 118, no. 1, pp. 54–64, 1994.

[27] Y. Ai, X. Sun, C. Zhang, and B. Wang, "Research on sensor data validation in aeroengine vibration tests," in *Proceedings of the International Conference on Measuring Technology and Mechatronics Automation (ICMTMA '10)*, vol. 3, pp. 162–166, March 2010.

[28] F. Pfaff, B. Noack, and U. D. Hanebeck, "Data validation in the presence of stochastic and set-membership uncertainties," in *Proceedings of the 16th International Conference on Information Fusion (FUSION '13)*, pp. 2125–2132, Istanbul, Turkey, July 2013.

[29] M. Staroswiecki, "Intelligent sensors: a functional view," *IEEE Transactions on Industrial Informatics*, vol. 1, no. 4, pp. 238–249, 2005.

[30] P. H. Ibargiengoytia, L. E. Sucar, and S. Vadera, "Real time intelligent sensor validation," *IEEE Transactions on Power Systems*, vol. 16, no. 4, pp. 770–775, 2001.

[31] J. Rivera-Mejýa, E. Arzabala-Contreras, and Á. G. Leýn-Rubio, "Approach to the validation function of intelligent sensors based on error's predictors," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference (I2MTC '10)*, pp. 1121–1125, IEEE, May 2010.

[32] M. P. Henry and D. W. Clarke, "The self-validating sensor: rationale, definitions and examples," *Control Engineering Practice*, vol. 1, no. 4, pp. 585–610, 1993.

[33] B. Mounika, G. Raghu, S. Sreelekha, and R. Jeyanthi, "Neural network based data validation algorithm for pressure processes," in *Proceedings of the International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT '14)*, pp. 1223–1227, July 2014.

[34] Q. Wang, Z. Shen, and F. Zhu, "A multifunctional self-validating sensor," in *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC '13)*, pp. 1283–1288, Minneapolis, Minn, USA, May 2013.

[35] Z. Shen and Q. Wang, "Data validation and validated uncertainty estimation of multifunctional self-validating sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 7, pp. 2082–2092, 2013.

[36] Z. Shen and Q. Wang, "Data validation and confidence of self-validating multifunctional sensor," in *Proceedings of the Sensors*, pp. 1–4, Taipei, Taiwan, October 2012.

[37] J. Doyle, A. Kealy, J. Loane et al., "An integrated home-based self-management system to support the wellbeing of older adults," *Journal of Ambient Intelligence and Smart Environments*, vol. 6, no. 4, pp. 359–383, 2014.

[38] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[39] B. Lamrini, E.-K. Lakhal, M.-V. Le Lann, and L. Wehenkel, "Data validation and missing data reconstruction using self-organizing map for water treatment," *Neural Computing and Applications*, vol. 20, no. 4, pp. 575–588, 2011.

[40] A. Pantelopoulos and N. G. Bourbakis, "A survey on wearable sensor-based systems for health monitoring and prognosis," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 1, pp. 1–12, 2010.

[41] S. Helal, J. W. Lee, S. Hossain, E. Kim, H. Hagras, and D. Cook, "Persim—simulator for human activities in pervasive spaces," in *Proceedings of the 7th International Conference on Intelligent Environments (IE '11)*, pp. 192–199, Nottingham, UK, July 2011.

[42] H. Eren, "Assessing the health of sensors using data historians," in *Proceedings of the IEEE Sensors Applications Symposium (SAS '12)*, pp. 208–211, February 2012.

[43] S. Oonk, F. J. Maldonado, and T. Politopoulos, "Distributed intelligent health monitoring with the coremicro Reconfigurable Embedded Smart Sensor Node," in *Proceedings of the IEEE AUTOTESTCON*, pp. 233–238, Anaheim, Calif, USA, September 2012.

[44] C.-M. Chen, R. Kwasnicki, B. Lo, and G. Z. Yang, "Wearable tissue oxygenation monitoring sensor and a forearm vascular phantom design for data validation," in *Proceedings of the 11th International Conference on Wearable and Implantable Body Sensor Networks (BSN '14)*, pp. 64–68, June 2014.

[45] F. J. Maldonado, S. Oonk, and T. Politopoulos, "Enhancing vibration analysis by embedded sensor data validation technologies," *IEEE Instrumentation and Measurement Magazine*, vol. 16, no. 4, pp. 50–60, 2013.

[46] E. Miluzzo, *Smartphone Sensing*, Dartmouth College, Hanover, New Hampshire, 2011.

[47] F. J. Maldonado, S. Oonk, and T. Politopoulos, "Optimized neuro genetic fast estimator (ONGFE) for efficient distributed intelligence instantiation within embedded systems," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '13)*, pp. 1–8, August 2013.

[48] B. Wallace, R. Goubran, F. Knoefel et al., "Automation of the validation, anonymization, and augmentation of big data from a multi-year driving study," in *Proceedings of the IEEE International Congress on Big Data*, pp. 608–614, New York, NY, USA, June 2015.

[49] J. Brownlee, *A Tour of Machine Learning Algorithms*, Machine Learning Mastery, 2013.

[50] R. Denton, "Sensor Reliability Impact on Predictive Maintenance Program Costs," 2010, Wilcoxon Research.

[51] O. Brand, G. K. Fedder, C. Hierold, J. G. Korvink, O. Tabata, and T. Tsuchiya, *Reliability of MEMS: Testing of Materials and Devices*, John Wiley & Sons, New York, NY, USA, 2013.

[52] G. Heredia, A. Ollero, M. Bejar, and R. Mahtani, "Sensor and actuator fault detection in small autonomous helicopters," *Mechatronics*, vol. 18, no. 2, pp. 90–99, 2008.

[53] G. J. Prescott and P. H. Garthwaite, "A simple Bayesian analysis of misclassified binary data with a validation substudy," *Biometrics*, vol. 58, no. 2, pp. 454–458, 2002.

[54] V. R. Basili, R. W. Selby Jr., and T.-Y. Phillips, "Metric analysis and data validation across fortran projects," *IEEE Transactions on Software Engineering*, vol. 9, no. 6, pp. 652–663, 1983.

[55] K. R. Sundaram and A. Jose, "Teaching: estimation of minimum sample size and the impact of effect size and altering the type-I & II errors on IT, in clinical research," in *Data and Context in Statistics Education: Towards an Evidence-Based Society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*, C. Reading, Ed., International Statistical Institute, Voorburg, The Netherlands, 2010.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 7436841, 17 pages
http://dx.doi.org/10.1155/2016/7436841

Research Article

Faulty Line Selection Method for Distribution Network Based on Variable Scale Bistable System

Xiaowei Wang,[1] Jie Gao,[2] Guobing Song,[1] Qiming Cheng,[2] Xiangxiang Wei,[3] and Yanfang Wei[4]

[1]*School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi Province 710049, China*
[2]*College of Automation Engineering, Shanghai University of Electric Power, Shanghai, China*
[3]*College of Information and Electrical Engineering, China Agricultural University, Beijing, China*
[4]*School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China*

Correspondence should be addressed to Jie Gao; iamgaojie1993@163.com

Received 17 June 2016; Accepted 17 July 2016

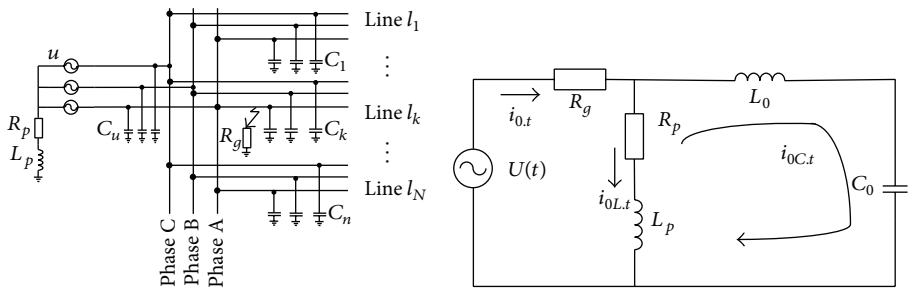
Academic Editor: Antonio Fernández-Caballero

Since weak fault signals often lead to the misjudgment and other problems for faulty line selection in small current to ground system, this paper proposes a novel faulty line selection method based on variable scale bistable system (VSBS). Firstly, VSBS is adopted to analyze the transient zero-sequence current (TZSC) with different frequency variety scale ratio and noise intensity, and the results show that VSBS can effectively extract the variation trends of initial stage of TZSC. Secondly, TZSC is input to VSBS for calculation with Runge-Kutta equations, and the output signal is chosen as the characteristic currents. Lastly, correlation coefficients of every line characteristic current are used as the index to a novel faulty line selection criterion. A large number of simulation experiments prove that the proposed method can accurately select the faulty line and extract weak fault signals in the environment with strong noise.

1. Introduction

As an important part of the power system, distribution network is closely associated with its users and also has direct impact on the users. Data show that 80% of fault occurring in distribution network is single phase-to-ground fault. When single phase-to-ground fault occurs, the line voltage value is still symmetrical, the fault current is weak, and it could run 1 to 2 hours after fault occurs, which significantly improves the reliability of power supply. However, during the single phase-to-ground fault period, nonfault phase voltage could rise, which will threaten the system insulation and result in interphase shortage, protection tripping, power supply outage, and other problems. Because of the weak fault signal and the harsh working condition, faulty line selection becomes difficult. Therefore, it is necessary to carry out further research in this area [1, 2].

At present, scholars have put forward various faulty line selection methods. Based on different characteristic components, faulty line selection methods for single phase-to-ground could be divided into 3 categories, that is, signal injection method [3], steady-state component method [4], and transient component method [5, 6]. The signal injection method needs additional signal device and its engineering realization is complex. In steady-state component method the characteristic signal is weak, which makes the result unreliable, while the transient characteristics method is more reliable and applicable because the transient characteristics component is larger than steady component and it will not be influenced by the arc suppression coil and it will need no additional devices [7–11]. Papers [7, 8] use wavelet transform to extract characteristic information for faulty line selection, but the wavelet transform is easily influenced by noise, and the chosen characteristic frequency band may be nonvalid



(a) The structure of small current to ground system

(b) TZSC equivalent circuit of single phase ground

Figure 1: The structure and TZSC equivalent circuit of small current to ground system.

transient faulty component. In addition, different wavelet basis function would lead to different extraction results and thus lead to error judgment. Paper [9] adopts Prony algorithm to fit TZSC signal when fault occurs. This method not only effectively avoids the effect of current transformer saturation flux density on collected signals but also improves the overall Prony fitting precision to a certain extent; but its calculation amount is large, its fitting order is difficult to determine, and the antinoise ability is not strong. In [10], the support vector machine shows its advantages in solving the problem of small sample, nonlinear, and high dimensional pattern recognition, but the recognition ability is easily influenced by its own parameters. Paper [11] uses Empirical Mode Decomposition (EMD) of TZSC to extract the five harmonic components in characteristic components and input them into Duffing oscillator to achieve faulty line selection according to the change of phase diagram. But when TZSC is greatly interfered with, modal aliasing phenomenon of EMD would arise and cause error judgment. Paper [12] employs the ratios of the first half-wave extreme and Spectral Kurtosis relative energy entropy from TZSC to build the stepped faulty line selection method. Paper [13] uses the S-transform to obtain the modulus and phase information of electric components at each frequency point, and this information is employed to detect the faulty line. In [14], Hilbert-Huang transform is used to decompose the TZSC, and then the most high frequency component of the intrinsic mode functions (IMF) can be obtained, and, based on this, the selection criterion is built; however, the decomposition process may cause modal aliasing. Paper [15] adopts evidence uncertainty reasoning and compared abnormal events to reduce computation amount and to improve the accuracy of faulty line selection. Paper [16] employs cross-correlation theory to calculate the integrated correlation coefficient of pure fault component of zero-modulus current for each line and takes the line with the smallest one as the faulty line. Paper [17] carries out the wavelet transform to decompose the transient zero-sequence current for each line, calculates the high and low frequency wavelet energies according to the wavelet coefficients, and selects the faulty line according to the maximum value of high or low frequency energy; however, in the strong noise background, the waveform and energy of weak TZSC will be affected.

In recent years, the research on stochastic resonance has made great progress. Stochastic resonance is a new practical technology which uses stochastic resonance principle to detect weak signal, and its research and application have spread into physical fields [18, 19], signal processing [20, 21], mechanical fault diagnosis [22], biology [23], neural network [24], and other academic fields; however, the research on this technology in power system is still needed. Therefore, with detailed study of the effect of TZSC on bistable system, this paper proposes a novel faulty line selection method for small current to ground system based on stochastic resonance theory. For signal feature extraction, the method employs VSBS to deal with TZSC and, then, choose the initial stage of output signal as characteristic current; for faulty line selection criterion, a novel faulty line selection criterion, which is based on cross-correlation coefficient sign, is proposed through calculating correlation coefficient of characteristic signal.

2. Characteristic Analysis of Single Phase-to-Ground Fault

The structure of small current to ground system is shown in Figure 1(a); when it experiences single phase-to-ground, the TZSC analysis circuit of faulty line is shown in Figure 1(b). In Figure 1, C_0 and L_0 are zero-sequence capacitance and inductance, respectively, R_g is transition resistance of ground point, R_p and L_p are, respectively, equivalent resistance and inductance of arc suppression coil, and $U(t)$ is zero-sequence voltage.

When distribution network fault occurs, from Figure 1(b), the TZSC flowing through the fault point $i_{0,t}$ is shown as [25]

$$\begin{aligned} i_{0,t} &= i_{0L,t} + i_{0C,t} \\ &= I_{Lm}\cos\varphi e^{-t/\tau_L} \\ &\quad + I_{Cm}\left(\frac{\omega_f}{\omega}\sin\varphi\sin\omega t - \cos\varphi\cos\omega_f t\right)e^{-\delta t}, \end{aligned} \tag{1}$$

where $i_{0L,t}$ and $i_{0C,t}$ are inductive current and capacitive current of TZSC, and its initial values are I_{Lm} and I_{Cm}, respectively ($I_{Cm} = U_{phm}\omega C$, $I_{Lm} = U_{phm}/\omega L$), U_{phm} is phase voltage amplitude, ω is angular frequency of power frequency,

ω_f and δ are oscillation angular frequency and attenuation coefficient of TZSC, τ_L is decay time constant of inductive current, and φ is initial phase angel.

From (1), when single phase-to-ground fault occurs in distribution network, the transient capacitance current has the characteristic of periodic attenuation oscillation. And [1] indicates that the free oscillation frequency of overhead line is within 300 Hz to 1500 Hz and the free oscillation frequency of cable lines is 1500 Hz~3000 Hz.

In addition, studies show that when single phase-to-ground fault occurs, the traveling wave pole is consistent with the overall changing trend of initial stage TZSC in transient process, so the mutation direction characteristic of initial stage TZSC can be used to replace the traveling wave polarity characteristic of TZSC, which can greatly reduce the hardware requirements and improve the reliability of faulty line selection [26]. Besides, whether it is big initial fault angle or small initial fault angle, the whole changing trend of faulty line is opposite to that of nonfaulty line in TZSC initial stage. But the introduction of arc suppression coil will greatly reduce ground fault current of distribution network, and when the fault occurs in voltage zero position, the changing trend of the initial stage TZSC is not easy to distinguish, which will make faulty line selection more difficult.

Hence, in some faulty conditions, it can be seen from the above analysis that TZSC of distribution network belongs to weak signal. As stochastic resonance (SR) theory has the unique advantage of amplification and detection of weak signals [27], it is helpful to employ the stochastic resonance theory to detect TZSC which is used to select the faulty line.

3. Signal-Detecting Ability of Variable Scale Bistable System

The bistable system for studying stochastic resonance is shown in [28]

$$s_p(t) = \frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{d\left(-ax^2/2 + bx^4/4\right)}{\mathrm{d}x} + s_i(t) + \Gamma(t), \quad (2)$$

where t is time, $s_i(t)$ is input signal, $\Gamma(t)$ is noise whose intensity is D, $s_p(t)$ is output signal, and x is the speed of Brownian particle.

According to Fokker-Planck equation, the probability distribution function of x is shown in (3) when $s_i(t)$ is $A\sin(2\pi f_0 t)$ and a and b are equal to μ and 1, where A and f_0 are the amplitude and frequency of periodic signal:

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[-\mu x - x^3 + A\sin\left(2\pi f_0 t\right)\rho(x,t)\right] + D\frac{\partial^2}{\partial x^2}\rho(x,t). \quad (3)$$

Since (3) has nonautonomous $-(\partial/\partial x)[A\sin(2\pi f_0 t)\rho(x, t)]$, it has no steady-state solution; that is, it can not have exact expression. However, in the adiabatic approximation condition with $A \ll 1$, $D \ll 1$, and $f_0 \ll 1$, the output signal-to-noise ratio of bistable system can be obtained and shown in

$$R_{\text{out}} = \frac{\sqrt{2}\mu^2 A^2 e^{-\mu^2/4D}}{4D^2}. \quad (4)$$

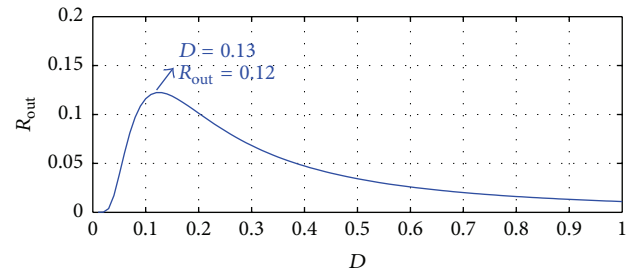


Figure 2: The curve of R_{out} with changed D.

Supposing μ is 1 and A is 0.2, the curve of R_{out} is shown in Figure 2 when D changed. The feature of Figure 2 is that, with the rise of D, R_{out} presents a trend of increasing and begins to decrease when D reached 0.13, which is the feature of stochastic resonance. So the bistable system can use noise to increase R_{out} of signal; that is, the weak signal is amplified and detected.

Under the small parameters of adiabatic approximation condition, the theoretical analysis of stochastic resonance of bistable system coincides with the numerical simulation of the bistable system [28]. However, it is improper to apply the method of small parameters stochastic resonance directly to the processing of signals with large parameters. Reference [27] introduces variable scale stochastic resonance to the process and gets better results; however, when the signal is TZSC, what change will happen to stochastic resonance feature of bistable system, and what rules can we get? This section will focus on the influence of each parameter on bistable system and try to figure out the VSBS characteristics under the effect of TZSC.

3.1. Variable Metric Algorithm and Its Evaluation Index. The principle of variable metric algorithm is to transform high frequency into low frequency in order to make the large parameter signal close to or meeting small parameters conditions of stochastic resonance, which means that the frequency is compressed and then detected by bistable system.

The Calculation Process of Variable Scale. According to the frequency and sampling frequency of signals, a frequency compression-scale ratio (CR) is determined, based on which the compression sampling frequency f_{sr} is defined ($f_{sr} = f_s/\text{CR}$). Then numerical calculation step h ($h = 1/f_{sr}$) is obtained from f_{sr}, and, finally, the response output of bistable system is numerically calculated.

Since TZSC are generally broadband signals and their frequency range is not confined to one or a small number of frequencies but distributed in a wide frequency band, to which the traditional signal-to-noise ratio measurement cannot be effectively applied, it is necessary to develop other measurement indexes.

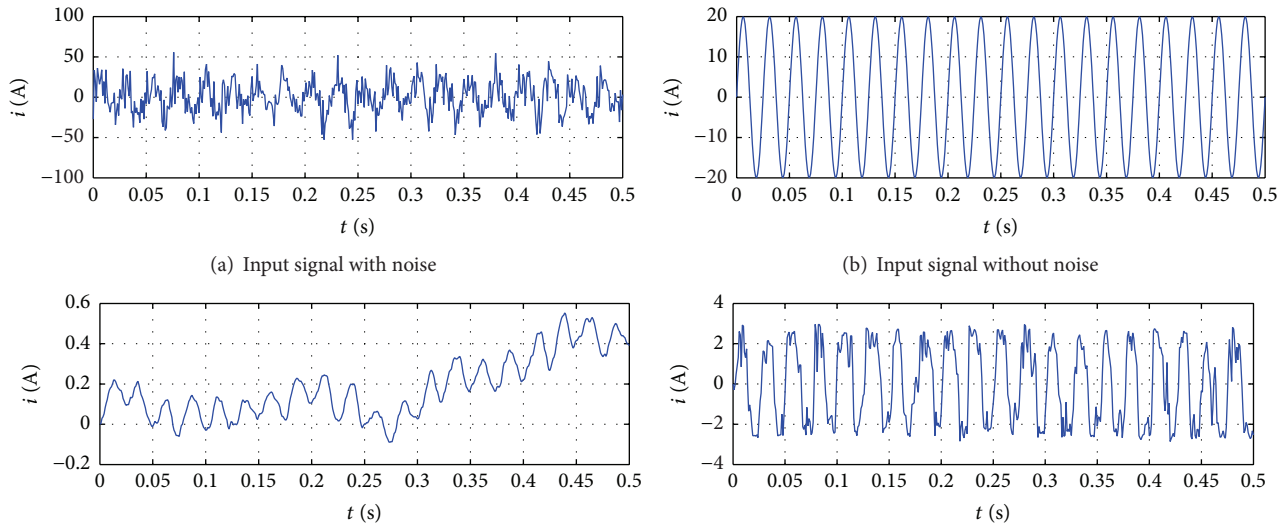


Figure 3: The simulation for bistable system.

Reference [27] shows that although nonlinear Langevin equation cannot accurately predict the motion of Brownian particles, it can well predict the statistical properties of the particle orbits. So, this paper uses cross-correlation coefficient as a measure to describe the response of VSBS whose input signal is weak aperiodic signal. The covariance $\mathrm{Cov}(s_p(t), s_i(t))$ and cross-correlation coefficients ρ_{pi} of two signals are shown in

$$\mathrm{Cov}\left(s_p(t), s_i(t)\right) = E_{(s_p \cdot s_i)} - E_{s_p} E_{s_i}$$

$$\rho_{pi} = \frac{\mathrm{Cov}\left(s_p(t), s_i(t)\right)}{\sqrt{D\left(s_p(t)\right)}\sqrt{D\left(s_i(t)\right)}}. \quad (5)$$

Additionally, in the initial faulty stage, the overall changing trend of TZSC of faulty line is opposite to that of nonfaulty line, so this paper focuses on the changing trend of input signal and output signal.

3.2. Simulation of Variable Scale Bistable System

3.2.1. Simulation of Nonintroducing Variable Scale. Supposing there is a set of measured signals, the sampling points are 500, the corresponding parameters of (2) are $a = b = 1$, $A = 20$ A, $f_0 = 40$ Hz, and $D = 100$ db, respectively, and the value of sampling frequency f_s is 1000 Hz. Fourth-order Runge-Kutta algorithm is adopted to calculate (2). And the value of cross-correlation coefficients ρ_{pi} of $s_i(t)$ and $s_p(t)$ is −0.0078, whose results are shown in Figure 3. Figure 3(a) shows the result of $s_i(t)$ with noise intensity as 100 db, Figure 3(b) shows $s_i(t)$ without noise, and Figure 3(c) shows $s_{p1}(t)$ without being solved by variable scale.

It can be known from Figure 3 and ρ_{pi} that when both weak signal frequencies f_0 and D are large parameters (larger than 1), the output and input of the system differ dramatically, and the information contained in the output signal will not be able to represent the original signal. That is why the stochastic resonance method with small parameters can not be directly applied to large parameters signal, so the detection is ineffective.

3.2.2. Simulation of Introducing Variable Scale. Bring in variable scale thought, choose CR as 100, $a = b = 1$, $A = 20$ A, $f_0 = 40$ Hz, $D = 100$ db, and the value of sampling frequency f_s is 1000 Hz; calculate (2) with fourth-order Runge-Kutta algorithm and cross-correlation coefficients ρ_{pi} of $s_i(t)$ and $s_p(t)$, $\rho_{p_2 i}$ will be obtained, and its value is 0.8088. The result is shown in Figure 3(d).

Figure 3(d) shows that after the treatment of VSBS, the waveform of output signal $s_p(t)$ becomes orderly. Compared to Figure 3(c), the cross-correlation coefficients between $s_i(t)$ and $s_p(t)$ have obviously improved as well as the amplitude value of $s_p(t)$; besides, $s_p(t)$ and $s_i(t)$ belong to strong correlation. Therefore, through frequency conversion, the disorganized large parameter signal is made clear and orderly; besides, ρ_{pi} is equal to 0.8088, which indicates that $s_p(t)$ could better represent changing trend of $s_i(t)$ submerged in noise, achieving the large parameter stochastic resonance or, exactly speaking, a kind of stochastic resonance.

3.3. Transient Zero-Sequence Current Detection.

In order to test whether VSBS can detect the TZSC, the ideal TZSC $i_z(t)$ [29] is defined as below:

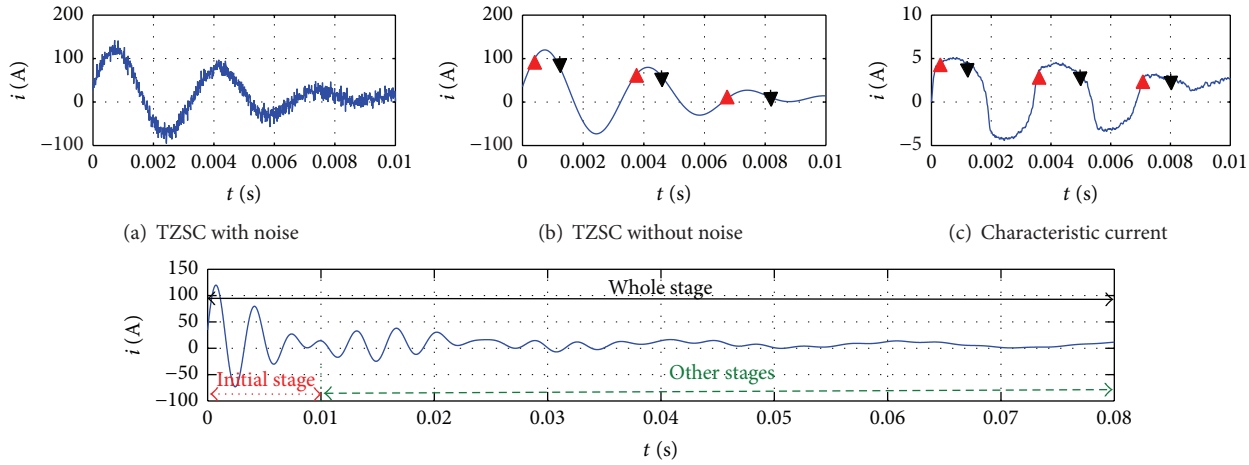
$$i_z(t) = x_1(t) + x_2(t) + x_3(t) + x_4(t) + \Gamma(t), \quad (6)$$

$$x_1(t) = 5.6\cos\left(2\pi \times 50t + 60^\circ\right), \quad (7)$$

$$x_2(t) = 40e^{-56t}\cos\left(2\pi \times 250t + 30^\circ\right), \quad (8)$$

$$x_3(t) = 72e^{-102t}\cos\left(2\pi \times 315t\right), \quad (9)$$

$$x_4(t) = 10e^{-5.5t}. \quad (10)$$



(a) TZSC with noise (b) TZSC without noise (c) Characteristic current

(d) TZSC without noise

Figure 4: Characteristic extraction of TZSC by VSBS.

It can be seen that $i_z(t)$, which consists of 5 signals, has the characteristic of multifrequency and attenuation; therefore it is a nonperiodic signal. Input it into (2), and its corresponding parameters are $a = b = 1$ and $D = 50$ db and sampling frequency is f_s =100000 Hz. CR is equal to 1000, and then the results of its numerical simulation are shown in Figure 4.

Definition of Characteristic Current $i_c(t)$. Characteristic current is the output signal obtained by solving VSBS with TZSC by fourth-order Runge-Kutta algorithm. Choose nonnoises $i_z(t)$ and $i_c(t)$ to calculate cross-correlation coefficient ρ_{cz}, and the value is 0.7628.

When only the first 0.01 s of $i_z(t)$ and $i_c(t)$ is chosen, as shown in Figure 4(a), the noise causes strong disturbance in the initial stage of $i_z(t)$, which makes the changing trend not so clear as the original signal. It is known from Figure 4(c) that, after VSBS treatment, the changing trend of $i_c(t)$ is similar to that of $i_z(t)$; then, their ρ_{cz} is calculated, and the value has improved to 0.8909. Therefore, VSBS can effectively extract TZSC changing trend of the initial stage.

This method can be used to better extract the change trend of TZSC in the initial stage. This paper defines 0~0.01 s as the initial stage of TZSC, 0.01 s~∞ as noninitial stage, and signal length as the whole stage. To put it vividly, TZSC from (6) is chosen as the label, and the results are shown in Figure 4(d).

3.4. The Detection Adaptability of TZSC. In order to test detection adaptability of VSBS for TZSC, the paper will analyze frequency compression-scale ratio, noise intensity, the initial value, and signal amplitude, respectively.

3.4.1. Frequency Compression-Scale Ratio (CR). Set $a = b = 1$ and $D = 50$ db and sampling frequency f_s equals 100000 Hz; set CR as 10, 100, 1000, and 5000, respectively, and the change of ρ_{cz} is shown in Table 1.

It can be seen from Table 1 that, with the increase of CR, ρ_{cz} between $i_z(t)$ and $i_c(t)$ first increases and then decreases;

Table 1: ρ_{cz} in different conditions.

Condition	Whole stage ρ_{cz}	Initial stage ρ_{cz}
CR = 10	−0.2220	−0.3395
CR = 100	0.4716	0.5300
CR = 1000	0.7628	0.8909
CR = 5000	0.6784	0.7944
$D = 0$	0.7638	0.8874
$D = 50$	0.7628	0.8909
$D = 100$	0.7518	0.8876
$D = 1000$	0.6470	0.8641
$D = 5000$	0.4286	0.7910
IV = 34.8	0.7628	0.8909
IV = 0	0.8806	0.9221
IV = −34.8	0.8555	0.8914
$\tau = 1/100$	0.4807	−0.0234
$\tau = 1/10$	0.7403	0.1820
$\tau = 1$	0.9221	0.7130
$\tau = 10$	0.7146	0.9916
$\tau = 100$	0.7150	0.9922

the reason is that the increase of CR can gradually compress the frequency band range of $i_z(t)$ into VSBS's detection range, and there may be a most suitable CR making the input and output most relevant, but when CR continues to increase, excessive frequency compression will also lead to decrease of gap between different frequencies of $i_z(t)$, showing the reduction of frequency species, which will further weaken the detection ability of VSBS. In addition, the calculation of cross-correlation coefficients of $i_z(t)$ and $i_c(t)$ in initial stage shows that ρ_{cz} has greatly improved, and the changing trend of $i_c(t)$ is the same as that of $i_z(t)$, which verifies that VSBS can effectively extract the changing trend of TZSC in initial stage.

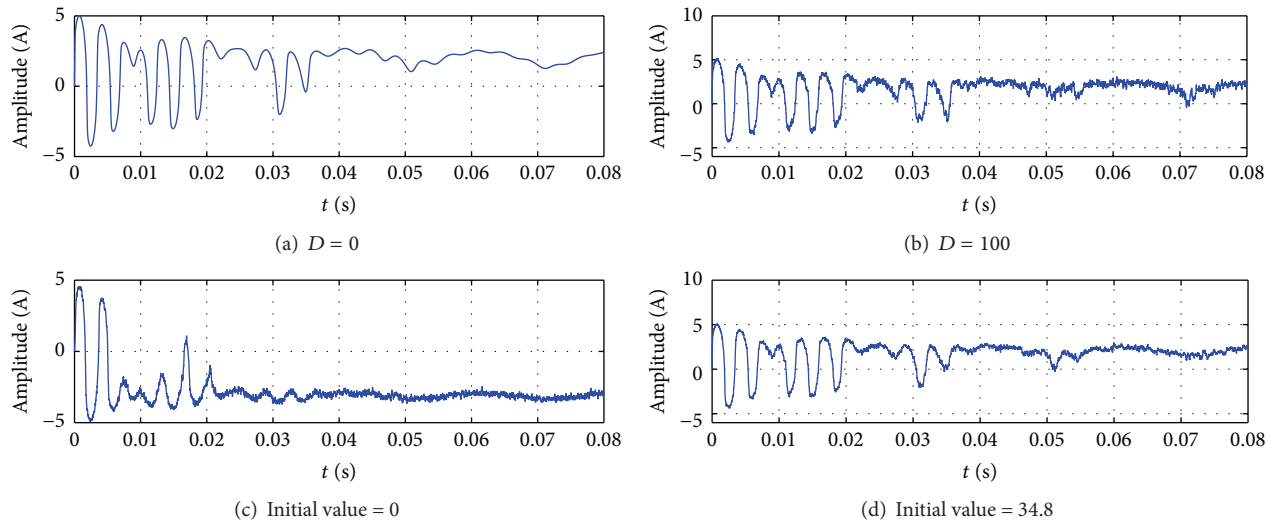


Figure 5: Characteristic current under the different conditions.

3.4.2. Noise Intensity (D). Set $a = b = 1$, sampling frequency $f_s = 100000$ Hz, and CR = 1000. Set noise intensity D as 0 db, 50 db, 100 db, 1000 db, and 5000 db, respectively, and the change of cross-correlation coefficient ρ_{cz} is shown in Table 1. The change of characteristic current $i_c(t)$ is shown in Figure 5(a) when the noise intensity of $i_z(t)$ is 0 db, and the change of characteristic current $i_c(t)$ is shown in Figure 5(b) when the noise intensity of $i_z(t)$ is 100 db.

It can be seen from waveform in Figure 5 that the amplitude of $i_c(t)$ increases with the increase of D, which means that part of the noise energy is transferred to $i_c(t)$ [27]. In addition, the increases of D made the latter part waveform disorderly; however, the changing trend of the waveform in initial stage is clear which shows no difference with the changing trend of nonnoise; therefore, this once again shows that the VSBS can extract initial stage TZSC. From ρ_{cz} in Table 1 we know that, within a certain range of noise intensity, the increase of D shows little effect on ρ_{cz} of whole stage and of initial stage, which indicates that VSBS can well extract changing trend of TZSC in initial stage with the disturbance of strong noise. However, excessive noise intensity will produce a wide range of interference frequency components, which will affect the existence of the original signal, resulting in the decrease of cross-correlation coefficient and fuzziness of the changing trend in initial stage. It is worth noting that when the noise intensity is 0, the VSBS can also predict the changing trend of TZSC.

3.4.3. Signal Initial Value (IV). Set $a = b = 1$, $f_s = 100000$ Hz, CR = 1000, and $D = 50$ db; set initial value of $i_z(t)$ as 0 and 34.8, respectively, where 34.8 is initial value (IV) of $i_z(t)$. Then carry out numerical simulation and the change of characteristic current $i_c(t)$ and cross-correlation coefficient ρ_{cz} is shown in Table 1 and Figure 5. Figure 5(c) is characteristic current $i_c(t)$ when initial value of $i_z(t)$ is 0, and Figure 5(d) is characteristic current $i_c(t)$ when initial value of $i_z(t)$ is 34.8.

It can be known from Figures 5(c) and 5(d) and Table 1 that when initial value of $i_z(t)$ is 0, the change trend of $i_c(t)$ is closest to that of $i_z(t)$, especially in the initial stage. Either increase or decrease of the initial value will decrease ρ_{cz}, because when initial value of VSBS is 0, any tiny disturbance is likely to cause it to move in the double potential well with large amplitude, so it can better reflect the moving trend of signals. However, when the initial value is too large, tiny disturbance may not be enough to cause a large amplitude motion in the double well potential or only small range of motion in a single potential well, therefore, it will weaken the detection ability of VSBS, and this is consistent with the decrease of ρ_{cz} in Table 1. Another reason for adopting TZSC to select faulty line in this paper is that when the initial value of TZSC is 0, VSBS and faulty line selection can be better combined [30].

3.4.4. Signal Amplitude. Set $a = b = 1$, $f_s = 100000$ Hz, CR = 1000, and $D = 50$ db; set the initial value as 0, increase the amplitude of $i_z(t)$, and set the amplitude factor τ as 1/100, 1/10, 1, 10, and 100, respectively.

It is known from Table 1 that, with the increase of amplitude $i_z(t)$, cross-correlation coefficient ρ_{cz} of whole stage and initial stage first increases and then decreases, but ρ_{cz} of noninitial stage increases. The reason is that $i_z(t)$ belongs to damped oscillation signal, and, compared with that of noninitial stage, the amplitude values of initial stage are always larger, and the detection ability of VSBS on $i_z(t)$ in noninitial stage is weaker. Therefore, appropriate increase of signal value would improve detection ability of VSBS.

Based on the above analysis, the features of BSVS detecting TZSC are summarized as follows:

(1) Appropriate frequency compression ratio can improve the signal detection performance.

(2) For small amplitude signal, appropriate increase of amplitude can improve signal detection performance.

(3) For the signal with zero initial value, VSBS has a better detection performance of changing trend in its initial stage.

(4) The cross-correlation coefficient of whole stage is always smaller than that of initial stage.

4. Faulty Line Selection Method

4.1. Parameter Setting. Based on the above analysis, this paper will select faulty line according to the following characteristics of VSBS detecting TZSC:

① The overall changing trend of TZSC in initial stage between faulty line and nonfaulty line is opposite.

② VSBS has excellent detection ability for the changing trend of TZSC in initial stage.

③ When single phase-to-ground fault occurs in small current to ground system, the free oscillation frequency of overhead lines is generally 300 Hz~1500 Hz, while the free oscillation frequency of cable lines is 1500 Hz~3000 Hz. In addition, different fault conditions may cause the TZSC spectrum to transfer into low frequency band [25].

④ Appropriate increase of signal amplitude helps to improve detection performance of VSBS.

According to ① and ②, this paper will focus on cross-correlation coefficient of different lines. Since the TZSC before failure is 0, when calculating cross-correlation coefficients, we will choose $T/4$ cycle after fault as the initial stage (0.02 s~0.025 s) in the paper; with ③ frequency varieties are compressed as much as possible to make frequency varieties into the frequency range which VSBS can detect, in order to enhance the adaptability of the method; therefore, the frequency compression-scale ratio (CR) is set as 1500. Based on ④ and simulation experiment, when the maximum amplitude of signal is less than 5, we first expand the amplitude by 10 times and then input it into VSBS. In addition, we find that TZSC amplitude before fault is not 0 but a very small value, which needs to be set to 0 before fault.

4.2. Pretreatment of Faulty Line Selection. Take line l_1 as an example:

① Choose TZSC of line number 1 from one cycle before fault to one cycle after fault as TZSC $i_z(t)$, and set the signal one cycle before fault as 0.

② Judge whether the maximum amplitude of $i_z(t)$ is smaller than 5, if it is, carry out ③, and if it is not, carry out ④.

③ Expand amplitude of $i_z(t)$ by 10 times and input it into VSBS to calculate characteristic current $i_c(t)$.

④ Directly input $i_z(t)$ into VSBS to calculate characteristic current $i_c(t)$.

Calculate $i_c(t)$ of all lines according to the above steps.

4.3. Steps of Faulty Line Selection. Take l_j as an example to explain the steps of faulty line selection.

Step 1. Calculate cross-correlation coefficients ρ_{jq} between line number l_j ($j = 1, 2, 3, \ldots$) and other lines l_q ($q = 1, 2, 3, \ldots q \neq j$) in initial stage.

Step 2. Count positive and negative signs of calculated ρ_{jq}:

(1) If all the signs of ρ_{jq} are the same, the output is "−1" and line l_j is judged as faulty line.

(2) If all the signs of ρ_{jq} are not the same, the output is "1" and line l_j is judged as nonfaulty line.

The remaining lines are also judged with the same steps.

5. Case Study

5.1. Simulation Model. ATP-EMTP is used in this paper to simulate the single phase-to-ground fault. The simulation model is shown in Figure 6 and the parameters of simulation model are the same as [31].

5.2. Simulation Results with Changing Phase and Resistance. Build the simulation model according to the parameters, make fault of line l_1 occur at the point 5 km from the bus, and change the initial fault angle (0°, 30°, 60°, and 90°) and ground resistance for simulation. It is known from [25] that when single phase-to-ground fault of the small current to ground system occurs, the fault resistance value is generally 0 kΩ to 2 kΩ; therefore, the maximum fault resistance is set as 2 kΩ in the paper. Then, select the faulty line according to the proposed method, and the parameters of VSBS are as follows: $a = b = 1$ and CR = 1500. In addition, we use (l_1, 0°, and 300 Ω) to indicate the fault occurrence in line number 1 when its initial angle is 0° and faulty resistance is 300 Ω. The results of their specific cross-correlation coefficients are shown in Table 2 in which ρ_{12} represents cross-correlation coefficient of characteristic currents l_1 and l_2.

The paper takes (l_1, 90°, and 2000 Ω) as an example to explicitly show the consequence of VSBS. Under this fault condition, the TZSC of l_2 and $i_c(t)$ of l_2 are shown in Figures 7(a) and 7(b); the cross-correlation coefficients are shown in Table 3.

The comparison of Figures 7(a) and 7(b) shows that, after the treatment of VSBS, the changing speed of $i_c(t)$ in initial stage becomes slow, and oscillation part is reduced, which makes the changing trend of $i_c(t)$ in initial stage easier to identify than that of TZSC in initial stage. It can be seen from Figure 7(b) that $i_c(t)$ waveform of faulty line is steadier than that of nonfaulty lines, because frequency compression makes the part in low frequency band and with strong intensity easy to detect by VSBS, and the part in high frequency band and with weak intensity may be ignored, besides, the intensity of low frequency band fault transient component in the faulty line is much larger than that of nonfaulty line, and, therefore, the characteristic current waveform of faulty line is steadier than that of nonfaulty line. Besides, through simulation, we discover that because the transient

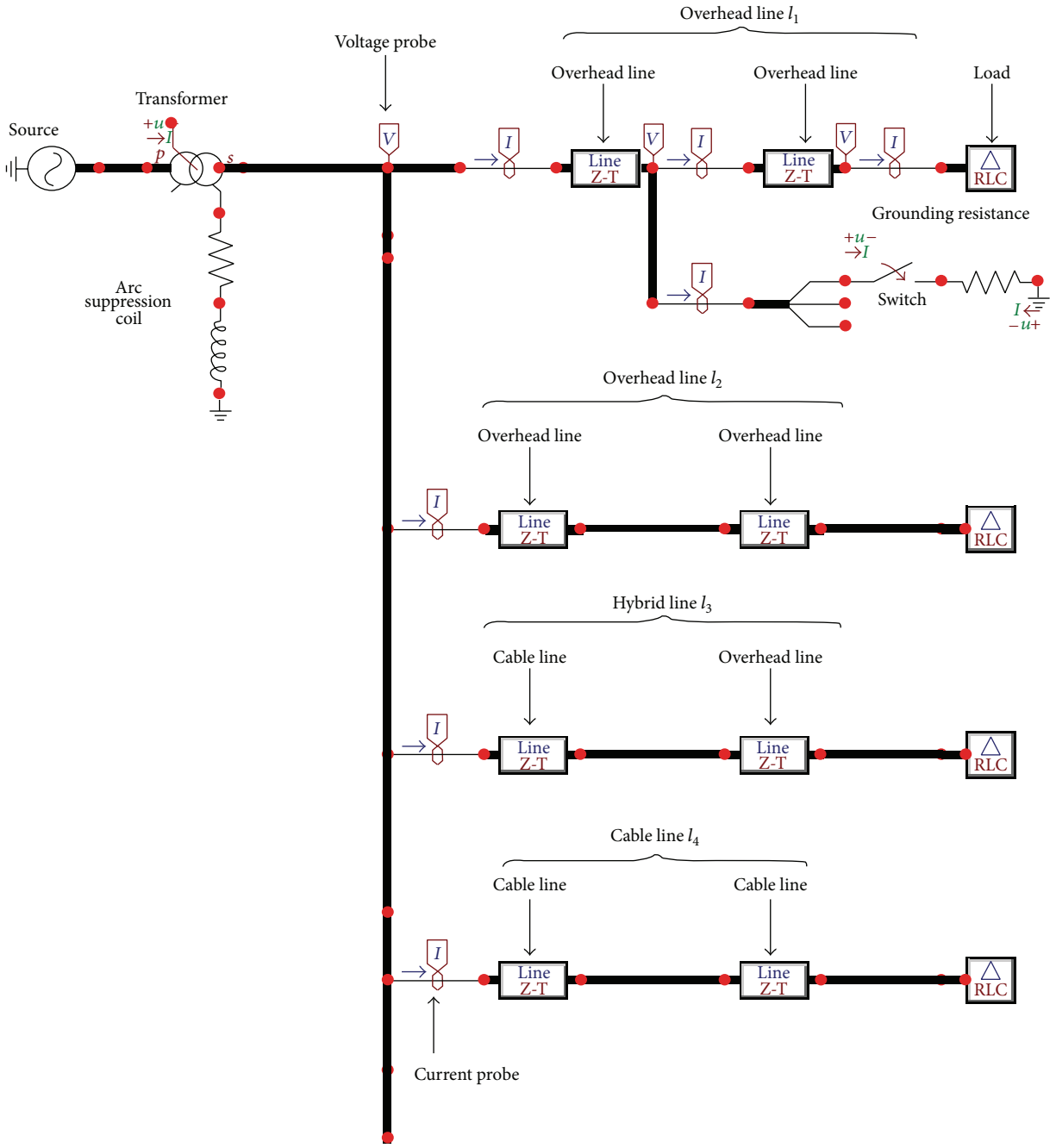


FIGURE 6: ATP simulation model.

free oscillation components and zero-sequence steady-state components are offset, the increase of fault resistance also makes waveform $i_c(t)$ of each line steadier.

It can be seen from Table 3 that ρ_{1q} of l_1 and other lines are equal to −0.9732, −0.8092, and −0.7535, respectively; since all of them are the same sign, the output is "−1"; ρ_{2q} of l_2 and other lines are equal to −0.9732, 0.7408, and 0.6597, respectively; since all of them are not the same sign, the output is "1"; therefore, we judged l_1 as faulty line, and the judging result is consistent with simulation.

In summary, we know from Table 2 that TZSC of various fault conditions is, respectively, input into VSBS, and the output judging results are consistent with actual fault situations. Therefore, the proposed method can accurately achieve faulty line selection under different fault resistance and initial fault angle.

5.3. Simulation Results of Fault with Random Gauss White Noise Added. Since signals collected in actual system with fault often carry noise with them, to verify the antinoise

Table 2: Results of different initial angle and ground resistance.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 10 Ω)	**−0.4942**	**−0.8054**	**−0.9092**	0.7852	0.6078	0.7973	l_1 fault
	(0°, 100 Ω)	**−0.6776**	**−0.7360**	**−0.7528**	0.8573	0.7034	0.8744	l_1 fault
	(0°, 500 Ω)	**−0.5710**	**−0.7146**	**−0.6549**	0.8999	0.8369	0.9032	l_1 fault
	(0°, 1000 Ω)	**−0.5991**	**−0.3511**	**−0.7291**	0.7810	0.7424	0.6944	l_1 fault
	(0°, 1500 Ω)	**−0.6845**	**−0.2796**	**−0.3330**	0.7057	0.5993	0.8385	l_1 fault
	(0°, 2000 Ω)	**−0.6928**	**−0.5382**	**−0.5511**	0.6465	0.5293	0.8454	l_1 fault
	(30°, 10 Ω)	**−0.5835**	**−0.8295**	**−0.9360**	0.8132	0.6362	0.7710	l_1 fault
	(30°, 100 Ω)	**−0.7767**	**−0.8968**	**−0.8948**	0.8668	0.7265	0.8485	l_1 fault
	(30°, 500 Ω)	**−0.7987**	**−0.8858**	**−0.8773**	0.8746	0.8609	0.9254	l_1 fault
	(30°, 1000 Ω)	**−0.7353**	**−0.6373**	**−0.7927**	0.8598	0.9000	0.7555	l_1 fault
	(30°, 1500 Ω)	**−0.6972**	**−0.5462**	**−0.5819**	0.8233	0.7478	0.8720	l_1 fault
	(30°, 2000 Ω)	**−0.7340**	**−0.4746**	**−0.4987**	0.8116	0.7052	0.8649	l_1 fault
	(60°, 10 Ω)	**−0.6113**	**−0.8418**	**−0.9606**	0.8201	0.6109	0.7682	l_1 fault
	(60°, 100 Ω)	**−0.8523**	**−0.9398**	**−0.9703**	0.9080	0.7461	0.8808	l_1 fault
	(60°, 500 Ω)	**−0.9027**	**−0.9822**	**−0.9407**	0.9268	0.9015	0.9374	l_1 fault
	(60°, 1000 Ω)	**−0.8862**	**−0.7425**	**−0.9420**	0.8639	0.9368	0.7823	l_1 fault
	(60°, 1500 Ω)	**−0.9137**	**−0.7322**	**−0.7302**	0.8450	0.7847	0.8721	l_1 fault
	(60°, 2000 Ω)	**−0.8855**	**−0.9129**	**−0.8872**	0.8027	0.7284	0.8894	l_1 fault
	(90°, 10 Ω)	**−0.9293**	**−0.8728**	**−0.9691**	0.9117	0.8795	0.7960	l_1 fault
	(90°, 100 Ω)	**−0.9464**	**−0.9790**	**−0.9820**	0.9549	0.8879	0.9346	l_1 fault
	(90°, 500 Ω)	**−0.9714**	**−0.8578**	**−0.9932**	0.8990	0.9487	0.8150	l_1 fault
	(90°, 1000 Ω)	**−0.9846**	**−0.8524**	**−0.8428**	0.8467	0.7932	0.8995	l_1 fault
	(90°, 1500 Ω)	**−0.9812**	**−0.8457**	**−0.7970**	0.8033	0.7265	0.9038	l_1 fault
	(90°, 2000 Ω)	**−0.9732**	**−0.8092**	**−0.7535**	0.7408	0.6597	0.9216	l_1 fault

Table 3: Cross-correlation coefficient with (l_1, 90°, and 2000 Ω) fault situation.

Faulty line	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Result
l_1	−0.9732	−0.8092	−0.7535	0.7408	0.6597	0.9216	l_1

performance of the method, we added 0.5 db or −0.5 db noise intensity to TZSC when fault in different line occurred and set signal before fault as 0. The selection results and specific cross-correlation coefficients are shown in Tables 4 and 5, respectively.

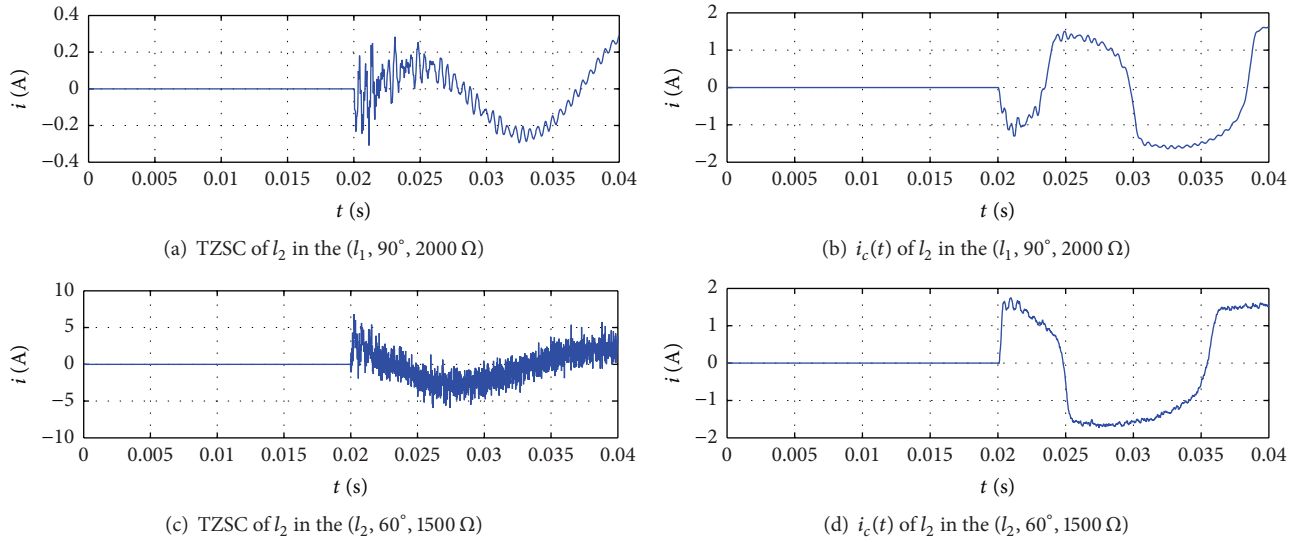
Signal-to-noise ratio equaling −0.5 db and fault situations as (l_2, 60°, 1500 Ω) are taken as an illustration. The TZSC with noise of l_2 and $i_c(t)$ of l_2 are shown in Figures 7(c) and 7(d), and the cross-correlation coefficients are shown in Table 6.

Firstly, from the faulty line selection method and Table 6, cross-correlation coefficients ρ_{2q} of l_2 and other lines are all negative, so the output is "−1" and ρ_{jq} of other lines are different, so the output is "1"; therefore, we judge line l_2 as faulty line, which is consistent with actual fault situation. Then, comparison of Figures 7(c) and 7(d) shows that, with the disturbance of strong noise, even if the TZSC of each line is submerged in strong noise, the proposed method is still able to effectively extract the changing trend of TZSC in initial stage and can accurately judge the faulty line. Finally, Tables 4 and 5 indicate that, with the disturbance of different noise intensity, the changing trends of characteristic currents in initial stage between faulty line and nonfaulty line still have a better discrimination after the treatment of VSBS, so we can say the method shows a good antinoise performance. The definition of signal-to-noise ratio [28] shows that the smaller the ratio is, the larger the noise intensity will be; therefore, antinoise performance of the method in this paper is much better than the one proposed in [8], which added 15 db and 0.5 db noise.

5.4. Adaptation Analysis of Faulty Line Selection Method

5.4.1. Different Faulty Lines. When fault occurs in l_2, l_3, and l_4, respectively, we carry out faulty line selection method proposed in the paper to verify its adaptability, and the results and specific cross-correlation coefficients are shown in Table 7.

We know from [25] that, with the introduction of cable lines, although the attenuation process of fault transient current becomes shorter, the frequency spectrum principal component of transient component will move to low frequency band, which helps to detect VSBS. Therefore, different line fault conditions will not affect selection results of the method, and excellent results can also be obtained with different fault resistance.



(a) TZSC of l_2 in the (l_1, 90°, 2000 Ω)

(b) $i_c(t)$ of l_2 in the (l_1, 90°, 2000 Ω)

(c) TZSC of l_2 in the (l_2, 60°, 1500 Ω)

(d) $i_c(t)$ of l_2 in the (l_2, 60°, 1500 Ω)

Figure 7: The TZSC and characteristic current of l_2 for different fault condition.

Table 4: Adding 0.5 db Gauss white noise.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 300 Ω)	**−0.5426**	**−0.7043**	**−0.7148**	−0.4045	0.3447	0.9060	l_1 fault
	(90°, 1300 Ω)	**−0.7183**	**−0.8165**	**−0.9733**	0.5301	0.6960	0.7200	l_1 fault
l_2	(30°, 400 Ω)	**−0.7562**	0.8406	0.8673	**−0.8435**	**−0.8573**	0.9282	l_2 fault
	(60°, 1500 Ω)	**−0.5732**	0.5756	0.5700	**−0.9466**	**−0.9019**	0.8969	l_2 fault
l_3	(60°, 200 Ω)	0.6424	**−0.7439**	0.7361	**−0.7997**	0.5996	**−0.8890**	l_3 fault
	(90°, 2000 Ω)	0.4531	**−0.1279**	0.1332	**−0.4213**	0.4255	**−0.9950**	l_3 fault
l_4	(0°, 600 Ω)	0.1898	0.2342	**−0.1300**	0.8282	**−0.0314**	**−0.3755**	l_4 fault
	(30°, 1700 Ω)	0.8354	0.8258	**−0.1221**	0.8843	**−0.1130**	**−0.0464**	l_4 fault

Table 5: Adding −0.5 db Gauss white noise.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 300 Ω)	**−0.5919**	**−0.6906**	**−0.6985**	0.8092	0.7142	0.9197	l_1 fault
	(90°, 1300 Ω)	**−0.4888**	**−0.8292**	**−0.9956**	−0.5549	0.4627	0.8320	l_1 fault
l_2	(30°, 400 Ω)	**−0.7433**	0.8314	0.8227	**−0.8555**	**−0.8542**	0.9247	l_2 fault
	(60°, 1500 Ω)	**−0.6892**	0.5295	0.7051	**−0.8987**	**−0.8958**	0.8216	l_2 fault
l_3	(60°, 200 Ω)	0.5730	**−0.4954**	0.3834	**−0.8823**	0.6574	**−0.8804**	l_3 fault
	(90°, 2000 Ω)	0.3280	**−0.5827**	0.6395	**−0.5173**	0.4422	**−0.9832**	l_3 fault
l_4	(0°, 600 Ω)	0.6865	0.7746	**−0.2193**	0.8570	**−0.0483**	**−0.4239**	l_4 fault
	(30°, 1700 Ω)	0.3201	0.6745	**−0.3900**	0.6954	**−0.5253**	**−0.7236**	l_4 fault

Table 6: Cross-correlation coefficients with (l_2, 60°, and 1500 Ω) fault situation.

Faulty line	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Result
l_2	**−0.6892**	0.5295	0.7051	**−0.8987**	**−0.8958**	0.8216	l_2

5.4.2. Different Fault Distance. Since the distance of fault point is different in actual fault situations, we carry out simulation of line l_1, with different distance from the bus line, and the fault distance is set as 4.5 km, 7.5 km, 10.5 km, and 13.5 km, respectively. Select the faulty line with the method and the results are shown in Table 8, with specific cross-correlation coefficients shown in Table 8.

It can be seen that the selection results are consistent with actual fault situation, which indicates that the method can also achieve faulty line selection of fault with different distance situation, especially with high ground resistance in the end of line.

Table 7: Results of fault in different line.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_2	(0°, 200 Ω)	**−0.6531**	0.9323	0.9305	**−0.6822**	**−0.6978**	0.9235	l_2 fault
	(0°, 1200 Ω)	**−0.7730**	0.4276	0.5390	**−0.2614**	**−0.3268**	0.6480	l_2 fault
	(30°, 300 Ω)	**−0.8426**	0.9380	0.9252	**−0.8706**	**−0.8636**	0.9173	l_2 fault
	(30°, 1600 Ω)	**−0.7976**	0.4731	0.5823	**−0.4084**	**−0.4836**	0.6704	l_2 fault
	(60°, 50 Ω)	**−0.9389**	0.9236	0.9220	**−0.9775**	**−0.9866**	0.9468	l_2 fault
	(60°, 500 Ω)	**−0.9552**	0.9591	0.9292	**−0.9702**	**−0.9210**	0.9186	l_2 fault
	(90°, 100 Ω)	**−0.9745**	0.9646	0.9611	**−0.9881**	**−0.9833**	0.9640	l_2 fault
	(90°, 2000 Ω)	**−0.3466**	0.3401	0.2875	**−0.9268**	**−0.9663**	0.8184	l_2 fault
l_3	(0°, 200 Ω)	0.2988	**−0.2088**	0.7533	**−0.5504**	0.3707	**−0.4434**	l_3 fault
	(0°, 1200 Ω)	0.4411	**−0.5978**	0.4780	**−0.1951**	0.4098	**−0.1930**	l_3 fault
	(30°, 300 Ω)	0.3693	**−0.4231**	0.7843	**−0.5670**	0.3883	**−0.6595**	l_3 fault
	(30°, 1600 Ω)	0.5930	**−0.5245**	0.5808	**−0.2329**	0.5754	**−0.3471**	l_3 fault
	(60°, 50 Ω)	0.5844	**−0.5267**	0.4706	**−0.8439**	0.7423	**−0.9695**	l_3 fault
	(60°, 500 Ω)	0.5631	**−0.7989**	0.8680	**−0.5192**	0.5302	**−0.8409**	l_3 fault
	(90°, 100 Ω)	0.6357	**−0.7711**	0.6803	**−0.6946**	0.5872	**−0.9779**	l_3 fault
	(90°, 2000 Ω)	0.7254	**−0.3280**	0.2684	**−0.7031**	0.6408	**−0.9910**	l_3 fault
l_4	(0°, 200 Ω)	0.8296	0.8473	**−0.0548**	0.9663	**−0.0947**	**−0.1597**	l_4 fault
	(0°, 1200 Ω)	0.7250	0.5219	**−0.4825**	0.8572	**−0.4005**	**−0.1097**	l_4 fault
	(30°, 300 Ω)	0.8040	0.8457	**−0.1849**	0.9636	**−0.2711**	**−0.2791**	l_4 fault
	(30°, 1600 Ω)	0.7594	0.5734	**−0.4455**	0.8304	**−0.3052**	**−0.1994**	l_4 fault
	(60°, 50 Ω)	0.6962	0.4713	**−0.1996**	0.7384	**−0.2860**	**−0.7616**	l_4 fault
	(60°, 500 Ω)	0.8955	0.9260	**−0.4905**	0.9621	**−0.5296**	**−0.5718**	l_4 fault
	(90°, 100 Ω)	0.8089	0.8493	**−0.8635**	0.9817	**−0.9673**	**−0.9873**	l_4 fault
	(90°, 2000 Ω)	0.8014	0.3311	**−0.4829**	0.6891	**−0.8129**	**−0.9470**	l_4 fault

Table 8: Results of fault with different distance.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1 (4.5 km)	(0°, 10 Ω)	**−0.4489**	**−0.8020**	**−0.8974**	0.7467	0.5894	0.7566	l_1 fault
	(0°, 100 Ω)	**−0.6667**	**−0.7359**	**−0.7554**	0.8437	0.6847	0.8680	l_1 fault
	(0°, 500 Ω)	**−0.5742**	**−0.7155**	**−0.6577**	0.9012	0.8383	0.9043	l_1 fault
	(0°, 1000 Ω)	**−0.6003**	**−0.3521**	**−0.7290**	0.7853	0.7477	0.6978	l_1 fault
	(0°, 1500 Ω)	**−0.6861**	**−0.2790**	**−0.3324**	0.7096	0.6051	0.8424	l_1 fault
	(0°, 2000 Ω)	**−0.7598**	**−0.2504**	**−0.2922**	0.6492	0.5343	0.8500	l_1 fault
l_1 (7.5 km)	(0°, 10 Ω)	**−0.6947**	**−0.9639**	**−0.9391**	0.7768	0.7641	0.9503	l_1 fault
	(0°, 100 Ω)	**−0.7154**	**−0.7370**	**−0.7452**	0.9116	0.7875	0.9062	l_1 fault
	(0°, 500 Ω)	**−0.5624**	**−0.7148**	**−0.6455**	0.8987	0.8371	0.8998	l_1 fault
	(0°, 1000 Ω)	**−0.5985**	**−0.3462**	**−0.7319**	0.7675	0.7272	0.6817	l_1 fault
	(0°, 1500 Ω)	**−0.6824**	**−0.2807**	**−0.3355**	0.6925	0.5808	0.8271	l_1 fault
	(0°, 2000 Ω)	**−0.7538**	**−0.2572**	**−0.2988**	0.6381	0.5127	0.8316	l_1 fault
l_1 (10.5 km)	(0°, 10 Ω)	**−0.5543**	**−0.9069**	**−0.9374**	0.7530	0.6665	0.9332	l_1 fault
	(0°, 100 Ω)	**−0.7379**	**−0.7404**	**−0.7442**	0.9511	0.8637	0.9347	l_1 fault
	(0°, 500 Ω)	**−0.5586**	**−0.7198**	**−0.6388**	0.9008	0.8435	0.8981	l_1 fault
	(0°, 1000 Ω)	**−0.6028**	**−0.3422**	**−0.7371**	0.7592	0.7194	0.6716	l_1 fault
	(0°, 1500 Ω)	**−0.6842**	**−0.2810**	**−0.3374**	0.6827	0.5661	0.8183	l_1 fault
	(0°, 2000 Ω)	**−0.7533**	**−0.2619**	**−0.3028**	0.6323	0.4989	0.8188	l_1 fault
l_1 (13.5 km)	(0°, 10 Ω)	**−0.9348**	**−0.9142**	**−0.9543**	0.9721	0.9623	0.9682	l_1 fault
	(0°, 100 Ω)	**−0.7530**	**−0.7467**	**−0.7521**	0.9639	0.9094	0.9508	l_1 fault
	(0°, 500 Ω)	**−0.5585**	**−0.7284**	**−0.6356**	0.9039	0.8536	0.8981	l_1 fault
	(0°, 1000 Ω)	**−0.6121**	**−0.3380**	**−0.7439**	0.7569	0.8223	0.6670	l_1 fault
	(0°, 1500 Ω)	**−0.6914**	**−0.2798**	**−0.3368**	0.6780	0.5608	0.8176	l_1 fault
	(0°, 2000 Ω)	**−0.7582**	**−0.2654**	**−0.3038**	0.6312	0.4937	0.8124	l_1 fault

Table 9: Selection results of different time length.

Time length	Sample size	Accuracy
0.005 s	112	110/112
0.01 s	112	107/112
0.015 s	112	71/112

5.4.3. Influence of Different Initial Stage Length on Selection Accuracy. From the moment of fault occurrence, choose time length of different initial stage as 0~0.005 s, 0~0.01 s, and 0~0.015 s, initial fault angle as 0°, 30°, 60°, and 90°, respectively, fault resistance as 10 Ω, 50 Ω, 100 Ω, 500 Ω, 1000 Ω, 1500 Ω, and 2000 Ω, and faulty line as l_1, l_2, l_3, and l_4, respectively, that is, a total of $4 \times 7 \times 4 = 112$ different fault conditions. Then, use the method proposed in the paper to carry out faulty line selection, and the selection results as shown in Table 9.

Table 9 shows that time length of the initial stage can affect the selection accuracy. The longer the length of initial stage is, the lower the selection accuracy will be. The main reasons are as follows:

(1) VSBS can well detect the changing trend of TZSC in initial stage, besides, TZSC is an oscillation attenuation signal whose initial value is 0, therefore, with smaller time length, and the cross-correlation coefficient has better representation.

(2) When single phase-to-ground fault occurs in distribution network, TZSC of each line will increase suddenly, and the TZSC mutation direction between faulty line and nonfaulty line is opposite. However, in the following $T/4$ time period, this situation will not happen, so the increase of signal length will affect the overall changing trend of the signal and ρ_1 and, then, cause wrong judgment.

(3) It is known from Section 3 that, with the increase of time length, ρ_1 would also decrease, which means that the characteristic current can not well extract the changing trend of TZSC, and it would lead to wrong judgment.

5.5. Adaptation Analysis of Faulty Line Selection Method. In order to compare with other faulty line selection methods, choose TZSC with (l_4, 90°, and 10 Ω) fault situation as an example, and demonstrate it from the following two cases, respectively: with noise and without noise. With the disturbance of noise, signal-to-noise ratio of the added noise is −0.5 db, and antinoise performances of existing methods are emphatically analyzed. At the end, in different faulty conditions, the selection results, which are from different faulty line selection methods, are given.

5.5.1. Without Disturbance of Noise

VSBS Method. According to the method in the paper, input TZSC of each line to VSBS, use fourth-order Runge-Kutta method for numerical simulation, and calculate cross-correlation coefficients of every line $i_c(t)$, the results of which are shown in Table 10. Choose characteristic current of l_3 and display its waveform in 0.019 s~0.021 s, as is shown in Figure 8(a).

Wavelet Packet Method. We use db10 wavelet packet to decompose TZSC of each line by four layers. Choose characteristic frequency band according to the maximum energy selection principle [32], restructure it with single branch, and calculate cross-correlation coefficient of every frequency band, whose results are shown in Table 10. Choose characteristic frequency band of l_3 and display its waveform in 0~50, as is shown in Figure 8(b).

Wavelet Method. We use db10 wavelet to decompose TZSC of each line by four layers. Choose the approximation coefficients of the four-layer wavelet of each line as characteristic signal, restructure it with single branch, and calculate cross-correlation coefficient after the restructuring of every characteristic signal, the results of which are shown in Table 10. Choose approximation coefficient waveform of line l_3 and display its waveform in 0.019 s~0.021 s which is shown in Figure 8(c).

EMD Method. We use EMD algorithm [33] to decompose TZSC of each line. Choose the first intrinsic mode components (IMF1 component) after treatment as characteristic mode component, calculate cross-correlation coefficient of each IMF1 component, and the results are shown in Table 10. Choose IMF1 component of line l_3 and display its waveform in 0.019 s~0.021 s, which is shown in Figure 8(d).

Firstly, for waveform, the changing trend of initial stage waveform in Figure 8(a) is clearer than that in Figure 8(b), indicating that VSBS can better describe the changing trend of TZSC compared to wavelet packet transform, because when the initial value is 0, the brown particles are in potential peak position of bistable system, and any small disturbance will make the brown particles of bistable system move drastically, so the bistable system can well track the signal changing trend. Oscillating components in Figure 8(b) are more abundant than that in Figure 8(a), which means that the characteristic signal processed by wavelet packet could contain more frequency components; the reason is that wavelet packet has such good capability of time-frequency analysis that it can elaborately divide the high frequency and low frequency of signals, while the frequency compression and transformation of VSBS will make some frequency components lost.

The waveform of Figure 8(a) changed after fault occurred, and the changing amplitude is larger than that before fault, while the waveform of Figure 8(c) changed before fault occurred, and the changing amplitude after fault occurred is smaller than that before fault, showing that VSBS can better reflect the changing time and trend of TZSC compared to wavelet transform. In addition, the oscillation degree of Figure 8(c) is smaller than that of Figure 8(b), indicating that although wavelet transform has good time-frequency localization, its high frequency resolution is poor.

The oscillation degree of Figure 8(d) is the strongest, because IMF component obtained by EMD contains

Table 10: Cross-correlation coefficient of different signal processing algorithm without noise.

Signal processing	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Result
VSBS	0.9065	0.6112	−0.5938	0.7582	−0.7103	−0.9876	l_4
Wavelet packet	−0.4887	−0.0218	−0.2531	0.6983	−0.5990	−0.9548	Error
Wavelet	0.4483	−0.1536	−0.6129	0.1580	−0.9549	0.0137	Error
EMD	0.1968	0.4415	−0.2457	0.3389	−0.8920	−0.3948	l_4

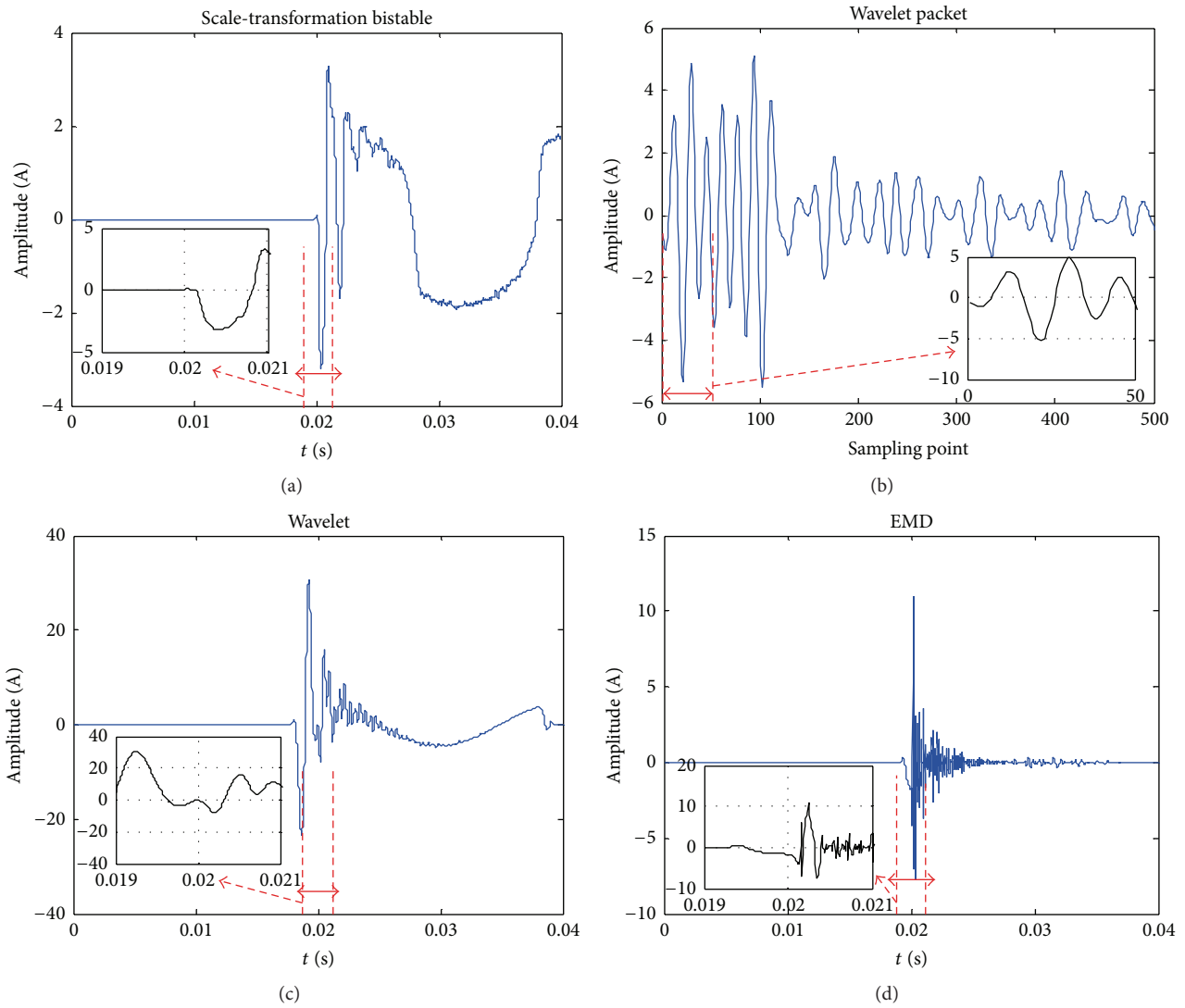


Figure 8: Characteristic signal of l_3 extracted by different signal processing algorithm without noise.

frequency component which changes with the signal itself and is more suitable for nonstationary signals like TZSC. However, similar to Figure 8(c), Figure 8(d) also changed before fault occurred, indicating that EMD algorithm has a weaker ability to describe changing time and trend of TZSC compared to VSBS.

Then, from the cross-correlation coefficient and faulty line selection results we can see that, with the method proposed in this paper, after processing with VSBS and EMD, only the cross-correlation coefficients of TZSC between line l_4 and other lines are all the same, so line l_4 is judged as faulty line, which is consistent with actual situations. However, processed by wavelet packet, the cross-correlation coefficients between characteristic signal of line l_1 and other lines are equal to −0.4887, −0.0218, and −0.2531, respectively, all of which are the same negative sign, and, in the same way, the cross-correlation coefficients between l_4 and other lines are equal to −0.2531, −0.5990, and −0.9548, respectively, which are also the same sign, so l_1 and l_4 are judged as faulty line, but this result is not consistent with actual fault situation.

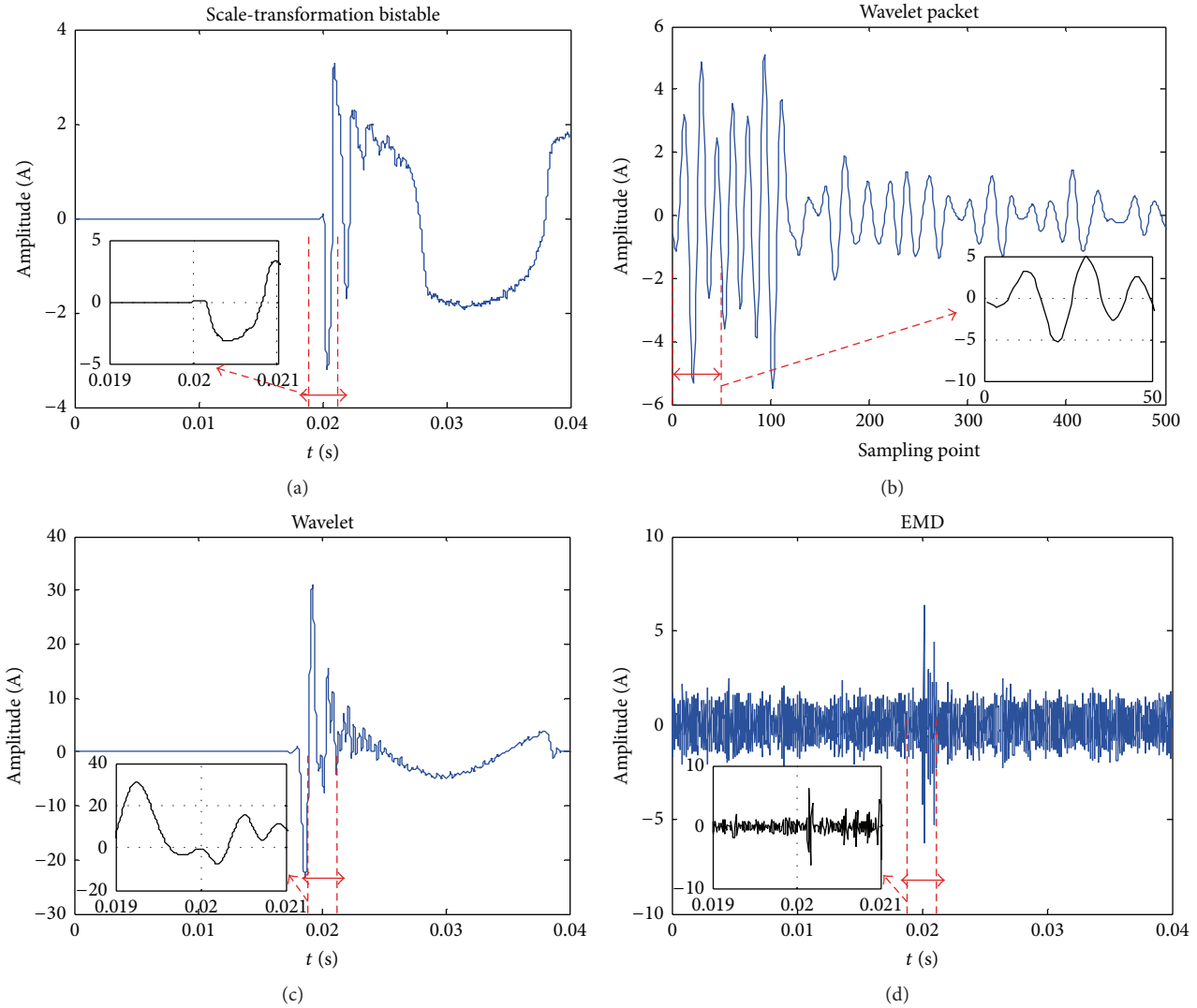


Figure 9: Characteristic signal of l_3 extracted by different signal processing algorithm with noise.

Table 11: Cross-correlation coefficient of different signal processing algorithm with noise.

Signal processing method	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Result
VSBS	0.9193	0.6179	−0.5949	0.7604	−0.7152	−0.9889	l_4
Wavelet packet	−0.4851	−0.0997	−0.2113	0.7048	−0.6111	−0.9429	Error
Wavelet	0.4328	−0.1579	−0.6183	0.1505	−0.9500	0.0231	Error
EMD	−0.4851	−0.0997	−0.2113	0.7048	−0.611	−0.9429	Error

And, then, processed by wavelet algorithm, none of the cross-correlation coefficients of characteristic signal between one line and other lines are the same sign, so all the lines are judged as healthy line, which, obviously, is not consistent with actual situation. This shows that wavelet transform and wavelet packet transform are not suitable for faulty line selection in this paper.

5.5.2. With Disturbance of Noise. With the same method and steps of Section 5.5.1, taking the waveform of line l_3 in 0.019 s~0.021 s as an example, we add a strong noise with signal-to-noise ratio as −0.5 db for simulation, the results of which are shown in Figure 9 and Table 11. Figure 9(a) is obtained by the process of VSBS, Figure 9(b) is obtained by the process of wavelet packet algorithm, Figure 9(c) is obtained by the process of wavelet transform algorithm, and Figure 9(d) is obtained by the process of EMD algorithm.

As to waveform, there are no obvious differences between other figures and Figure 8 except that Figure 9(d) is submerged in noise. As to cross-correlation coefficients, we will choose cross-correlation coefficient between line l_1 and line l_2 for analysis: without noise, processed in turn by VSBS,

Table 12: The faulty line selection results using VSBS.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 600 Ω)	**−0.5728**	**−0.7529**	**−0.6276**	0.8075	0.7007	0.8648	l_1 fault
	(90°, 1300 Ω)	**−0.3714**	**−0.7972**	**−0.9721**	0.4664	0.3556	0.7120	l_1 fault
l_2	(30°, 60 Ω)	**−0.8249**	0.8697	0.8300	**−0.9433**	**−0.9556**	0.9368	l_2 fault
	(60°, 700 Ω)	**−0.8470**	0.8171	0.7311	**−0.9788**	**−0.9223**	0.9154	l_2 fault
l_3	(90°, 1200 Ω)	0.6945	**−0.8165**	0.7977	**−0.6321**	0.6379	**−0.9853**	l_3 fault
	(30°, 80 Ω)	0.3286	**−0.4111**	0.3350	**−0.5162**	0.7632	**−0.6732**	l_3 fault
l_4	(60°, 800 Ω)	0.7923	0.5940	**−0.1569**	0.8881	**−0.4586**	**−0.5013**	l_4 fault
	(0°, 1000 Ω)	0.1455	0.4597	**−0.5422**	0.6710	**−0.0118**	**−0.6022**	l_4 fault

Table 13: The faulty line selection results using wavelet packet.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 600 Ω)	**−0.4176**	**−0.7475**	**−0.7392**	0.0358	0.4844	0.1128	l_1 fault
	(90°, 1300 Ω)	**−0.4293**	**−0.7013**	**−0.6931**	0.0363	0.4732	−0.0221	l_1 fault
l_2	(30°, 60 Ω)	−0.0995	−0.1246	−0.3036	−0.5007	−0.5424	−0.3091	Error
	(60°, 700 Ω)	−0.0659	−0.2447	0.0931	−0.6616	−0.3810	−0.4056	Error
l_3	(90°, 1200 Ω)	−0.1564	−0.1681	−0.2352	−0.0367	−0.3298	−0.8204	Error
	(30°, 80 Ω)	−0.0819	−0.2292	−0.3543	0.0630	−0.2489	−0.7990	Error
l_4	(60°, 800 Ω)	−0.4760	−0.0045	−0.6670	0.3079	−0.0135	−0.7307	Error
	(0°, 1000 Ω)	−0.4838	0.0016	−0.6757	0.2884	0.0106	−0.7269	l_4 fault

Table 14: The faulty line selection results using wavelet.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 600 Ω)	**−0.5330**	**−0.7158**	**−0.8579**	0.4851	0.2899	0.5469	l_1 fault
	(90°, 1300 Ω)	**−0.3189**	**−0.3250**	**−0.9168**	0.2064	0.2794	0.2527	l_1 fault
l_2	(30°, 60 Ω)	−0.6787	−0.0655	0.6617	0.0808	−0.9550	−0.2924	Error
	(60°, 700 Ω)	−0.5115	0.3300	0.4164	−0.6453	−0.8811	0.3380	Error
l_3	(90°, 1200 Ω)	0.3446	**−0.3254**	0.3109	**−0.3577**	0.2208	**−0.9037**	l_3 fault
	(30°, 80 Ω)	0.0951	−0.0220	0.1917	0.2228	0.1118	−0.1896	Error
l_4	(60°, 800 Ω)	0.3017	0.4761	**−0.4812**	0.6225	**−0.6560**	**−0.8082**	l_4 fault
	(0°, 1000 Ω)	0.0751	0.0991	−0.2497	0.6540	−0.2592	0.0464	Error

wavelet packet, and wavelet algorithm, the value is 0.9065, −0.4887, and 0.4483, respectively, while, with noise, processed in turn by VSBS, wavelet packet, and wavelet algorithm, the value is 0.9193, −0.4851, and 0.4328, respectively. Thus it can be seen that, with noise, the cross-correlation coefficient values by VSBS, wavelet packet, and wavelet algorithm are of little difference, so all of them have better antinoise ability. However, the cross-correlation coefficient processed by EMD algorithm without noise is 0.1968, and, with noise, the value is −0.4851, which changes from positive correlation to negative correlation, and the change is large, so combined with Figure 8(d) we can say that the antinoise ability of EMD algorithm is weak.

In summary, VSBS can extract the changing trend in initial stage of weak TZSC with the disturbance of strong noise, and its performance is better compared to wavelet transform, wavelet packet transform, and EMD algorithm; therefore, we choose VSBS to extract characteristic frequency band of TZSC in this paper.

5.5.3. Faulty Line Selection Results from Different Method. In strong noise background whose signal-to-noise ratio is −0.5 db, when different fault occurs including different lines, faulty resistance, and initial phase, the VSBS, wavelet packet, wavelet, and EMD are employed to select faulty line, respectively, and their faulty line selection results are shown in Tables 12–15, respectively.

Table 12 shows that VSBS has no misjudgment in strong noise background and different faulty conditions; that is, the VSBS method can select faulty line correctly. However, there are many misjudgments in Tables 13–15. These data indicate further that the antinoise performance of VSBS is better compared to wavelet transform, wavelet packet transform, and EMD algorithm.

Table 15: The faulty line selection results using EMD.

Faulty line	Fault situation	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	Judgment result
l_1	(0°, 600 Ω)	−0.0014	−0.0122	−0.0328	0.0103	−0.0362	−0.0239	Error
	(90°, 1300 Ω)	0.1435	−0.0413	0.0427	0.0506	−0.0252	0.0327	Error
l_2	(30°, 60 Ω)	−0.0428	−0.0414	0.1031	−0.0378	−0.1844	−0.0168	Error
	(60°, 700 Ω)	−0.0061	0.0327	0.0047	0.0046	0.0319	−0.0592	Error
l_3	(90°, 1200 Ω)	−0.0273	**−0.0076**	0.0364	**0.0416**	0.0595	**0.0010**	Error
	(30°, 80 Ω)	−0.0002	**−0.0232**	0.0214	**−0.0680**	0.0496	**−0.3613**	l_3 fault
l_4	(60°, 800 Ω)	−0.0207	0.0271	**0.0123**	−0.0448	**0.0256**	**0.0073**	l_4 fault
	(0°, 1000 Ω)	0.0370	−0.0199	−0.0206	−0.1296	0.0571	−0.0115	Error

6. Conclusions

This paper proposes a novel faulty line selection method for distribution network based on VSBS theory, and our research gets the following conclusions:

(1) VSBS has better recognition for TZSC, which can effectively extract the changing trend of TZSC in initial stage under different fault situations, and the method can accurately judge the faulty line. In addition, VSBS has better antinoise ability, which helps extract the changing trend of weak TZSC with the disturbance of strong noise, and its antinoise performance is better than that of EMD algorithm and harmonic selection criterion.

(2) The changing trend of TZSC in initial stage (0~0.005 s) is used to judge faulty line, which can reduce calculation time and the requirements for hardware. Besides, for the characterization capability of changing time and trend of TZSC in initial stage, the method in this paper is better than wavelet algorithm and wavelet packet algorithm.

(3) The inadequacies of this paper are as follows: the frequency compression ratio is obtained through experiment, which might cause deviation. In addition, high resistance to ground fault with −10 db strong noise needs further study owing to the insufficient sensitivity of the present research.

Appendix

Build the simulation model according to the parameters, make fault of line l_1 occur at the point 5 km from the bus, and change the initial fault angle (0°, 30°, 60°, and 90°) as well as ground resistance for simulation. Then, with the proposed selection method, the cross-correlation coefficients of each line and faulty line selection results are shown in Table 2.

Add 0.5 db or −0.5 db noise intensity to TZSC when fault in different lines occurs. And set signal before fault to 0. The selection results and specific cross-correlation coefficients are shown in Tables 4 and 5.

In Figure 5, l_3 is cable-overhead line, and l_4 is pure cable line; we carry out faulty line selection with the method proposed in the paper to verify its adaptability, the results of which are shown in Table 7, and specific cross-correlation coefficients are shown in Table 7.

Since the distance of fault point is different in actual fault situations, we carry out simulation of line l_1, with different distance from the bus line, and the fault distance is 4.5 km, 7.5 km, 10.5 km, and 13.5 km, respectively. Select the faulty line with the method and the results are shown in Table 8.

Notations

VSBS: Variable scale bistable system
TZSC: Transient zero-sequence current.

Competing Interests

The authors declare no conflict of interests.

Authors' Contributions

Xiaowei Wang and Jie Gao conceived and designed the experiments; Jie Gao performed the experiments; Qiming Cheng analyzed the data; Guobing Song, Xiangxiang Wei, and Yanfang Wei contributed reagents/materials/analysis tools; Jie Gao wrote the paper.

Acknowledgments

This work was supported by National Natural Science Fund (61403127) of China, Science and Technology Research (12B470003, 14A470004, and 14A470001) of Henan Province, and Control Engineering Lab Project (KG2011-15, KG2014-04) of Henan Province, China, and Doctoral Fund (B2014-023) of Henan Polytechnic University, China.

References

[1] H. K. Shu, *The Application of Electrical Engineering Signal Processing*, China Electric Power Press, Beijing, China, 2011.

[2] H. K. Shu, *Fault Line Selection of Distribution Power System*, China Machine Press, Beijing, China, 2008.

[3] Z. C. Pan, H. F. Zhang, F. Zhang, and Z. Sang, "Analysis and modification of signal injection based fault line selection protection," *Automation of Electric Power Systems*, vol. 31, no. 4, pp. 71–75, 2007 (Chinese).

[4] J. Liu, X. Q. Zhang, X. Y. Chen, B. Shen, X. Dong, and Z. Zhang, "Fault location and service restoration for distribution networks based on coordination of centralized intelligence and distributed intelligence," *Power System Technology*, vol. 37, no. 9, pp. 2608–2614, 2013 (Chinese).

[5] G. K. Ni, H. Bao, L. Zhang, and Y. Yang, "Criterion based on the fault component of zero sequence current for online fault location of single-phase fault in distribution network," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 30, no. 31, pp. 118–122, 2010.

[6] L. Zhang, P. Yang, D. M. Si, C. Qi, and Y. Yang, "Online fault location of neutral point ungrounded distribution network based on zero-sequence power direction," *Automation of Electric Power Systems*, vol. 32, no. 17, pp. 79–82, 2008 (Chinese).

[7] X. Z. Dong and S. X. Shi, "Identifying single-phase-to-ground fault feeder in neutral noneffectively grounded distribution system using wavelet transform," *IEEE Transactions on Power Delivery*, vol. 23, no. 4, pp. 1829–1837, 2008.

[8] X. Wang, J. Gao, X. Wei, and Y. Hou, "A novel fault line selection method based on improved oscillator system of power distribution network," *Mathematical Problems in Engineering*, vol. 2014, Article ID 901810, 19 pages, 2014.

[9] X. W. Wang, J. W. Wu, and R. Y. Li, "A novel method of fault selection based on voting mechanism of prony relative entropy theroy," *Electric Power*, vol. 46, no. 1, pp. 59–65, 2013.

[10] S. Zhang, Z.-Y. He, Q. Wang, and S. Lin, "Fault line selection of resonant grounding system based on the characteristics of charge-voltage in the transient zero sequence and support vector machine," *Power System Protection and Control*, vol. 41, no. 12, pp. 71–78, 2013.

[11] S. Q. Zhang, X. P. Zhai, X. Dong, L. Li, and B. Tang, "Application of EMD and Duffing oscillator to fault line detection in uneffectively grounded system," *Proceedings of the CSEE*, vol. 33, no. 10, pp. 161–167, 2013.

[12] X. Wang, X. Wei, J. Gao, Y. Hou, and Y. Wei, "Stepped fault line selection method based on spectral kurtosis and relative energy entropy of small current to ground system," *Journal of Applied Mathematics*, vol. 2014, Article ID 726205, 18 pages, 2014.

[13] J. Zhang, Z. Y. He, and Y. Jia, "Fault line identification approach based on S-transform," *Proceedings of CSEE*, vol. 31, no. 10, pp. 109–115, 2011.

[14] H. C. Shu, W. Y. Zhao, and S. X. Peng, "Faulty line selection based on HHT detection for hybrid distribution network," *Electric Power Automation Equipment*, vol. 29, no. 5, pp. 4–10, 2009.

[15] Q. Li and J. Z. Xu, "Power system fault diagnosis based on subjective Bayesian approach," *Automation of Electric Power Systems*, vol. 31, no. 15, pp. 46–50, 2007.

[16] H. C. Shu, L. Xu, and S. X. Peng, "Correlation analysis for faulty feeder detection in resonant earthed system," *Automation of Electric Power Systems*, vol. 28, no. 9, pp. 6–9, 2008.

[17] L. P. Wu, C. Huang, D. B. Lin, Z. Zhu, and H. Jiang, "Faulty line selection based on transient wavelet energy for non-solid-earthed network," *Electric Power Automation Equipment*, vol. 33, no. 5, pp. 70–75, 2013.

[18] Y. L. Meng and C. X. Pei, "Stochastic resonance in a bistable system driven by non-Gaussian noise and Gaussian noise," in *Proceedings of the IEEE Workshop on Electronics, Computer and Applications (IWECA '14)*, pp. 358–361, IEEE, Ottawa, Canada, May 2014.

[19] J. Fan, W.-L. Zhao, M.-L. Zhang, R.-H. Tan, and W.-Q. Wang, "Nonlinear dynamics of stochastic resonance and its application in the method of weak signal detection," *Acta Physica Sinica*, vol. 63, no. 11, Article ID 110506, 2014.

[20] J.-J. Tong, G.-L. Zhang, Q. Cai, J.-M. Jian, and X.-S. Guo, "Application of threshold stochastic resonance in low concentration gas detecting," *Journal of Zhejiang University (Engineering Science)*, vol. 49, no. 1, pp. 15–19, 2015.

[21] S. L. Lu, Q. B. He, F. Hu, and F. Kong, "Sequential multiscale noise tuning stochastic resonance for train bearing fault diagnosis in an embedded system," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 1, pp. 106–116, 2014.

[22] Y.-B. Li, M.-Q. Xu, H.-Y. Zhao, S.-Y. Zhang, and W.-H. Huang, "Application of cascaded bistable stochastic resonance and Hermite interpolation local mean decomposition method in gear fault diagnosis," *Journal of Vibration and Shock*, vol. 34, no. 5, pp. 95–101, 2015.

[23] P. E. Greenwood, L. M. Ward, D. F. Russell, A. Neiman, and F. Moss, "Stochastic resonance enhances the electrosensory information available to paddlefish for prey capture," *Physical Review Letters*, vol. 84, no. 20, pp. 4773–4776, 2000.

[24] H. Xin, "Theoretical study on stochastic resonance in chemical systems," *Chinese Journal of Chemical Physics*, vol. 13, no. 4, pp. 404–405, 2000.

[25] H. S. Zhang, Z. Y. He, and J. Zhang, "Frequency spectrum characteristic analysis of single-phase grounding fault in resonant grounded systems," *Automation of Electric Power Systems*, vol. 36, no. 6, pp. 79–84, 2012.

[26] H. C. Shu, L. Gao, and R. M. Duan, "A novel hough transform approach of fault line selection in distribution networks using the total zero-sequence current," *Automation of Electric Power Systems*, vol. 37, pp. 1–7, 2013.

[27] Y. G. Leng, *Mechanism Analysis of the Large Signal Scale-Transformation Stochastic Resonance and Its Engineering Application Study*, Tianjin Unibersity, Tianjin, China, 2004.

[28] N. Q. Hu, *The Theory and Method of Detection of Weak Characteristic Signal Based on Stochastic Resonance*, National Defense Industry Press, Beijing, China, 2012.

[29] X. N. Kang, X. Liu, and J. L. Sounan, "New method for fault line seclection in non-solidly grounded system based on matrix pencil method," *Automation of Electric Power Systems*, vol. 36, no. 12, pp. 88–93, 2012.

[30] M. F. Guo, S. D. Liu, and G. J. Yang, "A new approach to detect fault line in resonant earthed system based on Hilbert spectrum band-pass filter and transient waveform recognition," *Advanced Technology of Electrical Engineering and Energy*, vol. 32, no. 3, pp. 67–74, 2013.

[31] Z. J. Kang, D. D. Li, and X. L. Liu, "Faulty line selection with non-power frequency transient components of distribution network," *Electric Power Automation Equipment*, vol. 31, no. 4, pp. 1–6, 2011.

[32] Z. Y. He, *Application of Power Transient Signal Based on Wavelet Analysis*, China Electric Power Press, Beijing, China, 2011.

[33] X. W. Wang, Y. D. Li, and S. Tian, "A novel single-phase to ground fault location method based on EMD and ApEn algorithm for small current to ground system," *Journal of Computational Information Systems*, vol. 8, no. 13, pp. 5629–5637, 2012.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 3251632, 10 pages
http://dx.doi.org/10.1155/2016/3251632

Research Article

The Fuzzy Feedback Scheduling of Real-Time Middleware in Cyber-Physical Systems for Robot Control

Feng Tang, Ping Zhang, and Fang Li

South China University of Technology, Guangzhou, China

Correspondence should be addressed to Feng Tang; fengtang@scut.edu.cn

Received 4 March 2016; Revised 20 July 2016; Accepted 21 July 2016

Academic Editor: Antonio Fernández-Caballero

Cyber-physical systems for robot control integrate the computing units and physical devices, which are real-time systems with periodic events. This work focuses on CPS task scheduling in order to solve the problem of slow response and packet loss caused by the interaction between each service. The two-level fuzzy feedback scheduling scheme is designed to adjust the task priority and period according to the combined effects of the response time and packet loss. Empirical results verify the rationality of the cyber-physical system architecture for robot control and illustrate the feasibility of the fuzzy feedback scheduling method.

1. Introduction

Cyber-physical systems (CPSs) are integrations of the physical world, computation, communication, and control designs [1]. Many cyber-physical systems are equipped with distributed computing units that communicate with each other through the network [2]. As [3, 4] propose, application architecture, control technique, and resource allocation are important research topics in CPSs. One representative CPSs application architecture is stated in [5]. Another application of CPSs for neutrally controlled artificial legs is put forward in [6]. Reference [7] outlines the architecture of passive control CPSs. All these systems embody the features of CPS like distributed computing.

Distributed robot control systems consist of many computing nodes, actuators, sensing devices, and network topologies [8–10]. These are typical applications of distributed CPSs that must handle many aperiodic and periodic events. The computing nodes or devices have different interfaces. Applying middleware can realize network interconnection, data integration, and integrated application, which accounts for the increasing significance of middleware technologies with respect to distributed robot control CPSs. Real-time CORBA [11] is a vital real-time middleware frequently used in distributed systems. A reconfigurable real-time middleware is proposed in [12], which provides a flexible software platform for CPSs with timing constraints.

The scheduling models of traditional real-time CORBA are completely independent of each node [13]. That is to say, the scheduling in each node is independent. The scheduling method does not take nodal performance into account; therefore, it fails to guarantee global optimum. For that, the fuzzy feedback scheduling, a global scheduling framework based on end-to-end real-time CORBA scheduling model, is proposed in this paper. According to the nodal performance index, this method adjusts the task priority and period to improve the overall performance of CPSs for robot control and enhance the performance of the end-to-end real-time system.

2. Systems Architecture of CPSs for Robot Control Based on Real-Time Middleware

As the core mechanism of CORBA, ORB (Object Request Broker) realizes transparent interaction between objects residing at various computing nodes in heterogeneous environments. Using naming service of CORBA, the clients and the CRRBA objects are connected according to the objects' names. This method facilitates programming and establishes seamless connection between objective systems.

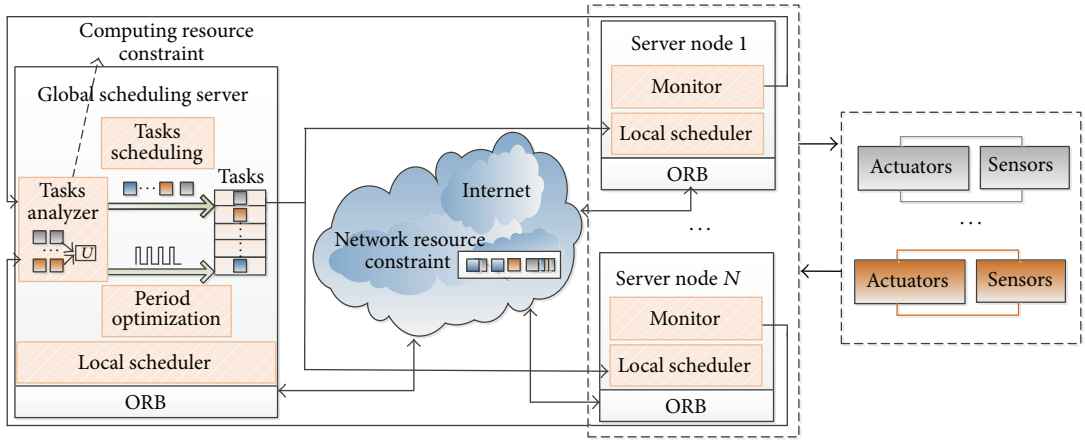


Figure 1: Systems architecture of CPSs for robot control.

Figure 1 shows the architecture of CPSs for robot control, which mainly consists of actuators, sensors, server nodes, and global scheduling server. These components communicate with each other via ORB. The tasks that travel along the servers compete for network resources of ORB soft bus. Once the requests are received, global scheduling server schedules the tasks to control multiple actuators, which can be regarded as a task management. Be they aperiodic or periodic events, the control tasks are assigned to server nodes by global scheduling server, resulting in competition for computing resources of global scheduling server. Moreover, implemented in different server nodes, the tasks in one server node compete for computing resources within their own server node while the monitor in charge of this server node monitors performance of this node. Then feedback information is sent to global scheduling server. The global scheduling server decides on the task priority and period according to the feedback information and assigns the tasks to server nodes in the light of priority. Server nodes map task priority to local operating system. This work focuses on resource scheduling and two-level scheduling method is illustrated later in the article.

3. Fuzzy Feedback Scheduling Method

The CPSs for robot control are an evolution of the real-time networked control systems [14, 15]. The tasks scheduling of CPSs can be based on the involved research results. Scheduling strategy can be divided into static scheduling and dynamic scheduling according to the implementation method. Liu and Layland proposed RM (Rate Monotonic) scheduling algorithm and EDF (Earliest Deadline First) real-time scheduling algorithm [16]. They are the most representative of the research results in scheduling algorithm. At present, they have become the research foundation of scheduling algorithm. RM algorithm is a static priority assignment algorithm which assigns priorities to tasks according to their request rates; the more frequent the task is, the higher in priority the task is. Reference [16] proves that the RM algorithm is the best one in the static scheduling algorithm. EDF scheduling algorithm is a dynamic priority scheduling algorithm. it assigns priorities to tasks according to their absolute deadline. A task will be assigned the highest priority if the deadline of its current request is the nearest and will be assigned the lowest priority if the deadline of its current request is the furthest. In the EDF scheduling algorithm, tasks with the highest priority level are always to be executed first, while the lower-priority tasks are to be preempted.

Because the robot control system environment is complex, RM scheduling algorithm has poor environment suitability [16]. Reference [17] makes it clear that dynamic scheduling algorithm outperforms static scheduling algorithm in changeable environments. Therefore, this work centers on EDF real-time scheduling algorithm. Reference [18] presents an improved dynamic EDF scheduling method for the network control systems, where task priority is changed according to the control error. References [19–21] demonstrate similar scheduling methods to set the task priority and reduce packet loss.

Studies have shown that feedback scheduling can better utilize resources and improve control performance in the distributed systems [22]. So both EDF dynamic scheduling and feedback schedule are taken into consideration. Reference [23] illustrates a fuzzy feedback scheduler that makes better use of resources according to the system performance while [24] shows an online feedback scheduling algorithm for the sake of error reduction by optimizing the object function for a robot manipulator.

3.1. The Fuzzy Feedback Scheduling Framework. Figure 2 presents the fuzzy feedback scheduling framework. Two-level fuzzy feedback scheduling method is designed in this paper where the object function of the method includes the response time and transmission error.

The effect of the first level scheduling is setting priority. An EDF dynamic scheduling algorithm is designed at this level. First of all, requirements of resource utilization U_i for each close-loop by fuzzy feedback controller are estimated. Then, the priority configurator computes the task priority in consideration of deadline, importance, and requirements of resource utilization. Lastly, tasks are transferred in the light of priority order and mapped to the operating system.

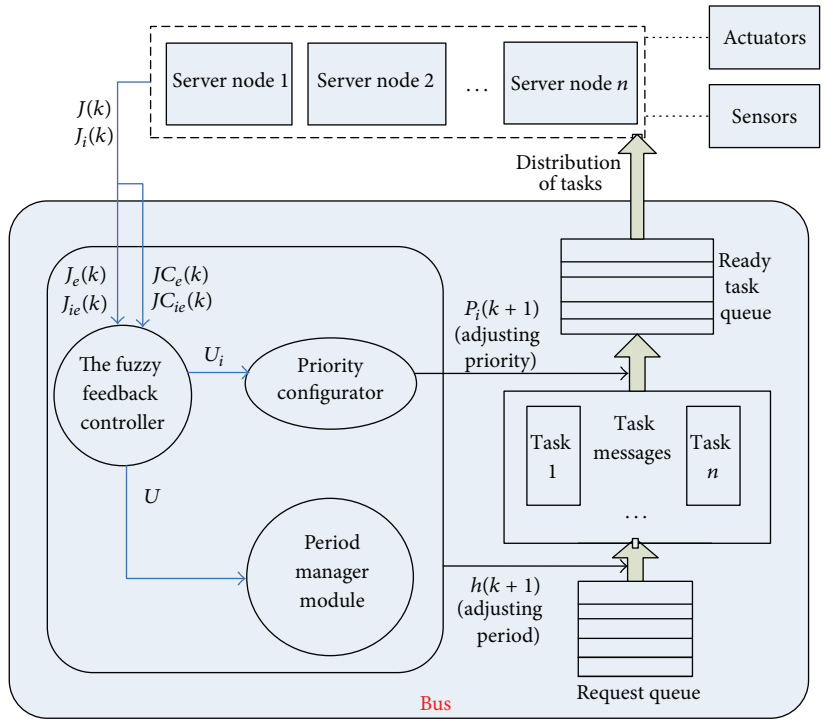


Figure 2: The fuzzy feedback scheduling framework.

The effect of the second level scheduling is adjusting period. Firstly, the total resource utilization U is computed by fuzzy feedback controller, and then period manager module adjusts the task period according to the total resource utilization. This method can address network congestion and achieve better resource utilization.

The feedback performance criteria function of the ith control loop is shown as follows: $J_i(k) = \alpha E_i(k) + \beta\tau_i(k)$, $E_i(k) = (X_i(k) - X_{iS}(k))/X_i(k)$. $X_i(k)$ representing all packets of the ith control loop at the kth cycle is computed by the monitor of the global scheduling server. $X_{iS}(k)$ referring to the packets that have been transmitted successfully is calculated by the monitor of the server nodes and sent to the global scheduling server. $\tau_i(k)$ represents the response time ratio and follows the equation $\tau_i(k) = H_i(k)/H_{i\max}$. $H_i(k)$ is the execution time of the ith task. Global scheduling server computes the execution time via the starting time and terminal time that are reported from the actuators. $H_{i\max}$ is the maximum execution time, while α and β are the weight of the packet loss ratio and response time ratio following the equation of $\alpha + \beta = 1$.

The deviation of performance indicator of the ith control loop is described as follows: $J_{ie}(k) = J_{ir} - J_i(k) = \alpha(e_{ir} - E_i(k)) + \beta(l_{ir} - \tau_i(k))$. J_{ir} represents the desired performance criteria and follows the equation $J_{ir} = \alpha e_{ir} + \beta l_{ir}$. The differential of performance criteria follows the formula $JC_{ie}(k) = J_{ie}(k) - J_{ie}(k-1)$, which echoes the direction of change for performance criteria.

Similarly, the overall feedback performance criteria function of all control loops are defined as follows: $J(k) = \alpha E(k) + \beta\tau(k)$, $E(k) = (X(k) - X_S(k))/X(k)$. $X(k)$ represents all packets of the whole control loops, while $X_S(k)$ represents the packets that have been transmitted successfully of the whole control loops. In $\tau(k) = (\sum_{i=1}^{n} H_i(k))/(\sum_{i=1}^{n} H_{i\max})$, $\sum_{i=1}^{n} H_i(k)$ refers to the total response time of all control loops and $\sum_{i=1}^{n} H_{i\max}$ the total maximum response time of all control loops. Deviation of overall performance indicator is described as $J_e(k) = J_r - J(k) = \alpha(e_r - E(k)) + \beta(l_r - \tau(k))$. Differential of overall performance criteria follows the equation $JC_e(k) = J_e(k) - J_e(k-1)$.

In this paper, the fuzzy feedback scheduling framework adjusts the task priority and period according to the feedback performance. The greater the priority is, the sooner the task should be handled. The worse the performance of the control loops is, the more resources they receive. When the total resource utilization U is great, it implies network congestion and the period should be shortened, and vice versa.

3.2. Fuzzy Feedback Control. The deviation of performance indicators $J_{ie}(k)$, $J_e(k)$ and the differential of performance criteria $JC_{ie}(k)$, $JC_e(k)$ are the inputs into the fuzzy feedback controller. Total resource utilization U and the ith resource utilization U_i are accessible through fuzzy feedback control. The relations between inputs and outputs are illustrated in the following equations: $U = f(J_e(k), JC_e(k))$, $U_i = f(J_{ie}(k), JC_{ie}(k))$.

There are four steps in fuzzy feedback control: (1) fuzzification; (2) establishing fuzzy control rules; (3) fuzzy composition; and (4) fuzzy judgment. The ith resource utilization U_i is taken as an example to illustrate these processes.

Table 1: Subordinate degree value for $J_{ie}(k)$.

Variable	$J_{ie}(k)$						
	−0.6	−0.4	−0.2	0	0.2	0.4	0.6
NB	1.0	0.5	0	0	0	0	0
NM	0.5	1	0.5	0	0	0	0
NS	0	0.5	1	0.5	0	0	0
Z	0	0	0.5	1	0.5	0	0
PS	0	0	0	0.5	1	0.5	0
PM	0	0	0	0	0.5	1	0.5
PB	0	0	0	0	0	0.5	1

Table 2: Subordinate degree value for $JC_{ie}(k)$.

Variable	$JC_{ie}(k)$						
	−0.3	−0.2	−0.1	0	0.1	0.2	0.3
NB	1.0	0.5	0	0	0	0	0
NM	0.5	1	0.5	0	0	0	0
NS	0	0.5	1	0.5	0	0	0
Z	0	0	0.5	1	0.5	0	0
PS	0	0	0	0.5	1	0.5	0
PM	0	0	0	0	0.5	1	0.5
PB	0	0	0	0	0	0.5	1

Table 3: Subordinate degree value for U_i.

Variable	U_i		
	0.6	0.4	0.2
PB	1.0	0.5	0
PM	0.5	1	0.5
PS	0	0.5	1

3.2.1. Fuzzification. The fuzzy sets of $J_{ie}(k)$ and $JC_{ie}(k)$ are expressed as {NB, NM, NS, Z, PS, PM, PB} which, respectively, denote negative big, negative medium, negative small, zero and positive small, positive medium, positive big. The fuzzy sets of $J_{ie}(k)$ and $JC_{ie}(k)$ are quantified as $\{-0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6\}$ and $\{-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3\}$, respectively, whose subordinate degree values are correspondingly presented in Tables 1 and 2. The fuzzy set of the output variable U_i is expressed as {PS, PM, PB}, which, respectively, denotes small, medium, and big values. The fuzzy set is quantified as $\{0.2, 0.4, 0.6\}$, whose subordinate degree values are shown in Table 3.

3.2.2. Establishing Fuzzy Control Rules. The basic idea of fuzzy control rules is shown in Table 4. When the performance criteria and its differential value both are positives or negatives and their absolute values are relatively big, the system's performance is diverging from the calibrated value with a tendency of continuous deviation. Under such circumstance, the system performs poorly. So a bigger value should be assigned to the requirement of resource utilization U_i. When the performance criteria and its differential values are smaller, the system's performance is close to the calibrated value, which means the system performs well. In this case, a smaller value should be assigned to U_i. When the performance criteria and its differential are valued one positive and another negative, the system's performance is approaching to the calibrated value with a tendency of improvement. A medium value should be assigned to U_i. Fuzzy control rules can be expressed in fuzzy conditional statements, for example,

Table 4: Fuzzy control rules.

U_i	$JC_{ie}(k)$						
	NB	NM	NS	Z	PS	PM	PB
$J_{ie}(k)$							
NB	PS	PS	PS	PS	PS	PM	PM
NM	PS	PS	PS	PS	PM	PM	PM
NS	PS	PS	PS	PM	PM	PM	PB
Z	PS	PS	PM	PM	PM	PB	PB
PS	PS	PM	PM	PM	PB	PB	PB
PM	PM	PM	PM	PB	PB	PB	PB
PB	PM	PM	PB	PB	PB	PB	PB

$$\begin{aligned} &\text{if } J_{ie}(k) = \text{NB or NM or NS},\\ &JC_{ie}(k) = \text{NB or NM or NS}\\ &\text{then } U_i = \text{PS}. \end{aligned} \tag{1}$$

3.2.3. Fuzzy Composition. Compositional rule of fuzzy inference expresses a fuzzy relation by means of the Cartesian product. For instance, "if A and B then C" is expressed as $R = (A \cap B) \rightarrow C$. According to the rule of Mamdani inference, the following expressions are concluded, where A^L refers to the transpose of A [25]:

$$R = (A \cap B) \times C = \left[A^L \circ B\right]^L \circ C. \tag{2}$$

There are 49 fuzzy control rules in Table 4, so the fuzzy relation is expressed as (3) shows [26]:

$$R = R_1 \cup R_2 \cup \cdots R_{49}. \tag{3}$$

$R_1, R_2, \ldots, R_{49}$ can be calculated from the following formula:

$$\begin{aligned} R_1 &= D_1^L \times (\text{PB})_{U_i},\\ D_1 &= (\text{PB})_{J_{ie}(k)} \times (\text{PB})_{JC_{ie}(k)},\\ &\vdots\\ R_{49} &= D_{49}^L \times (\text{NB})_{U_i},\\ D_{49} &= (\text{NB})_{J_{ie}(k)} \times (\text{NB})_{JC_{ie}(k)}. \end{aligned} \tag{4}$$

D_1 can be calculated from the following equation:

$$D_1 = (\mathrm{PB})_{J_{ie}(k)} \times (\mathrm{PB})_{JC_{ie}(k)} = (\mathrm{PB})^L_{J_{ie}(k)} \circ (\mathrm{PB})_{JC_{ie}(k)}$$
$$= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 1 \end{bmatrix} \circ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.5 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 1 \end{bmatrix}. \quad (5)$$

R_1 can be figured out from (6) by putting (5) into (4):

$$R_1 = D_1^L \times (\mathrm{PB})_{U_i} = \left[(\mathrm{PB})^L_{J_{ie}(k)} \circ (\mathrm{PB})_{JC_{ie}(k)}\right]^L \circ (\mathrm{PB})_{U_i}$$
$$= \begin{bmatrix} 40 \times 1 & 0 \\ 0.5 \\ 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 1 \end{bmatrix} \circ \begin{bmatrix} 1 & 0.5 & 0 \end{bmatrix} = \begin{bmatrix} & 40 \times 3 & 0 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ & 5 \times 3 & 0 \\ 0.5 & 0.5 & 0 \\ 1 & 0.5 & 0 \end{bmatrix}. \quad (6)$$

It is found from (6) that R_1 is a 49×3 matrix. R can be calculated out of (7) on the base of (3) and (4):

$$R = R_1 \cup R_2 \cup \cdots R_{49} = \begin{bmatrix} 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \\ & \vdots & \\ 1 & 0.5 & 0 \\ 1 & 0.5 & 0 \end{bmatrix}. \quad (7)$$

Then the requirement of resource utilization U_i can be computed through fuzzy synthetic operation:

$$U_i = \left(J_{ie}(k) \times JC_{ie}(k)\right) \circ R. \quad (8)$$

For example, if,

$$J_{ie}(k) = (\mathrm{NB})_{J_{ie}(k)} = \begin{bmatrix} 1.0 & 0.5 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$
$$JC_{ie}(k) = (\mathrm{NB})_{JC_{ie}(k)} = \begin{bmatrix} 1.0 & 0.5 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad (9)$$

then

$$U_i = \left(J_{ie}(k) \times JC_{ie}(k)\right) \circ R$$
$$= \begin{bmatrix} 1.0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \vdots 35 \times 0 \end{bmatrix}$$
$$\circ \begin{bmatrix} 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \\ & \vdots & \\ 1 & 0.5 & 0 \\ 1 & 0.5 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & 1 \end{bmatrix}. \quad (10)$$

3.2.4. Fuzzy Judgment. There are mainly two judgment methods: weighted average method and max membership function. The weighted average method is expressed as follows:

$$u = \frac{\sum_{i=1}^{n} u_i \mu(u_i)}{\sum_{i=1}^{n} \mu(u_i)}. \quad (11)$$

In this paper, weighted average method is adopted to determine the requirements of resource utilization U_i, while the max membership function is used to study the total resource utilization U. For example, if $U_i = \begin{bmatrix} 0 & 0.5 & 1 \end{bmatrix}$, then U_i is calculated by the weighted average method as follows:

$$U_i = \frac{\sum_{i=1}^{3} u_i \mu(u_i)}{\sum_{i=1}^{3} \mu(u_i)} = \frac{0 \times 0.6 + 0.5 \times 0.4 + 1 \times 0.2}{0 + 0.5 + 1} \approx 0.267. \quad (12)$$

If $U = \begin{bmatrix} 0 & 0.5 & 1 \end{bmatrix}$, then U is figured out by the max membership function as follows:

$$U = \frac{0}{0.6} + \frac{0.5}{0.4} + \frac{1}{0.2}. \quad (13)$$

Since $\mu(0.2) = 1$ is the largest of all subordinates, $U = 0.2$ is selected.

3.3. Priority Assignment Algorithm. Once the requests are received, the global scheduling server will schedule the tasks. Priority is given to each task according to the following scheduling algorithm: the task message (including the priority attributes such as the task's arrival time and deadline) is arranged to the ready task queue according to the priority and then it is distributed to trigger execution of each task in server node. The priority attributes of each task and its subtasks should be consistent within task cycle h_i.

The traditional EDF algorithm cannot promptly adjust the task priority in an uncertain operation environment. In the meanwhile, the feedback scheduling fails to dispatch in time as a result of an introduced feedback mechanism

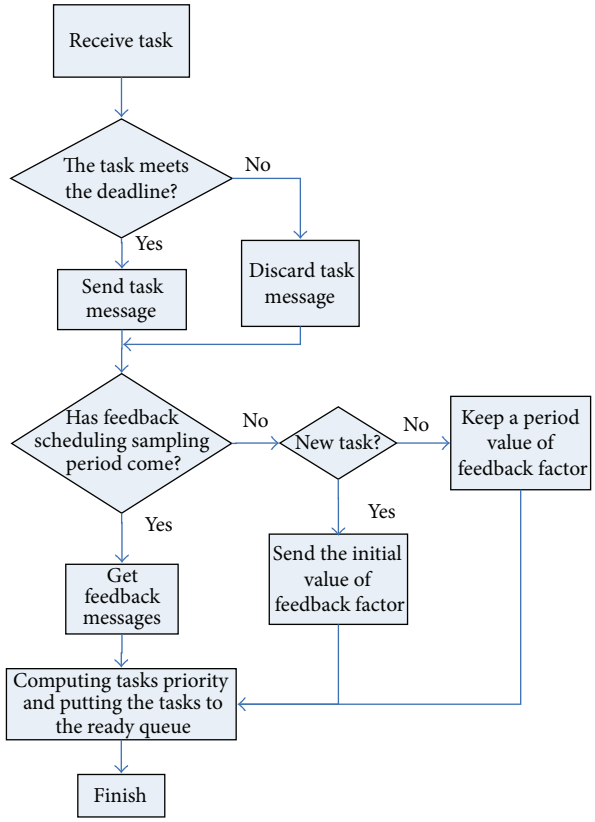


Figure 3: Task scheduling process.

(scheduling period should not be too short so as to avoid frequent switching task priority). "The fuzzy feedback improved EDF scheduling policy" proposed in this work is aimed at assigning the priority of each task and sorting the task messages in the ready task queue in progressive priority order. The basic idea is described as follows (Figure 3):

(1) When a task message arrives, its schedulability is analyzed at first to judge whether it applies to the deadline or not.

(2) When the scheduling sampling period comes, the fuzzy scheduling method as Section 3.2 shows is adopted to calculate the requirement of resource utilization U_i according to the feedback information $(J_{ie}(k), JC_{ie}(k))$. Then U_i is used to compute the priority as the next step proposes.

(3) Task T_i, typically, has a residual deadline de_i, resource utilization U_i, and the importance of the task im_i. Its particular priority $P(T_i, \mathrm{de}_i, U_i, \mathrm{im}_i)$ $(i = 1, \ldots, n)$ is represented as P_i.

Resource utilization U_i is introduced as one of the metrics to determine the priority of task T_i together with the task's importance im_i and residual deadline de_i. The task priority is expressed in the following equation:

$$P_i = \zeta U_i \times \frac{\mathrm{im}_i}{\mathrm{de}_i}. \tag{14}$$

```
Now adjusting the list!!
Size = 6
1: service ID = 16202        deadline = 129964974490228663
2: service ID = 30186        deadline = 129964974493067141
3: service ID = 7363         deadline = 129964974488790034
4: service ID = 8679         deadline = 129964974491642807
5: service ID = 12856        deadline = 129964974491628663
6: service ID = 9832         deadline = 129964974494505122
Scheduling ! ! ! ! ! ! ! ! !
Service 7363 : priority = 0.018733
Service 16202 : priority =  0.017710
Service  12856 : priority = 0.016687
Service   8679 : priority = 0.017632
Service 30186 : priority = 0.016609
Service 9832 : priority = 0.015589
Size = 6
1: service ID = 7363         deadline = 129964974491628663
2: service ID = 16202        deadline = 129964974493067141
3: service ID = 12856        deadline = 129964974488790034
4: service ID = 8679         deadline = 129964974494505122
5: service ID = 30186        deadline = 129964974490228663
6: service ID = 9832         deadline = 129964974491642807
```

Figure 4: A scheduling example.

If there comes a periodic task, the initial value of the resource utilization U_i is set at 0.3. As for an aperiodic task, U_i is set at 0.6. The adjustment coefficient ζ needs to be determined by an experimental method. It is possible that multiple tasks share the same priority. In this case, the tasks can be sorted according to the feedback performance or deadline. All in all, the shorter the deadline lasts, the greater resource utilization and task importance it embodies, and therefore the greater the priority is, and vice versa. This method excels itself in the full consideration of multifaceted attributes of the task and the presence of multiple parameters in the priorities. This method bridges the gap of traditional EDF algorithm which only takes deadline into account.

3.4. Period Adjustment Strategy. Too frequent periodic adjustment makes the system complicated and unstable. Therefore, three kinds of period $(P = 4, 8, 16)$ are developed. To judge the total resource utilization U, the max membership function is employed to produce the fuzzy feedback controller outputs of $U = 0.2, 0.4, 0.6$. The period manager module computes the period P according to the mapping relations between resource utilization and task period which is presented in Table 5 and then writes the period to the task messages and transfers the message to related services. Finally, the services adjust the period according to the received message.

4. Algorithm Simulation

4.1. Priority Assignment Examples. A scheduling example is shown in Figure 5; the arrangement of tasks on ready task queue before scheduling stands at the upper part of Figure 4. To be specific the tasks ID are 16202, 30186, 7363, 8679, 12856, and 9832.

The lower part of this figure is the arrangement of tasks after scheduling. It shows that task 1 and task 3 embrace

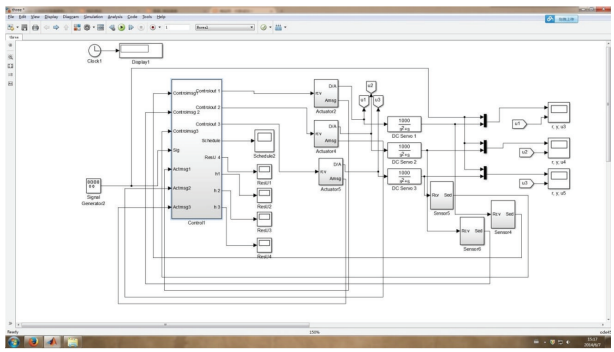
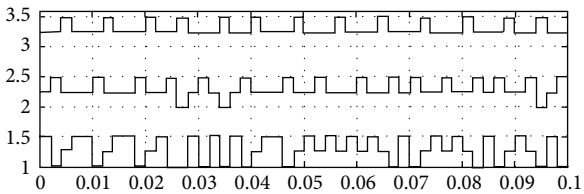
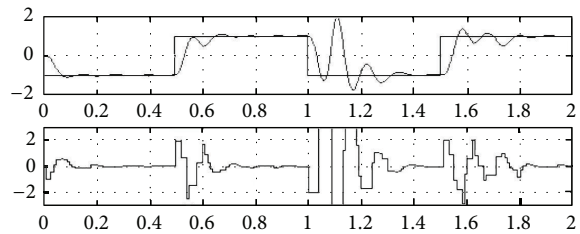


Figure 5: Simulation model for the scheduling system.



(a) Task execution



(b) The output curve of the actuators and controller

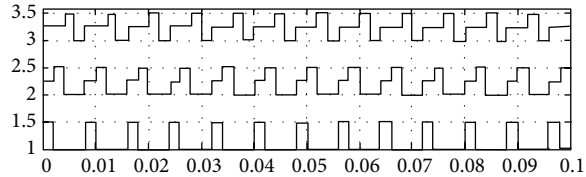
Figure 6: Simulation results without using fuzzy feedback scheduling.

Table 5: Mapping relations between resource utilization and tasks period.

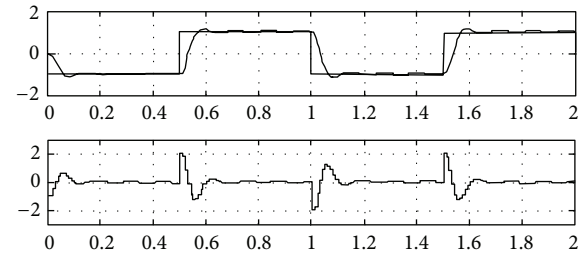
U	Task period
0.2	4
0.4	8
0.6	16

greater importance. In this regard, the task with ID "16202" and that with ID "7363" stand at the first place, even though they have longer deadline. But the deadline of the task with ID "12856" is right behind the aforementioned two, thus making this task prior to that with ID "8679." The task with ID "8679" performs worst. So it is prior to the tasks with ID "16202" and "7363." With the smallest deadline, the task with ID "16202" is prior to that with ID "7363."

4.2. Period Adjustment Examples. A robot control system simulation platform based on CPS architecture is developed by means of toolbox of Matlab/TrueTime, which is shown in Figure 5. The module consists of three major parts: global scheduling server, server nodes. and network. The toolbox



(a)



(b)

Figure 7: Simulation results using fuzzy feedback scheduling.

of TrueTime Kernel contributes to this integrated model. The global scheduling server separately controls the three actuators; each close-loop control system represents a task as T_i. Sensors send the outputs of actuators to the global scheduling server which then computes the priority and adjusts the period according to the feedback information.

A comparison study of using and not using fuzzy feedback scheduling is conducted. Suppose the interpolation period of three control loops is 4 ms. The three tasks cannot be scheduled without using fuzzy feedback scheduling. The task execution is shown in Figure 6(a) and the output curve of three actuators and controller in Figure 6(b). By using fuzzy feedback scheduling, the interpolation period of three control loops is adjusted to 8 ms. The three tasks can be scheduled. The task execution is presented in Figure 7(a) and the output curve of three actuators and controller in Figure 7(b). Obviously, the fuzzy feedback scheduling method improves task schedulability and reduces the error of robot control system.

5. Experimental Results

There is one global scheduling server and two server nodes. They control two 6-DOF mechanical arms under such operating environment: YANXIANG EC3-1711CLDNA IPC; CPU: Genuine Intel(R) processor 600 Mhz; Memory: 599 Mhz, 480 MB; Hard Disk: 60 G; Operating System: Windows XP, Ardence RTX 8.1. The experimental operation object (Figure 8) is six-degree-of-freedom robot arm of Googol Tech (GRB3016).

5.1. Real-Time Test and Analysis. After testing a dozen sets of data, the response time of the control tasks is recorded in Figure 9. As Figure 9 shows, the average response time of control task is 8.323 ms with the presence of fuzzy feedback scheduling method and 10.2387 ms with the absence of the scheduling method. The average response time is much shorter when the scheduler is in place. It implies that the scheduling module can, to some extent, mitigate delay

Figure 8: 6-DOF mechanical arm.

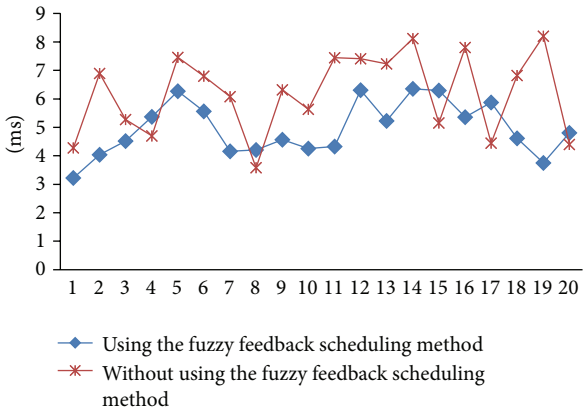


Figure 9: The response time of the control task.

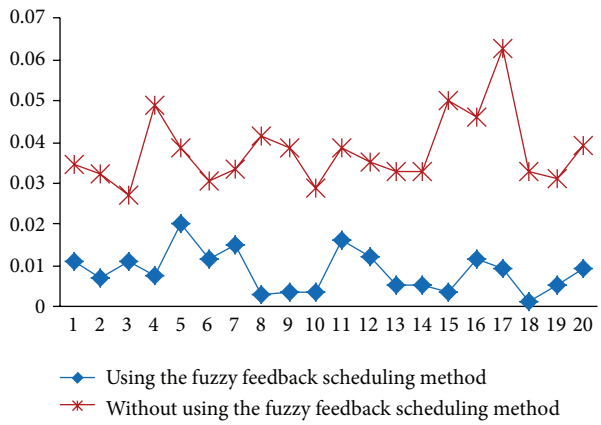


Figure 10: Packet loss rate test without using the scheduling module.

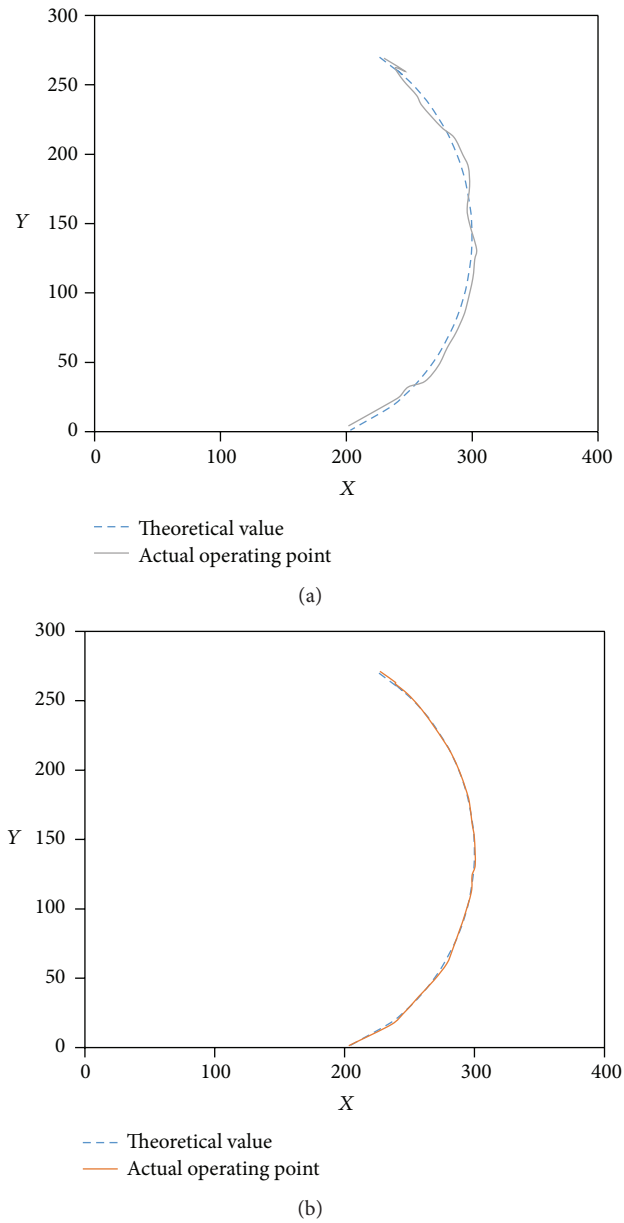


Figure 11: Experimental trajectory of circular interpolation.

between tasks caused by transmission congestion and ensures the real-time performance of the system.

5.2. Packet Loss Ratio Test. The deadline of the tasks is set at 20 ms and a test program for interference is available. Figure 10 shows a total packet loss ratio of 3.69% without using the scheduling method and 0.76% with the use of the scheduling method. The latter is significantly lower than the former. The experiment proves that the scheduling method not only ensures lower packet loss ratio but also effectively improves the utilization.

5.3. Comparison of Operating Error. Three teaching points are selected: (200.482, 282.073), (299.878, 141.112), and (200, 0). There are 39 interpolation points. The experimental trajectory of circular interpolation is described in Figure 11. Figure 11(a) depicts the experimental results without using fuzzy feedback scheduling method and Figure 11(b) with the use of of scheduling method. The dashed line refers to the theoretical value while the full line represents the actual operating point.

From the error graph (Figure 12), it is learned that the maximum error is 1.86 mm with the use of the fuzzy feedback scheduler. The manipulator can accord to the predetermined path in a smaller error range. On the contrary, not using the fuzzy feedback scheduler results in the maximum error of 5.879 mm. Evidently, the fuzzy feedback scheduling method can reduce the error of robot control system and improve the system's performance.

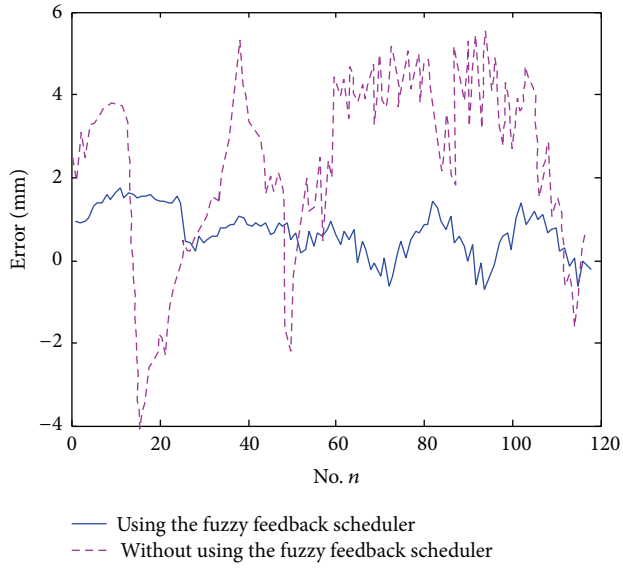


FIGURE 12: Error curve of circular interpolation.

6. Conclusion

The fuzzy feedback scheduling method in this paper can be applied to the cyber-physical systems for robot control. The proposed two-level scheduling framework dynamically adjusts the priority and period. Priority is adjusted according to the task's deadline, its importance, and performance feedback, while the period is adjusted according to the total resource utilization. The method controls the traffic of the tasks to ensure schedulability, reduce packet loss and delay, and realize sounder performance compared with the system with a fixed period. The experimental results indicate that the method can effectively handle slow response and high packet loss caused by the introduction of bus into the network.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61262013), Science and Technology Program of Guangzhou (2013Y2-00100), and the Fundamental Research Funds for the Central Universities (2014ZM0048).

References

[1] J. Wan, H. Yan, H. Suo, and F. Li, "Advances in cyber-physical systems research," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 11, pp. 1891–1908, 2011.

[2] E. A. Lee and S. A. Seshia, *Introduction to Embedded Systems—A Cyber-Physical Systems Approach*, LeeSeshia.org, 2011.

[3] J. F. Wan, D. Li, Y. Q. Tu, P. Zhang, and F. Li, "A survey of cyber physical systems," in *Proceedings of the IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, Kunming, China, March 2011, http://www.jiafuwan.net/publications.html.

[4] J. Z. Li, H. Gao, and B. Yu, "Concepts, features, challenges, and research progresses of CPSs," *Development Report of China Computer Science*, pp. 1–17, 2009.

[5] Y. Tan, M. C. Vuran, and S. Goddard, "Spatio-temporal event model for cyber-physical systems," in *Proceedings of the 29th IEEE International Conference on Distributed Computing Systems Workshops (ICDCS '09)*, pp. 44–50, Montreal, Canada, June 2009.

[6] H. Huang, Y. Sun, Q. Yang et al., "Integrating neuromuscular and cyber systems for neural control of artificial legs," in *Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS '10)*, pp. 129–138, ACM, April 2010.

[7] X. Koutsoukos, N. Kottenstette, J. Hall et al., "Passivity-based control design for cyber-physical systems," in *Proceedings of the International Workshop on Cyber-Physical Systems-Challenges and Applications (CPS-CA '08)*, Santorini Island, Greece, 2008.

[8] M. Chen, J. Wan, and F. Li, "Machine-to-machine communications: architectures, standards and applications," *KSII Transactions on Internet and Information Systems*, vol. 6, no. 2, pp. 480–497, 2012.

[9] C. Zou, J. Wan, M. Chen, and D. Li, "Simulation modeling of cyber-physical systems exemplified by unmanned vehicles with WSNs navigation," in *Proceedings of the 7th International Conference on Embedded and Multimedia Computing Technology and Service*, pp. 269–275, Gwangju, Republic of Korea, 2012.

[10] J. Wan, M. Chen, F. Xia, D. Li, and K. Zhou, "From machine-to-machine communications towards cyber-physical systems," *Computer Science and Information Systems*, vol. 10, no. 3, pp. 1105–1128, 2013.

[11] Object Management Group, *Real-Time CORBA Specification*, Version 1.1, 2002.

[12] Y. Zhang, C. Gill, and C. Lu, "Reconfigurable real-time middleware for distributed cyber-physical systems with aperiodic events," in *Proceedings of the 28th International Conference on Distributed Computing Systems (ICDCS '08)*, pp. 581–588, IEEE, Beijing, China, July 2008.

[13] OMG, *Real-Time CORBA Specification Version 2.0*, 2003.

[14] J. F. Wan, D. Li, H. H. Yan, and P. Zhang, "Fuzzy feedback scheduling algorithm based on central processing unit utilization for a software-based computer numerical control system," *Journal of Engineering Manufacture*, vol. 224, no. 7, pp. 1133–1143, 2010.

[15] J. F. Wan and D. Li, "Fuzzy feedback scheduling algorithm based on output jitter in resource-Constrained embedded systems," in *Proceedings of the International Conference on Challenges in Environmental Science and Computer Engineering (CESCE '10)*, pp. 457–460, Wuhan, China, March 2010.

[16] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *Journal of the Association for Computing Machinery*, vol. 20, no. 1, pp. 46–61, 1973.

[17] K. M. Zuberi and K. G. Shin, "Scheduling messages on controller area network for real-time CIM applications," *IEEE Transactions on Robotics and Automation*, vol. 13, no. 2, pp. 310–316, 1997.

[18] Z.-P. Chen and H.-Q. Xu, "Research on network control system using improved EDF dynamic scheduling algorithm," *Advanced Materials Research*, vol. 403-408, pp. 2420–2423, 2012.

[19] J. Wang, "A scheduling algorithm based on communication delay for wireless network control system," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 20, pp. 3891–3895, 2012.

[20] O. Esquivel-Flores, H. Benítez-Pérez, P. Méndez-Monroy, and J. Ortega-Arjona, "Bounded communication between nodes of a networked control system as a strategy of scheduling," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 27, no. 6, pp. 481–502, 2012.

[21] X. Zhu, P.-C. Huang, S. Han, A. K. Mok, D. Chen, and M. Nixon, "MinMax: A sampling interval control algorithm for process control systems," in *Proceedings of the 18th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA '12)*, pp. 68–77, Seoul, South Korea, August 2012.

[22] F. Xia, X. Dai, Z. Wang, and Y. Sun, "Feedback based network scheduling of networked control systems," in *Proceedings of the 5th International Conference on Control and Automation (ICCA '05)*, pp. 1231–1236, Budapest, Hungary, June 2005.

[23] Z.-X. Li, W.-L. Wang, B.-C. Lei, and H.-Y. Chen, "Message scheduling based on fuzzy feedback in networked control systems," *Acta Automatica Sinica*, vol. 33, no. 11, pp. 1229–1232, 2007.

[24] D. Simon, D. Robert, and O. Sename, "Robust control/scheduling co-design: application to robot control," in *Proceedings of the 11th IEEE Real Time and Embedded Technology and Applications Symposium*, pp. 118–127, San Francisco, Calif, USA, March 2005.

[25] S.-G. Liu, J.-M. Wei, and Z.-C. Zhu, *Fuzzy Control Technology*, China Textile Press, Beijing, China, 2001.

[26] Q.-G. Han and X.-Q. Wu, *Computer Fuzzy Control Technology and Instrumentation*, China Measurement Press, Beijing, China, 1999.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 5908459, 9 pages
http://dx.doi.org/10.1155/2016/5908459

Research Article

Estimation of Individual Cylinder Air-Fuel Ratio in Gasoline Engine with Output Delay

Changhui Wang and Zhiyuan Liu

Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, China

Correspondence should be addressed to Changhui Wang; wang_changhui@126.com

Received 1 March 2016; Revised 23 June 2016; Accepted 14 July 2016

Academic Editor: José A. Somolinos

The estimation of the individual cylinder air-fuel ratio (AFR) with a single universal exhaust gas oxygen (UEGO) sensor installed in the exhaust pipe is an important issue for the cylinder-to-cylinder AFR balancing control, which can provide high-quality torque generation and reduce emissions in multicylinder engine. In this paper, the system dynamic for the gas in exhaust pipe including the gas mixing, gas transport, and sensor dynamics is described as an output delay system, and a new method using the output delay system observer is developed to estimate the individual cylinder AFR. With the AFR at confluence point augmented as a system state, an observer for the augmented discrete system with output delay is designed to estimate the AFR at confluence point. Using the gas mixing model, a method with the designed observer to estimate the individual cylinder AFR is presented. The validity of the proposed method is verified by the simulation results from a spark ignition gasoline engine from engine software enDYNA by Tesis.

1. Introduction

Cylinder-to-cylinder air-fuel ratio (AFR) balancing control in internal combustion engines with multiple cylinders is one of the technology trends to satisfy the increasingly stringent emission regulations, which can also improve the engine performance, such as thermal efficiency and fuel economy. The AFR of each cylinder is decided by the aspirated air mass, the injected fuel mass, and the residual gas from the prior cycle, in which the combustion stroke of each cylinder sequentially occurs along the rotational angle of the crankshaft. Due to the air breathing variability and injector variability, there exists AFR imbalance between cylinders, leading to adverse impacts on emission performance using the conventional controllers [1, 2].

In order to improve the AFR control accuracy, there has been a great deal of research that focuses on the AFR control of individual cylinders [3–10]. In fact, the estimation of the individual cylinder AFR with a single universal exhaust gas oxygen (UEGO) sensor installed in the exhaust pipe is one of the key technology trends for the individual cylinder AFR control. The digital filtering techniques are employed to extract the AFR imbalance signals from oxygen sensor voltage signals in [11], in which the oxygen sensor voltage signal is processed to determine imbalanced cylinder identification and AFR cylinder imbalance levels. In [12], a modeling method to estimate the individual fuel-gas ratio is proposed to estimate AFR, which is used for an adaptive generalized predictive control approach to balance the individual cylinder characteristics in the static engine operation mode. A static steady state observer based on the individual cylinder AFR model along the air mass flow, gas mixing, and sensor dynamics in an exhaust manifold can be found in [13]. In the diesel engines, a nonlinear observer is proposed to estimate the individual cylinder AFR [14]. In [15], a PI compensator is designed to compensate the cylinder-by-cylinder variations, in which an input observer is proposed to estimate individual cylinder AFR.

However, the transport delay and sensor delay from the exhaust confluence point to UEGO sensor output are ignored in the proposed observers from the above papers, which may reduce the accuracy of the AFR estimation of each cylinder. In order to improve the individual cylinder AFR estimation accuracy, the system dynamics in the exhaust pipe including gas transport and sensor dynamics is described as an augmented discrete system with output delay in this paper, in which the AFR at confluence point is augmented as

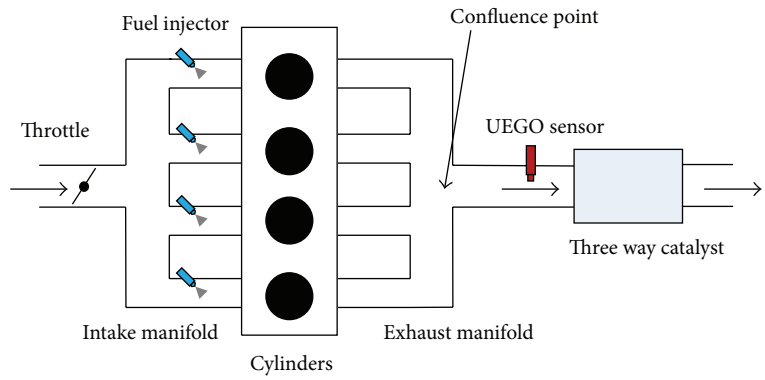


Figure 1: Schematic of the 4-cylinder SI gasoline engine.

a system state. Then, an observer for the augmented discrete system with output delay is designed. With the combination of the designed observer and the gas mixing model at confluence point, the method to estimate the individual cylinder AFR is presented. The performance of the proposed method is validated against the simulation result from engine software enDYNA provided by Tesis, and a comparison with existing method is given during an urban driving cycle, which demonstrates that the proposed method can improve the accuracy of the individual cylinder AFR estimation.

This paper is organized as follows. In Section 2, the system dynamics in the exhaust pipe including gas transport and UEGO sensor dynamics is described as an augmented system with output delay. In Section 3, an observer for the output delay system is designed, and the method to estimate the individual cylinder AFR is presented. Simulation results from enDYNA are presented in Section 4, and the conclusions are summarized in Section 5.

2. Problem Formulation

A schematic diagram of a 4-cylinder spark-ignited (SI) gasoline engine is shown in Figure 1, where the fuel injectors equipped at the inlet port near to the intake valve are controlled individually. The fuel mass burnt in each cylinder is injected by the corresponding injector, and the fuel injection command is delivered to the injector of each cylinder serially along the crank angle. The AFR of each cylinder is $\lambda_i = m_{\text{cyli}}/m_{\text{air}}$, $i = 1, 2, 3, 4$, where m_{cyli} is the fuel mass into the cylinder and m_{air} is the air mass into the cylinder.

After combustion, the combusted gas of each cylinder is exhausted into the corresponding runner during the exhaust stroke of each cylinder and passes through their runners and confluence in the public exhaust manifold. Then, the AFR of the mixed gas is measured by a UEGO sensor, and the mixed gas runs to the outside passing through the catalyst. The system dynamics in the exhaust pipe includes the gas mixing, gas transport, and sensor dynamics, in which the transfer function from the confluence point to sensor output can be given by [13]

$$G_{\text{exh}}(s) = \frac{\lambda_{\text{sen}}(s)}{\lambda_c(s)} = \frac{e^{-\delta s}}{(1+\tau_{\text{mix}}s)(1+\tau_{\text{sen}}s)}, \quad (1)$$

where λ_c is the AFR at the exhaust confluence point, λ_{sen} is the measured AFR of the UEGO sensor, $\delta = \delta_{\text{mix}} + \delta_{\text{sen}}$ is the time delay including the transport delay δ_{mix} and sensor delay δ_{sen}, τ_{mix} is the time constant of the mixing process, and τ_{sen} is the sensor time constant.

For an engine with 4 cylinders, the combined UEGO sensor signal is sampled at the exhaust top dead center of each cylinder, and the sampling period T_s is related to the engine cycle period as $T_s = T_c/4$, in which the engine cycle period is $T_c = 120/n_e$, where n_e is the engine speed in rpm. Consisting of the zero-order holder and the mixing and sensor dynamics, the discrete form of model (1) with the sampling period T_s can be given by [13]

$$Q_{\text{exh}}(z) = z^{-m}\left(1 + a\frac{z-1}{z-\alpha_{\text{mix}}} + b\frac{z-1}{z-\alpha_{\text{sen}}}\right), \quad (2)$$

where

$$\begin{aligned} \alpha_{\text{mix}} &= e^{-T_s/\tau_{\text{mix}}}, \\ \alpha_{\text{sen}} &= e^{-T_s/\tau_{\text{sen}}}, \\ m &= \frac{\delta}{T_s}, \\ a &= \frac{-\tau_{\text{mix}}}{\tau_{\text{mix}} - \tau_{\text{sen}}}, \\ b &= \frac{\tau_{\text{sen}}}{\tau_{\text{mix}} - \tau_{\text{sen}}}, \\ z &= e^{T_s s}. \end{aligned} \quad (3)$$

Furthermore, (2) can be written as a state-space form with input delay [15]:

$$\begin{aligned} x_q(l+1) &= A_q x_q(l) + B_q \lambda_c(l-m), \\ \lambda_{\text{sen}}(l) &= C_q x_q(l), \end{aligned} \quad (4)$$

where

$$A_q = \begin{pmatrix} \alpha_{\text{mix}} & 0 \\ 0 & \alpha_{\text{sen}} \end{pmatrix},$$

$$B_q = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$C_q = \begin{pmatrix} a(\alpha_{\text{mix}} - 1) & b(\alpha_{\text{sen}} - 1) \end{pmatrix},$$

$$x_q(l) = \begin{pmatrix} x_{q1}(l) \\ x_{q2}(l) \end{pmatrix},$$

$$x_{q1}(l) = \frac{\lambda_{\text{sen}}(l) - b(\alpha_{\text{sen}} - 1) x_{q2}(l)}{a(\alpha_{\text{mix}} - 1)}, \tag{5}$$

$$x_{q2}(l) = \frac{1}{(\alpha_{\text{sen}} - \alpha_{\text{mix}}) b(\alpha_{\text{sen}} - 1)} (\alpha_{\text{sen}} \lambda_{\text{sen}}(l) - \alpha_{\text{mix}} \alpha_{\text{sen}} \lambda_{\text{sen}}(l-1) + (1 - 2a\alpha_{\text{sen}} - 2b\alpha_{\text{mix}} - \alpha_{\text{mix}} - \alpha_{\text{sen}} + a + b) \cdot \lambda_c(l-m) + (a\alpha_{\text{sen}} + b\alpha_{\text{mix}} + \alpha_{\text{mix}} \alpha_{\text{sen}}) \cdot \lambda_c(l-m-1)).$$

Define the new system state as $\overline{x}_q(l) = x_q(l+m)$, and (4) with input delay can also be rewritten as the following discrete system with output delay:

$$\begin{aligned} \overline{x}_q(l+1) &= A_q \overline{x}_q(l) + B_q \lambda_c(l), \\ \lambda_{\text{sen}}(l) &= C_q \overline{x}_q(l-m). \end{aligned} \tag{6}$$

$\lambda_c(l)$ in (6) is unknown, which can be considered as a system state due to the small rate of change of $\lambda_c(l)$. Then, an augmented discrete system with output delay from (6) can be obtained:

$$\begin{aligned} x(l+1) &= Ax(l), \\ y(l) &= Cx(l-m), \end{aligned} \tag{7}$$

where

$$x(l) = \begin{pmatrix} \overline{x}_q(l) \\ \lambda_c(l) \end{pmatrix},$$

$$A = \begin{pmatrix} \alpha_{\text{mix}} & 0 & 1 \\ 0 & \alpha_{\text{sen}} & 1 \\ 0 & 0 & 1 \end{pmatrix}, \tag{8}$$

$$y(l) = \lambda_{\text{sen}}(l),$$

$$C = \begin{pmatrix} a(\alpha_{\text{mix}} - 1) & b(\alpha_{\text{sen}} - 1) & 0 \end{pmatrix}.$$

Equation (7) indicates that the estimation of the AFR λ_c at the exhaust confluence point becomes the state estimation of the discrete system with output delay (7).

3. Observer Design for Discrete System with Output Delay

The observer for the output delay system (7) is given:

$$\begin{aligned} \widehat{x}(l+1) &= A\widehat{x}(l) + L(y(l) - C\widehat{x}(l-m)), \\ \widehat{y}(l) &= C\widehat{x}(l-m), \end{aligned} \tag{9}$$

where $\widehat{x} \in \mathbb{R}$ is the state estimate and $L \in \mathbb{R}^{3\times 1}$ is the feedback gain matrix. The asymptotical stability of the proposed observer (9) is analyzed in the following theorem.

Theorem 1. *There exist matrices L, $P = P^T > 0$, $Q = Q^T \geq 0$, and $Z = Z^T \geq 0$, such that the following linear matrix inequality (LMI) is feasible:*

$$\Xi = \begin{pmatrix} -P + Q + M^T + M & -M^T + N & A^T P & M^T & (A^T - I) Z \\ -M + N^T & -Q - N^T - N & -C^T L^T P & N^T & -C^T L^T Z \\ PA & -PLC & -P & 0 & 0 \\ M & N & 0 & -m^{-1} Z & 0 \\ Z(A - I) & -ZLC & 0 & 0 & m^{-1}(Z - 2P) \end{pmatrix} < 0; \tag{10}$$

then observer (9) is asymptotically stable.

Proof. Set the estimation error $e(l) = \widehat{x}(l) - x(l)$, and the error dynamic system between (7) and (9) is obtained:

$$e(l+1) = Ae(l) - LCe(l-m). \tag{11}$$

Denote $\eta(l) = e(l+1) - e(l)$; then we have

$$\begin{aligned} e(l-m) &= e(l) - \sum_{n=l-m}^{l-1} \eta(n), \\ \eta(l) &= (A - I) e(l) - LCe(l-m). \end{aligned} \tag{12}$$

Choose a Lyapunov functional candidate as

$$V(l) = e^T(l) Pe(l) + \sum_{i=l-m}^{l-1} e^T(i) Qe(i) + \sum_{j=-m+1}^{0} \sum_{i=l-1+j}^{l-1} \eta^T(i) Z\eta(i). \tag{13}$$

Define $\Delta V = V(l+1) - V(l)$; then along the solution of (11) and (12) we have

$$\Delta V(l) = \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix}^T \cdot \begin{pmatrix} -P + Q + A^T PA & -A^T PLC \\ -C^T L^T PA & -Q + C^T L^T PLC \end{pmatrix} \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix} + m\eta^T(l) Z\eta(l) - \sum_{n=l-m}^{l-1} \eta^T(n) Z\eta(n). \tag{14}$$

According to Lemma 1 in [16], for any constant matrix $Z > 0$, M, and N, the following inequality holds:

$$-\sum_{n=l-m}^{l-1} \eta^T(n) Z\eta(n) \le \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix}^T \cdot \begin{pmatrix} M^T + M & -M^T + N \\ -M + N^T & -N^T - N \end{pmatrix} \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix} + m \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix}^T \cdot \begin{pmatrix} M^T \\ N^T \end{pmatrix} Z^{-1} \begin{pmatrix} M & N \end{pmatrix} \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix}. \tag{15}$$

With the combination of (14) and (15), we have

$$\Delta V(l) \le \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix}^T \Omega \begin{pmatrix} e(l) \\ e(l-m) \end{pmatrix}, \tag{16}$$

where

$$\Omega = \begin{pmatrix} -P + Q + M^T + M + A^T PA + m(A-I)^T Z(A-I) + mM^T Z^{-1} M & -M^T + N - A^T PLC - m(A-I)^T ZLC + mM^T Z^{-1} N \\ * & -Q - N^T - N + C^T L^T PLC - mC^T L^T ZLC + mN^T Z^{-1} N \end{pmatrix}. \tag{17}$$

By Schur complement [17], the following LMI (18) guarantees $\Omega < 0$, which can guarantee $\Delta V(l) < 0$ and the asymptotical stability of observer (9):

$$\Pi = \begin{pmatrix} -P + Q + M^T + M & -M^T + N & A^T P & M^T & (A^T - I) Z \\ -M + N^T & -Q - N^T - N & -C^T L^T P & N^T & -C^T L^T Z \\ PA & -PLC & -P & 0 & 0 \\ M & N & 0 & -m^{-1} Z & 0 \\ Z(A-I) & -ZLC & 0 & 0 & -m^{-1} Z \end{pmatrix} < 0. \tag{18}$$

Now, condition (10) guaranteeing (18) must be proved. Define $W = \mathrm{diag}(I, I, I, I, PZ^{-1})$, and we have

$$W^T \Pi W = \begin{pmatrix} -P + Q + M^T + M & -M^T + N & A^T P & M^T & (A^T - I) P \\ -M + N^T & -Q - N^T - N & -C^T L^T P & N^T & -C^T L^T P \\ PA & -PLC & -P & 0 & 0 \\ M & N & 0 & -m^{-1} Z & 0 \\ P(A-I) & -PLC & 0 & 0 & -m^{-1} PZ^{-1} P \end{pmatrix}. \tag{19}$$

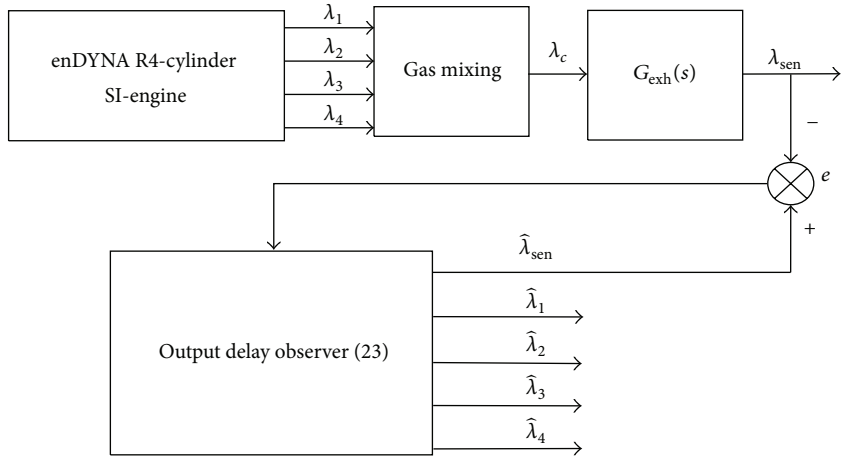


Figure 2: Schematic diagram for the estimation of the individual cylinder AFR.

Because of the fact that $(P - Z)Z^{-1}(P - Z) \geq 0$, we have $-PZ^{-1}P \leq Z - 2P$. Therefore, condition (10) can guarantee $W^{\mathrm{T}}\Pi W < 0$; then (18) holds.

The estimation of the AFR λ_c at the exhaust confluence point can be obtained according to observer (9). In order to obtain the AFR of each cylinder, the relationship between the AFR of each cylinder and the AFR at the exhaust confluence point is analyzed in the following.

The combusted gas of each cylinder is discharged into the corresponding exhaust port and flows to the exhaust confluence point in the exhaust manifold, in which we assume that exhaust gas mixing in the individual exhaust runner can be neglected. Hence, the AFR in the exhaust runner is constant during one engine cycle, and the AFR λ_c at the exhaust confluence point in the kth engine cycle can be given by [15]

$$\begin{aligned}\lambda_c\left(kT_c + (i-1)T_s\right) &= \frac{\sum_{j=1}^{i}\dot{m}_{cj}\left(kT_c + (i-1)T_s\right)\lambda_j(k)}{\sum_{l=1}^{4}\dot{m}_{cl}\left(kT_c + (i-1)T_s\right)} \\ &+ \frac{\sum_{j=1}^{4}\dot{m}_{cj}\left(kT_c + (i-1)T_s\right)\lambda_j(k-1)}{\sum_{l=1}^{4}\dot{m}_{cl}\left(kT_c + (i-1)T_s\right)},\end{aligned} \tag{20}$$

where $\dot{m}_{ci}$ is the exhaust air mass flow in the ith exhaust manifold passing through the confluence point and λ_i is the AFR in the ith cylinder. Furthermore, under the assumption that air mass aspirated into cylinders in each cycle is constant, exhaust air flow has the same shape between successive cycles; then a periodic function can be obtained:

$$\begin{aligned}\gamma_{ij}(k) &= \frac{\dot{m}_{cj}\left(kT_c + (i-1)T_s\right)}{\sum_{l=1}^{4}\dot{m}_{cl}\left(kT_c + (i-1)T_s\right)} \\ &= \frac{\dot{m}_{cj}\left((i-1)T_s\right)}{\sum_{l=1}^{4}\dot{m}_{cl}\left((i-1)T_s\right)}.\end{aligned} \tag{21}$$

Therefore, the gas mixing behavior (20) can be rewritten in the T_s domain as

$$\lambda_c(l) = \sum_{i=0}^{3}\gamma_{[l][l-i]}\lambda_{[l-i]}(l-i), \tag{22}$$

where $[l] = (i \bmod 4) + 1$. The relationship between the AFR of each cylinder and the AFR at the exhaust confluence point can be obtained by (22).

The algorithm to estimate the individual cylinder AFR is as follows: First, the AFR λ_c at the exhaust confluence point is obtained according to observer (9). Then, the individual cylinder AFR can be calculated through (22). With the combination of (9) and (22), the method for the estimation of each cylinder AFR can be given as follows:

$$\begin{aligned}\hat{x}(l+1) &= A\hat{x}(l) + L\left(y(l) - C\hat{x}(l-m)\right), \\ \hat{\lambda}_c(l) &= \begin{pmatrix}0 & 0 & 1\end{pmatrix}\cdot\hat{x}(l), \\ \hat{\lambda}_{[l]}(l) &= \frac{1}{\gamma_{[l][l]}}\left(-\sum_{i=1}^{3}\left(\gamma_{[l][l-i]}\cdot\hat{\lambda}_{[l-i]}(l-i)\right) + \hat{\lambda}_c(l)\right).\end{aligned} \tag{23}$$

□

4. Simulation Studies

In this section, the simulation study of the estimation of the individual cylinder AFR is presented in the environment of a 2.0 L 4-cylinder SI gasoline engine from enDYNA [18, 19]. The enDYNA is a professional software tool for the real-time simulation of internal combustion engines, providing ready-to-use models for all common engine types comprising crank angle synchronous combustion, gas path, fuel system, cooling system, drivetrain, driver, and soft-ECU. The R4-cylinder SI-engine is an example in enDYNA to simulate a 4-cylinder SI gasoline engine, whose specifications are given in Table 1. The observer architecture is illustrated in Figure 2.

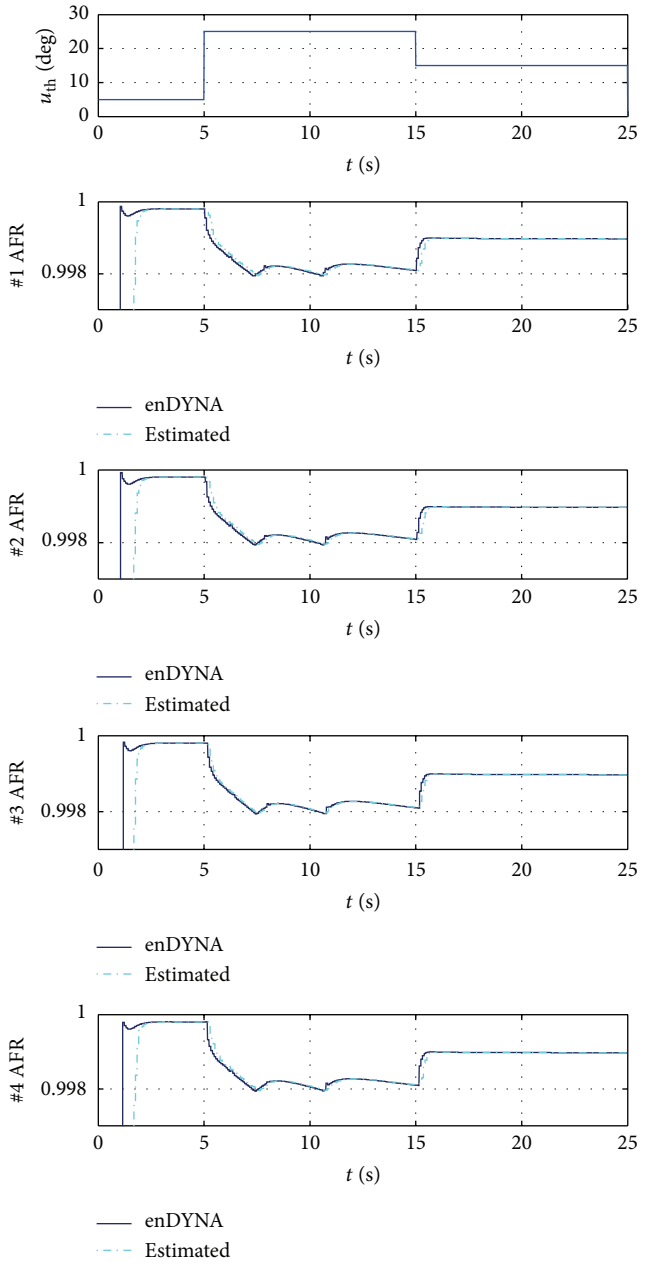


Figure 3: Individual cylinder AFR estimation results.

Table 1: Engine specifications.

Fuel system	Direct injection
Displacement (m^3)	0.002
Intake manifold volume (m^3)	0.004
Exhaust manifold volume (m^3)	0.0015
Max engine speed (rad/s)	785

The parameters of the system dynamics in the exhaust pipe (2) and the gas mixing (22) are presented as follows [15]:

$$\tau_{\text{mix}} = 0.01 \text{ s},$$

$$\tau_{\text{sen}} = 0.12 \text{ s},$$

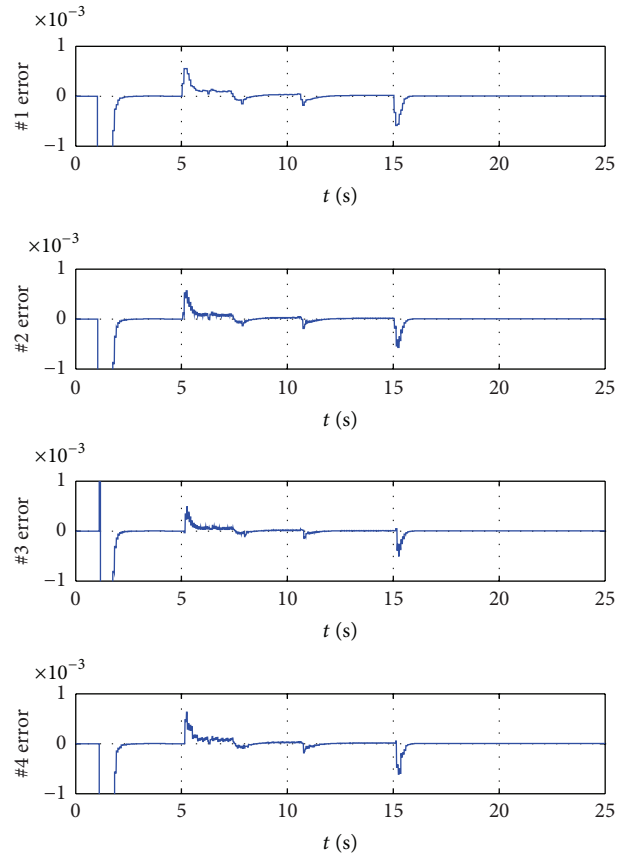


Figure 4: Evolution of the estimation error.

$$m = 1,$$

$$T_c = 0.08 \text{ s},$$

$$T_s = 0.02 \text{ s},$$

$$\Gamma = \begin{pmatrix} 0.55 & 0.05 & 0.15 & 0.25 \\ 0.25 & 0.6 & 0.05 & 0.1 \\ 0.15 & 0.25 & 0.5 & 0.1 \\ 0.05 & 0.1 & 0.15 & 0.7 \end{pmatrix}. \tag{24}$$

Then, the system matrix can be obtained:

$$A = \begin{pmatrix} 0.1353 & 0 & 1 \\ 0 & 0.8465 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \tag{25}$$

$$C = \begin{pmatrix} -0.0786 & 0.1675 & 0 \end{pmatrix}.$$

According to the inequality (10), the gain matrix can be given by $L = \begin{pmatrix} 0.97 & 3.79 & 0.73 \end{pmatrix}^T$.

Here, the input of the throttle angle in enDYNA is designed as a step signal presented in Figure 3. Accordingly, the estimation results of the individual cylinder AFR by the proposed method are shown in Figure 3, and the estimation errors are plotted in Figure 4. When the throttle changes

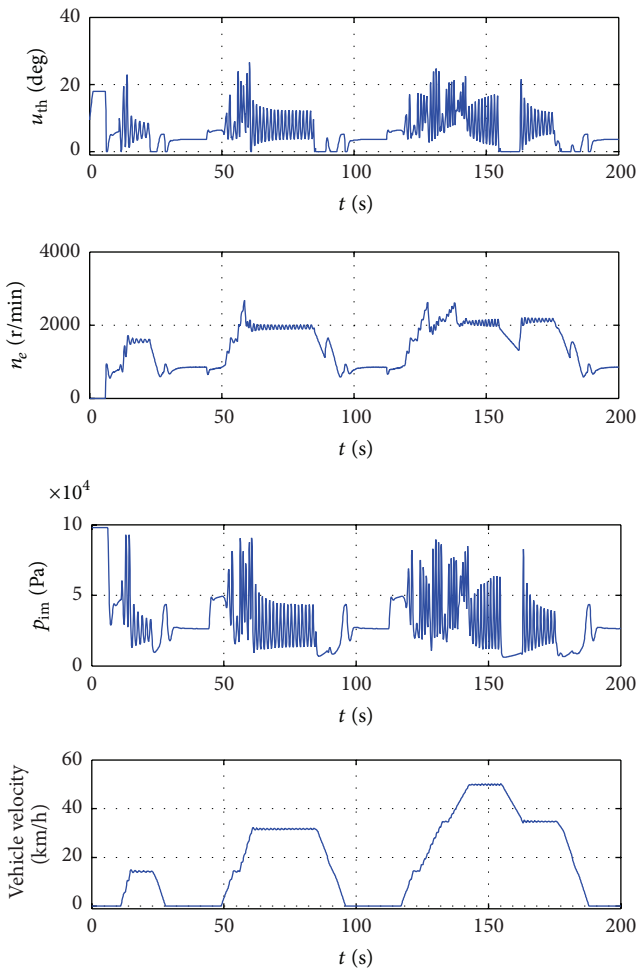


FIGURE 5: Evolution of throttle angle u_{th}, engine speed n_e, intake manifold pressure p_{im}, and vehicle velocity under ECE cycle.

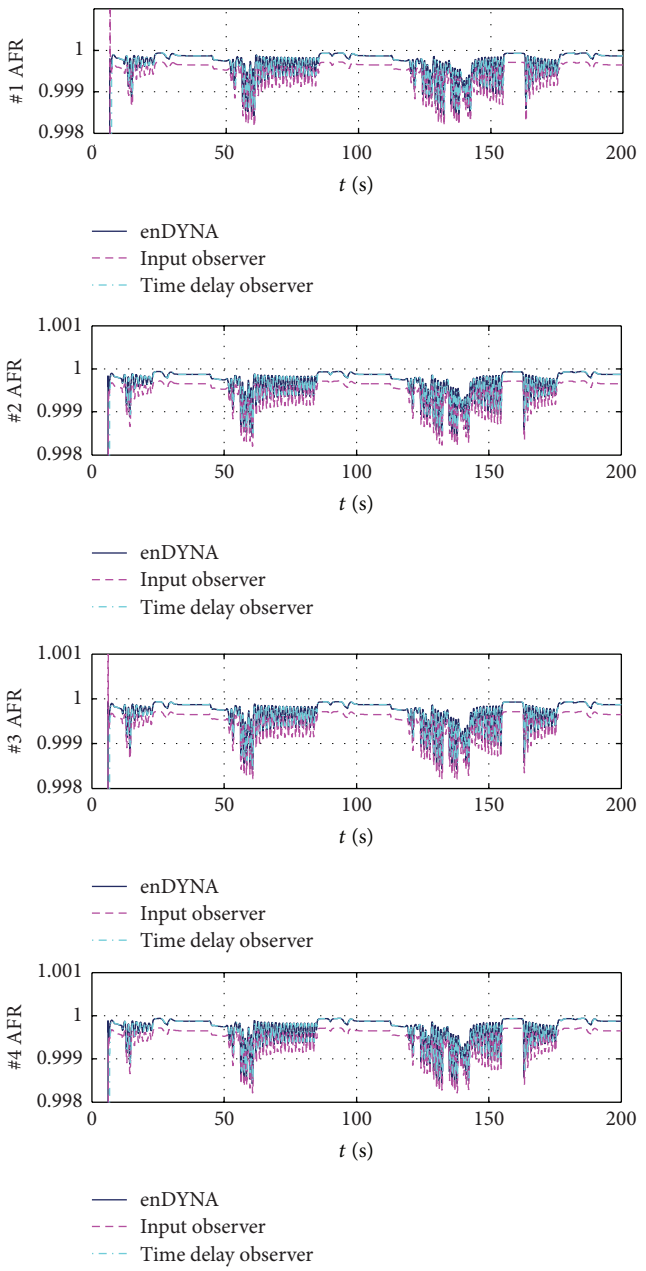


FIGURE 6: Individual cylinder AFR estimation results of the proposed method and unknown input observer under ECE cycle.

suddenly, the estimation errors of the individual cylinder AFR are about 0.06%, and then the steady state errors tend to decay within 0.01%.

In order to verify the effectiveness of the proposed method under driving cycle condition, one segment of the urban driving cycle ECE (Economic Commission for Europe) is used [19], under which the throttle angle u_{th}, engine speed n_e, intake manifold pressure p_{im}, and vehicle velocity are plotted in Figure 5. Accordingly, the comparison of the individual cylinder AFR estimation between the proposed method and the input observer in [15] are presented in Figure 6, and the estimation errors are plotted in Figure 7. Clearly, the error of the proposed method is smaller than the input observer when the AFR changes slowly, in which the steady state error of the input observer is 0.03%. When the AFR changes severely, there exist fluctuations of the AFR estimation error from both the proposed method and input observer in [15]. However, the estimation error from the proposed method is smaller. It is demonstrating that the proposed method considering time delay can improve the accuracy of the individual cylinder AFR estimation.

5. Conclusion

An efficient method for the estimation of the individual cylinder AFR with a single UEGO sensor was developed to improve the estimation accuracy. The system dynamics in the exhaust pipe was described as an augmented discrete system with output delay, in which the AFR at confluence point was augmented as a system state and beneficial to be estimated comparing the system with input delay. Then, an observer for the augmented system with output delay was designed to estimate the AFR at confluence point, which can avoid accurately inverting the engine model including delays. Using the gas mixing model, a method to estimate

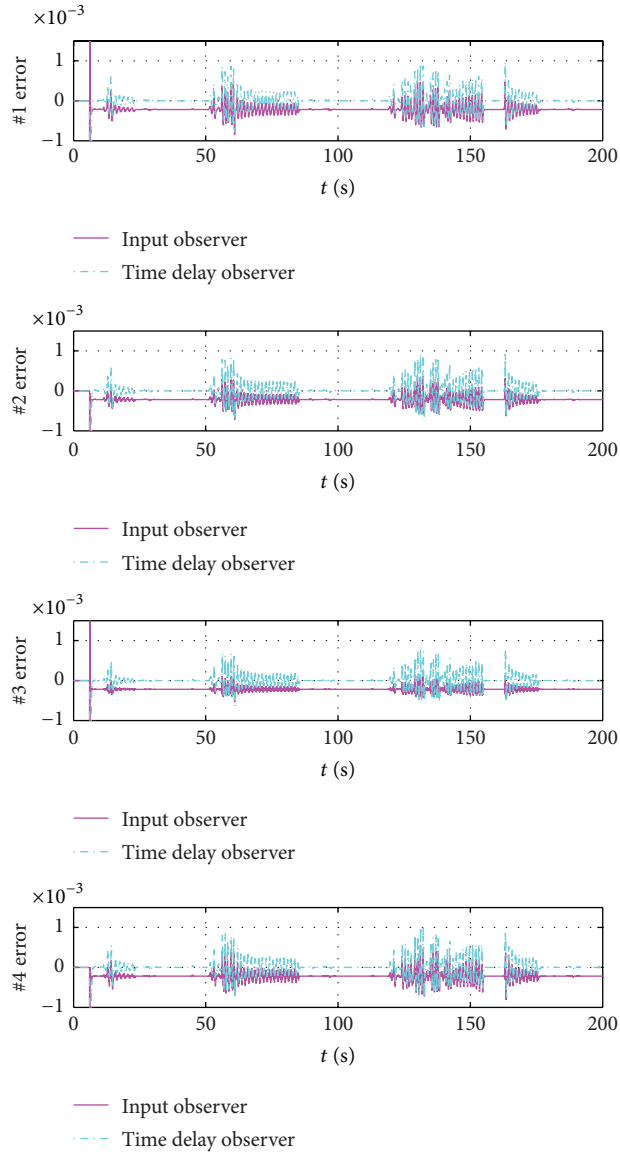


FIGURE 7: Evolution of the estimation error of the proposed method and unknown input observer.

the individual cylinder AFR based on the proposed observer was presented. The performance of the proposed method was validated by the simulation data from engine software enDYNA provided by Tesis, and a comparison with existing method was obtained during ECE cycle, demonstrating that the proposed method considering time delay from exhaust gas transport and UEGO sensor dynamics can improve the accuracy of the individual cylinder AFR estimation.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by China Automobile Industry Innovation and Development Joint Fund (no. U1564213).

References

[1] J. B. Heywood, *Internal Combustion Engine Fundamentals*, vol. 930, McGraw-Hill, New York, NY, USA, 1988.

[2] J. W. Grizzle, K. L. Dobbins, and J. A. Cook, "Individual cylinder air-fuel ratio control with a single EGO sensor," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 1, pp. 280–286, 1991.

[3] J. Chauvin, P. Moulin, G. Corde, N. Petit, and P. Rouchon, "Real-time nonlinear individual cylinder air fuel ratio observer on a diesel engine test bench," in *Proceedings of the 16th Triennial World Congress of International Federation of Automatic Control (IFAC '05)*, pp. 194–199, Prague, Czech Republic, July 2005.

[4] J. Chauvin, P. Moulin, G. Corde, N. Petit, and P. Rouchon, "Kalman filtering for real-time individual cylinder air fuel ratio observer on a diesel engine test bench," in *Proceedings of the American Control Conference*, pp. 1886–1891, IEEE, Minneapolis, Minn, USA, June 2006.

[5] J. Chauvin, N. Petit, and P. Rouchon, "Six degrees crankshaft individual air fuel ratio estimation of diesel engines for cylinder balancing purpose," Sae Technical Papers, 2006.

[6] Y. Liu and T. Shen, "Modeling and experimental validation of air-fuel ratio under individual cylinder fuel injection in gasoline engines," *IEEJ Journal of Industry Applications*, vol. 1, no. 3, pp. 155–163, 2012.

[7] J. F. Burkhard, "Individual cylinder fuel control for a turbocharged engine," SAE Technical Paper 0148-7191, SAE International, 2014.

[8] H. Li, Y. Huang, G. Li, and Y. Yang, "Research on the cylinder-by-cylinder variations detection and control algorithm of diesel engine," SAE Technical Paper 0148-7191, 2015.

[9] S. Nakagawa, A. Numata, and T. Hori, "Individual cylinder control for air-fuel ratio cylinder imbalance," SAE Technical Paper 2015-01-1624, 2015.

[10] M. Kassa, C. Hall, A. Ickes, and T. Wallner, "In-cylinder oxygen mass fraction estimation method for minimizing cylinder-to-cylinder variations," SAE Technical Paper 0148-7191, 2015.

[11] N. Qiao, C. Krishnamurthy, and N. Moore, "Determine air-fuel ratio imbalance cylinder identification with an oxygen sensor," *SAE International Journal of Engines*, vol. 8, no. 3, pp. 1005–1011, 2015.

[12] K. Suzuki, T. Shen, J. Kako, and S. Yoshida, "Individual A/F estimation and control with the fuel-gas ratio for multicylinder IC engines," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 4757–4768, 2009.

[13] L. Benvenuti, M. D. Di Benedetto, S. Di Gennaro, and A. Sangiovanni-Vincentelli, "Individual cylinder characteristic estimation for a spark injection engine," *Automatica*, vol. 39, no. 7, pp. 1157–1169, 2003.

[14] J. Chauvin, G. Corde, P. Moulin, N. Petit, and P. Rouchon, "High frequency individual cylinder estimation for control of diesel engines," *Oil and Gas Science & Technology*, vol. 61, no. 1, pp. 57–72, 2006.

[15] B. He, T. Shen, J. Kako, and M. Ouyang, "Input observer-based individual cylinder air-fuel ratio control: modelling, design and validation," *IEEE Transactions on Control Systems Technology*, vol. 16, no. 5, pp. 1057–1065, 2008.

[16] W.-A. Zhang and L. Yu, "Stability analysis for discrete-time switched time-delay systems," *Automatica*, vol. 45, no. 10, pp. 2265–2271, 2009.

[17] S. P. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, vol. 15, SIAM, 1994.

[18] TESIS DYNAware, *en-DYNA® THERMOS® 2.0 Block Reference Manual*, 2006.

[19] T. DYNAware, *en-DYNA® THERMOS® 2.0 User Manual*, 2006.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 4250746, 8 pages
http://dx.doi.org/10.1155/2016/4250746

Research Article

Multipath Load Balancing Routing for Internet of Things

Chinyang Henry Tseng

Computer Science and Information Engineering, National Taipei University, New Taipei 237, Taiwan

Correspondence should be addressed to Chinyang Henry Tseng; tsengcyt@gm.ntpu.edu.tw

Received 22 May 2016; Revised 23 June 2016; Accepted 27 June 2016

Academic Editor: Antonio Fernández-Caballero

In the next-generation technology, Internet of Things (IoT), billions of smart objects will communicate with one another to make human lives more convenient. IoT is based on wireless sensor network (WSN), and Zigbee is one of the most popular WSN protocols. A mature IoT environment involves heavy WSN data transmission causing bottleneck problems. However, Zigbee's AODV routing stack does not have load balance mechanism to handle bursty traffic. Therefore, we develop Multipath Load Balancing (MLB) Routing to substitute Zigbee's AODV routing. Our proposed MLB consists of two main designs: LAYER_DESIGN and LOAD_BALANCE. LAYER_DESIGN assigns nodes into different layers based on node distance to IoT gateway. Nodes can have multiple next-hops delivering IoT data. All neighboring layer nodes exchange flow information containing current load, used by LOAD_BALANCE to estimate future load of next-hops. With MLB, nodes can choose the neighbors with the least load as the next-hops and thus can achieve load balance and avoid bottlenecks. Compared with Zigbee's AODV and multipath version AODV (AOMDV), experiment results demonstrate that MLB achieves better load balance, lower packet loss rate, and better routing connectivity ratio in both grid and random uniform topologies. MLB provides a more convincing routing solution for IoT applications.

1. Introduction

Recently, the demands of Internet of Things (IoT) [1] keep growing. In the beginning, wireless sensor network (WSN) [2] enables ubiquitous sensing technologies. As the WSN technology evolves, the proliferation and application of these sensing devices create the Internet of Things (IoT). IoT is the next revolution, where the interconnection among smart objects creates an intelligent environment. It is estimated and expected to reach 24 billion IoT devices by 2020 [1]. As more and more IoT devices are connected and communicated, IoT applications generate tremendous IoT traffic. Since IoT traffic is for the communication between objects, the transmission reliability is critical, especially in a relatively unstable WSN, compared with wired network. As Figure 1 shows, IoT technology is applied in many domains, including environmental monitoring, transportation, automotive vehicles, industry [3], medical technology [4], healthcare, smart home [5], and smart city [6].

WSN is the most essential component of IoT, which comprises everything of WSN plus a thick layer of software installed across computational devices and the cloud. In other words, IoT is developed based on WSN, in which Zigbee is one of the most popular WSN protocols. In IoT, the low end sensors rely on WSN where data is transmitted from sensors (things) to the sink node (IoT gateway) using a multihop fashion. More static and mobile sinks can be deployed to collect data from sensors. Multiple sensor networks are connected together over the Internet. Therefore, performing data management is important. IoT research needs to find more efficient and effective ways of data management, such as collecting, modelling, reasoning, and distribution. We focus on data transmission reliability between things and IoT gateways.

We focus on Zigbee instead of Wifi because Zigbee is more health-friendly. Since IoT makes humans surrounded by wireless-connected objects, it is important to make all smart objects with low radio transmission power to make the environment healthier. Zigbee is with low transmission power, 1 mW, and is an appropriate option for IoT. Zigbee stack [7] adopts Ad Hoc On-Demand Distance Vector (AODV) [8] to automatically construct an ad hoc network as routes are needed. AODV optimizes routing paths to be the

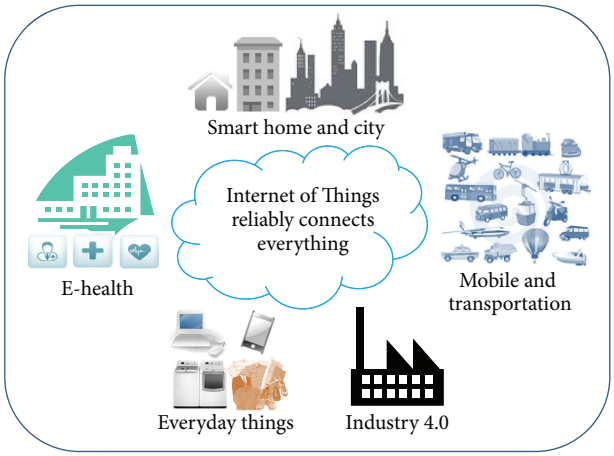


Figure 1: IoT application domains.

shortest but does not support multipath routing. Multipath routing is important to perform load balancing by selecting the less busy channel as the next-hop when network is under heavy traffic. In addition, once traffic bottleneck occurs, the unsuccessful delivery will trigger AODV route error (RERR) messages, which might generate more REER messages to jam the network. In the worst case, excessive AODV RERR messages can paralyze the network, particularly the links close to Zigbee sink. Therefore, we intend to enhance Zigbee routing by substituting AODV with our proposed routing protocol, Multipath Load Balancing, MLB, Routing.

In order to provide a reliable routing service for data-intensive IoT applications, we propose Multipath Load Balancing (MLB) Routing. Instead of cluster design, we use distributed architecture for MLB to avoid the situation that cluster heads become bottlenecks. MLB takes traffic load as the cost function and adaptively updates load information with neighbors to calculate the least busy routes. MLB offers multiple paths for the next-hop options to enhance reliability while evenly distributing traffic. The main compared target of MLB is Ad Hoc On-Demand Multipath Distance Vector (AOMDV) [9], which is a multipath version of AODV by giving equivalent paths with the same hop counts to IoT gateway. The shortcoming of AOMDV is that its equivalent paths must be thoroughly disjoint and cannot share nodes on their distinct paths. This limits the number of available alternatives. AOMDV does not consider traffic load when selecting the sending path because it is designed simply for multiple paths without considering load balancing issue. To become a better multipath solution for load balancing than AOMDV to support data-intensive IoT services, MLB has a reliable layered architecture and utilizes traffic load as the cost function. Layered architecture allows the routing computation to be done among neighbors locally and taking traffic load as the cost solves imbalanced load more directly than AOMDV. Thus, MLB can enhance network reliability by providing multiple next-hops and guarantees the shortest paths selected.

MLB consists of two main components: *LAYER_DESIGN* and *LOAD_BALANCE*. In LAYER_DESIGN, IoT gateway is the top level and we define that nodes *closer* to IoT gateway are in the *inner* layers and nodes *farther* from IoT gateway are in the *outer* layers. Each sensor node may play both roles of outer-layer and inner-layer nodes depending on the relative distance to IoT gateway compared with their neighbors. Each outer-layer node only needs to know the *local* information of next-hop nodes in the immediate inner layer to IoT gateway and the path from source to IoT gateway is constructed hop by hop. This structured and inductive two-layer relationship establishes the reliable routing service. Furthermore, LOAD_BALANCE allows each outer-layer node to calculate which inner-layer node is with the least traffic load. Consequently, the inner-layer node with less traffic is selected as the next-hop to IoT gateway. Through the cooperation of LAYER_DESIGN and LOAD_BALANCE, load balancing optimization is accomplished. In MLB routing table, multiple paths are recorded and allow more fault-tolerance once some next-hop fails. In case that any node fails to operate normally, MLB allows outer-layer nodes to recalculate their best inner-layer nodes toward IoT gateway without broadcasting route error messages. Therefore, MLB can quickly adapt sensor nodes to dynamic flow change and malfunctioned links.

The main contributions of MLB are multipath routing with load balancing, robustness, and reliability. First, load balancing is done by selecting the best inner-layer node with the least traffic load. Second, robustness is achieved because the synergy of LAYER_DESIGN and LOAD_BALANCE provides multiple inner-layer next-hops to IoT gateway for each outer-layer nodes, and ROUTE_RECOVERY can detect link failure for quick link switch. Since MLB eliminates bottlenecks by load balancing design and provides multipath routing, MLB provides a much more reliable routing service than current Zigbee's AODV related solutions.

To evaluate the performance of load balancing and reliability, simulation results are demonstrated based on three evaluation metrics, Load Balance Degree (LBD), packet loss rate (PLR), and connectivity ratio (CR). LBD illustrates load balancing performance for each layer in the routing topology and shows MLB balances network traffic more effectively than AODV and AOMDV, especially in the first layer. PLR directly shows the reliability of data delivery and CR shows the reliability of the entire routing topology. Because of the effectiveness of load balancing, PLR in MLB is much lower than PLR in AODV and a little lower than PLR in AOMDV. CR shows the routing service in MLB is very stable with perfect connectivity with larger among of data traffic. On the other hand, CR shows the routing services in AODV and AOMDV suffering from different degrees of disconnections due to the unbalanced traffic loading. Therefore, simulation results show that MLB provides better load balancing with more reliable packet delivery comparing with Zigbee's AODV and AOMDV.

The rest of this paper is organized as follows. Section 2 discusses the related works. Section 3 illustrates how Zigbee-MLB assists IoT communication. Section 4 presents MLB. Section 5 demonstrates simulation results. Finally, Section 6 concludes this paper.

2. Related Works

To avoid traffic congestion occurring at some bottleneck point, many load balance wireless routings are proposed. Pure cluster solutions [10, 11] aim at building as uniformly small-sized clusters as possible to achieve cluster load balancing. In minimum radius clustering algorithm [10], each cluster head extracts and sends local information to sink. After aggregating information from all cluster heads, SINK generates a balancing module to achieve cluster load balancing. Zhang and Yang [11] proposed a distributed algorithm to ensure the mean square deviation value of the number of nodes within each cluster as small as possible.

Advanced cluster solutions [12, 13] balance energy consumption on each node and indirectly achieve load balancing. Liao et al. [12] proposed a load-balanced clustering algorithm, DSBCA, for wireless sensor networks. To build a more balanced clustering structure and avoid forming large clusters, DSBCA considers the connectivity density and distance between nodes and base station. In each cluster, the node with the highest weight is selected as the cluster head. Weight calculations include the residual energy of node, the initial energy node, and the times of the node being elected as cluster head. During data transmission, each cluster head needs to aggregate all the data from its cluster members and send to the base station. However, data is sent to the base station through the cluster head which might become another potential bottleneck. Wu and Liu [13] proposed a centralized power efficient routing algorithm, EHGUC-OARP, for energy harvesting-wireless sensor networks. The base station uses EHGUC algorithm to form clusters of unequal sizes and select cluster heads of all formed clusters. When EHGUC is applied, the clusters with smaller size are closer to the BS. Subsequently, the base station uses OAPR to construct an optimal routing among all cluster heads.

Multipath Energy Aware AODV (ME-AODV) routing [14] utilizes network topology to divide the network into one or more logical clusters and restricts the flooding of route request outside the cluster. ME-AODV uses the remaining battery power of the nodes as the cost function instead of the hop count used in traditional AODV and adds multipath concept. The node uses all available paths in a round robin fashion in order to evenly distribute the energy consumption over the entire network. ME-AODV still has the drawback that cluster headers and cluster gateways have heavier load and become the bottleneck candidate. Neighbor-aware Adaptive Load Balancing algorithm [15] uses the information of parent and child nodes along with a probability factor to balance the traffic and prolongs network lifetime. All nodes send traffic load information to the gateway, which subsequently calculates and broadcasts the determined probability factor used by the whole network.

Different from mentioned solutions above, MLB uses traffic load as the cost function and noncluster structure can prevent potential bottlenecks occurring at cluster heads. MLB does not consider energy in its design because we focus on the IoT applications where sensors contain more energy and higher computation ability to send large amount of network data. Therefore, reliable routing services become more critical than saving energy for MLB.

3. How MLB Assists IoT Communication

As shown in Figure 2, WSN is the most essential component of IoT. IoT comprises everything that WSN plus a thick layer of software installed across computational devices and the cloud. Therefore, IoT can be explained as a general purpose WSN. In other words, WSN is a part of the IoT while IoT is not a part of SN. With regard to IoT communication, IoT follows the architecture of a three-layer WSN. Data is transmitted from Phase I sensors (things) to Phase II sink node (IoT gateway) using a multihop fashion. More static and mobile sinks can be deployed to collect data from sensors. WSN data is then sent to Phase III computational devices for further data analysis and IoT applications. Multiple sensor networks may be connected together over PHASE-IV-Internet.

Our work focuses on the reliability of data transmission of Phase I to successful data collection of PHASE II. For example, hospital may let patients wear Electrocardiography sensors to real-time monitor patients' heart health. Large amount of Electrocardiography data is continuously transmitted over WSN and cannot be lost because it involves human life. Only successful data transmission and collection of Phases I and II can provide computational devices complete data for correct data analysis and application. To find more efficient and effective ways of data transmission, we focus on data transmission reliability between things and IoT gateways (sensors and sinks).

MLB is proposed to cooperate with a large-scale wireless Zigbee network. When the traffic load increases significantly due to large number of IoT objects, routers with more neighbors might experience much heavier traffic load and become the bottlenecks, especially for the routers close to IoT gateway. To prevent bottlenecks from happening, MLB guides Zigbee routers to select the next-hops with the least traffic loads. Each MLB router in a Zigbee network serves network traffic with equal possibility to prevent itself from becoming a potential bottleneck.

Take a closer look at MLB in Figure 3. MLB provides an alternative routing service for Zigbee network without modifying existing Zigbee stack. Once a Zigbee router forwards IoT data, MLB guides Zigbee network layer to choose a next-hop with the least load toward IoT gateway. Besides, MLB also ensures that the current router's next-hop is closer to IoT gateway than the router itself to prevent routing loops. Therefore, MLB design can easily cooperate with the existing Zigbee stack.

4. MLB: Multipath Load Balancing Routing

As IoT applications grow rapidly, IoT sensors may deliver massive and critical data to IoT gateway so reliable IoT routing service is highly desirable. Current solutions, such as Zigbee and relative works in Section 1, cannot avoid bottlenecks, which may paralyze the entire network if the network traffic grows and congestion occurs nearby IoT gateway. To solve this problem, we propose Multipath Load Balancing (MLB) Routing to provide a reliable routing service for IoT applications, particularly data-intensive applications. Compared with AODV's improved multipath version, AOMDV, MLB

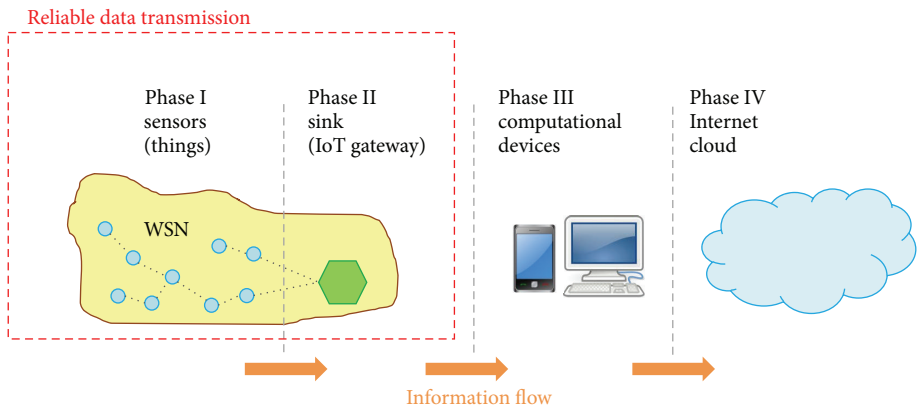


Figure 2: Architecture of Internet of Things.

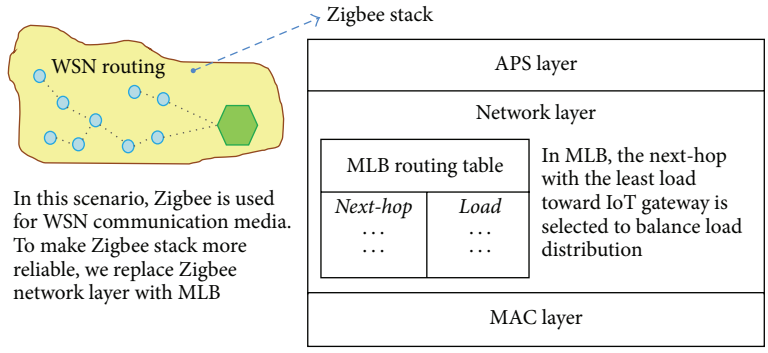


Figure 3: MLB routing in Zigbee stack.

has the same advantage of multipath but has better traffic load distribution and network reliability, MLB consists of two main designs: LAYER_DESIGN and LOAD_BALANCE. MLB is introduced from these three aspects in the following.

4.1. LAYER_DESIGN. When a node forwards a data packet to the gateway, it requires a routing service to generate the best next-hop choice for data forwarding. If the chosen next-hop fails to operate, a traditional single-hop routing service requires long response time to compute a new next-hop. Multipath routing services provide several next-hop choices so these services can quickly provide a new next-hop in case of the failure of current next-hop. As a result, multipath routing services provide a more flexible and reliable data forwarding services than traditional single-hop routing services.

For IoT applications, sensors require a reliable routing service to forward sensor data to the IoT gateway. In MLB, LAYER_DESIGN provides a reliable multipath data forwarding service with simple layered routing design. In LAYER_DESIGN, layer value presents the number of hops to the IoT gateway for each node. Layer 1 nodes, which are 1 hop away from the IoT gateway, have direct wired connections to the IoT gateway to avoid the gateway becoming a bottleneck. If all layer 1 nodes send packets via Zigbee wireless links to the gateway for a period of time, the gateway will encounter traffic jams because of shared media among wireless links. Besides, data traffic usually accumulates at layer 1 nodes so wired links are required. Then layer 1 nodes can collect data from other nodes through Zigbee wireless links. The IoT gateway can have several layer 1 nodes collecting data from Zigbee links simultaneously without becoming a bottleneck itself.

To allow other sensor nodes to join LAYER service, layer 1 nodes broadcast their beacon messages to present layer 1 routing service. Other nodes hearing these beacon messages from layer 1 nodes can claim their layer number as 2 and announce their layer number through their beacon messages to present layer 2 routing service. Then layer 3 routing service can be presented in the same way and so on. A beacon message consists of its address, layer value, and network loading. The usage of network loading is defined in LOAD_BALANCE. While receiving beacons from neighbors, each node updates its neighbor table, which records neighbors' information including address, layer value, and network loading. Through beacon messages and neighbor tables, nodes can quickly establish their LAYER services toward the IoT gateway.

Figure 4 shows an example topology for presenting LAYER service. For layer 2 nodes, since layer 1 nodes are closer to the gateway, layer 1 becomes inner layer for layer 2 and layer 2 becomes outer layer for layer 1. Again, layer 2 can be the inner layer for layer 3 so a node may play both roles of outer-layer and inner-layer nodes in different relationships. Nodes can forward their packets to their neighboring inner-layer nodes as their next-hops until the packets reach the gateway.

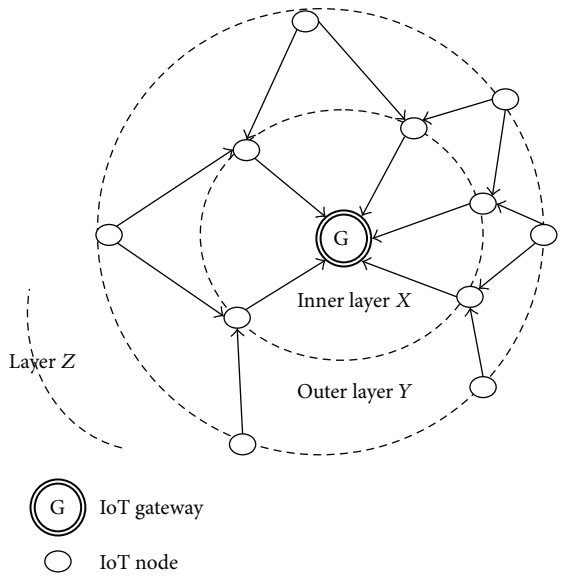


Figure 4: Multipath layer routing (LAYER_DESIGN) in MLB.

Since each node may have several neighboring inner-layer nodes, it can have multiple next-hops to forward packets. Again, these next-hops may have multiple forwarding choices from their inner-layer nodes so multipath routing establishes. For example, layer 3 node may have 3 neighboring layer 2 nodes, and each of these layer 2 nodes may have 3 neighboring layer 1 nodes, and thus 9 possible paths exist for this layer 3 node. If one of these paths becomes unavailable, the rest of 8 paths can be used. For traditional single path routing protocols, such as AODV, the node can have 1 path at a time, and it must generate another path in case this path becomes unavailable. Therefore, LAYER_DESIGN can provide a lot more paths than AODV and provide more reliable routing service. In addition, AOMDV requires that its multipaths must be disjoint paths. In other words, these paths cannot share the same nodes so AOMDV may have less available paths than LAYER_DESIGN according to network topologies.

In LAYER_DESIGN, given a node X, if one of the inner nodes becomes unavailable, node X can still use other inner nodes as its next-hop so LAYER_DESIGN can quickly adjust its path locally. If all inner nodes become unavailable, node X searches for layer values of other nodes in its neighbor table. At this time, the nodes with the highest layer value are usually peer of nodes for node X, and node X uses them as its new inner layer. Then node X updates its layer value, which is usually larger than its old value by 1 and announces it in its new beacon message immediately.

4.2. LOAD_BALANCE. To accomplish load balancing, each node chooses the next-hop node in the inner layer with lowest network loading while forwarding data packets. Each node announces its network loading in its beacon message to allow its outer-layer nodes to retrieve its network loading value. As a node forwards a data packet, the node chooses the next-hop with the lowest network loading among neighboring nodes in the inner layer. Since each data packet is forwarded to the node with the lowest network loading, LOAD_BALANCE is done based on LAYER_DESIGN service.

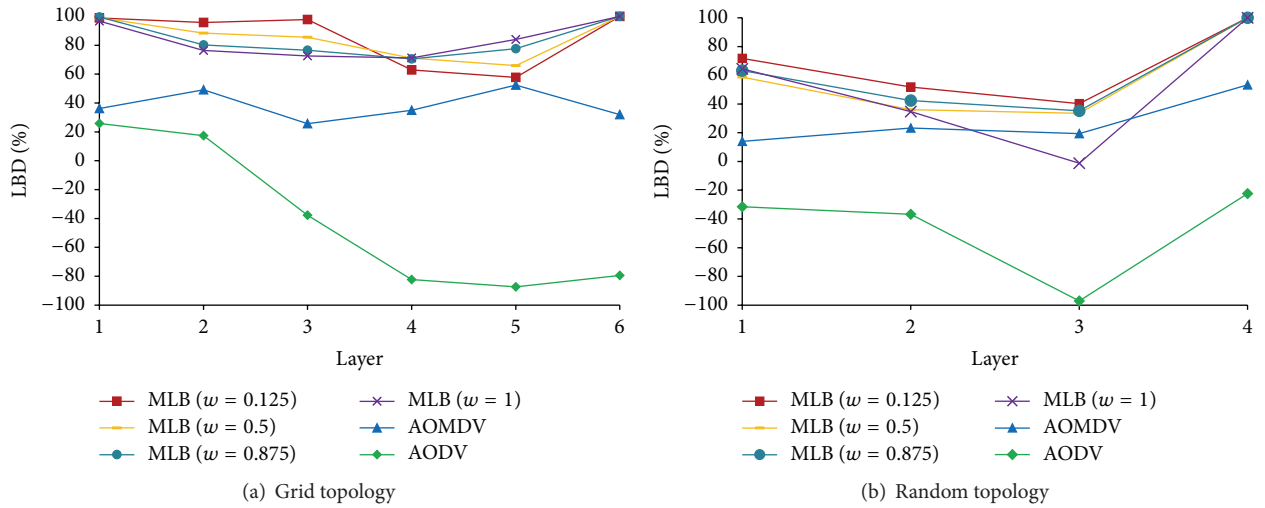
If the network loading is determined based on current network loading of a node during a short period of time, nodes change their next-hops too often. This can trigger network loading dramatically and cause potential bottlenecks. So LOAD_BALANCE determines the network loading by the Estimated Network Loading based on exponential weighted moving average formula [16], in which newer data has heavier weighting and higher influence on next estimation value and the influence of data decreases exponentially with time. Therefore, the estimated loading can reflect a long term network loading so nodes switch next-hops smoothly.

Given a time slot x, its Estimated Network Loading (ENL) is denoted as ENL_x, and current Sample Network Loading (SNL) is denoted as SNL_x. If SNL_x is not 0, ENL_x becomes $(1 - w) * \mathrm{ENL}_{x-1} + w * \mathrm{SNL}_x$, where w is the weight of SNL to determine the influence of current traffic load for long term traffic estimation. If SNL_x is 0, ENL_x is set to 1/2 ENL_{x-1} to prevent ENL_x from becoming 0 in case w is 1. ENL_0 is initially set to SNL_0. If w is large, SNL has high impact of ENL so ENL changes fast, and ENL changes slow if w is small. Experiment results in Section 5.1 can show the impact of w to determine the best practice for LOAD_BALANCE. By using LOAD_BALANCE upon LAYER_DESIGN, traffic goes through nodes with lowest long term traffic loading dynamically. Therefore, MLB accomplishes loading balancing with reliable multipath layered routing.

5. Experiment Results for Evaluating MLB

Section 5 presents experiment results evaluating MLB compared with AOMDV and AODV based on the three evaluation metrics, LBD, PLR, and CR, which are presented in the following subsections. The experiment platform is ns2 with the following parameters: simulation time is 300 seconds, MAC layer is 802.15.4, field size is 250 meters × 250 meters, and transmission range is 50 meters. Data traffic type is constant bit rate, data packet size is 100 bytes, and data sending interval is 1 second. In order to show data traffic patterns in different types of network topologies, the experiment is conducted in a grid topology with 85 nodes and a topology with 100 nodes deployed by random uniform distribution. Layer 1 nodes have direct wired link with the IoT gateway. In order to demonstrate extreme situations as stress tests, this experiment launches data traffic from half of the nodes in the two topologies. The layer number of the data sending nodes is larger than the layer number of the other nodes, which only forward data traffic. Each topology uses MLB, AOMDV, and AODV separately to evaluate their performance in terms of LBD, PLR, and CR.

5.1. Load Balance Degree. In the experiment, nodes in layer 1 have direct access to the IoT gateway, and nodes from other layers transmit data via Zigbee connections to the nodes in layer 1. So the data traffic accumulated at layer 1 is much larger than other layers, and thus load balance in layer 1 is critical.



(a) Grid topology (b) Random topology

Figure 5: Load Balance Degree in two topologies.

To evaluate the performance of load balance, we present LBD, which is calculated by standard deviation (SD) of traffic load and Load Average (LA) among all nodes in a layer:

$$\mathrm{LBD}_Y = \left[1 - \left(\frac{\mathrm{SD}_Y}{\mathrm{LA}_Y}\right)\right] \times 100\%, \quad (1)$$

where and LBD_Y, SD_Y, LA_Y denote LBD, SD, and LA in the layer Y.

If SD_Y is 0, it shows all nodes in layer Y have the same traffic load. In this case the load balance performance is the best and LBD_Y is 100%. If LBD_Y is 0%, it shows SD_Y equals LD_Y. In other words, the deviation of traffic load in layer Y equals the average load so the load balancing in this layer is poor. If LBD_Y is negative, the load balancing in this layer is even worse because it indicates traffic load of some nodes is even larger than average plus standard deviation.

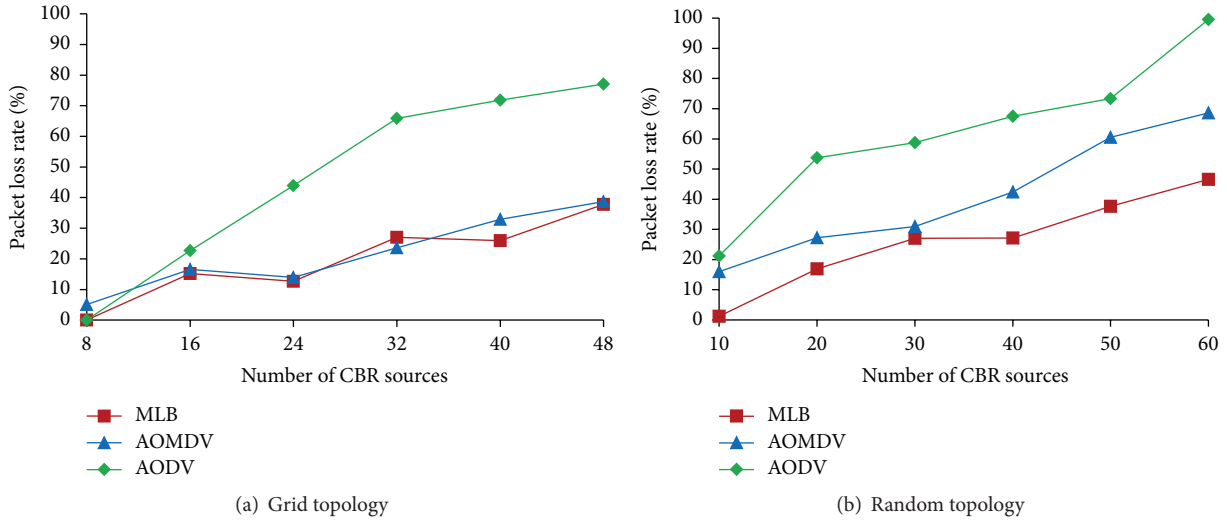
Figure 5 shows LBD in two types of topologies using AODV, AOMDV, and MLB with four different w (w = 0.125, w = 0.5, w = 0.875, and w = 1). Since load balance in layer 1 is critical, LBD_1 is the key observation point. In both grid and random topologies, MLB with w = 0.125 provides the best LBD, particularly the inner layers, layers 1, 2, and 3.

In the grid topology, LBD_1 in MLB cases is 100% so MLB can provide the best load balance performance in the most critical layer, layer 1, which usually accumulates all data traffic. In addition, MLB with w = 0.125 can keep such an optimal load balancing performance through layers 1 to 3. On the other hand, LBD in AOMDV case is about 40% and is much worse than MLB. LBD in AODV case even quickly drops to negative value.

In the random topology, the node distribution is not uniformly deployed compared with the grid topology so the load balance task becomes much more than challenging. The number of connections of grid topology is fixed and the traffic is easier to be predicted and optimized. In general, LBD values in the random topology become worse than those in the grid topology. LBD in MLB with w = 0.125 case is still the best and its LBD_1 is about 75%, which is much better than the case of AOMDV and AODV. Therefore, MLB shows much better load balancing performance than both AOMDV and AODV do. AOMDV provides multiple paths for data traffic so its LBD is much better than AODV, which is used in Zigbee stack. MLB can even provide better load balancing performance than AOMDV. In the grid topology, MLB can even provide the optimal load balancing performance in the inner layers. In the most outer layer, LBD values of all MLB cases are 100 % for both topologies. It shows MLB can perform optimal load balancing performance as data traffic initiates. When data traffic aggregates and arrives at layer 1, MLB can still perform great load balancing performance.

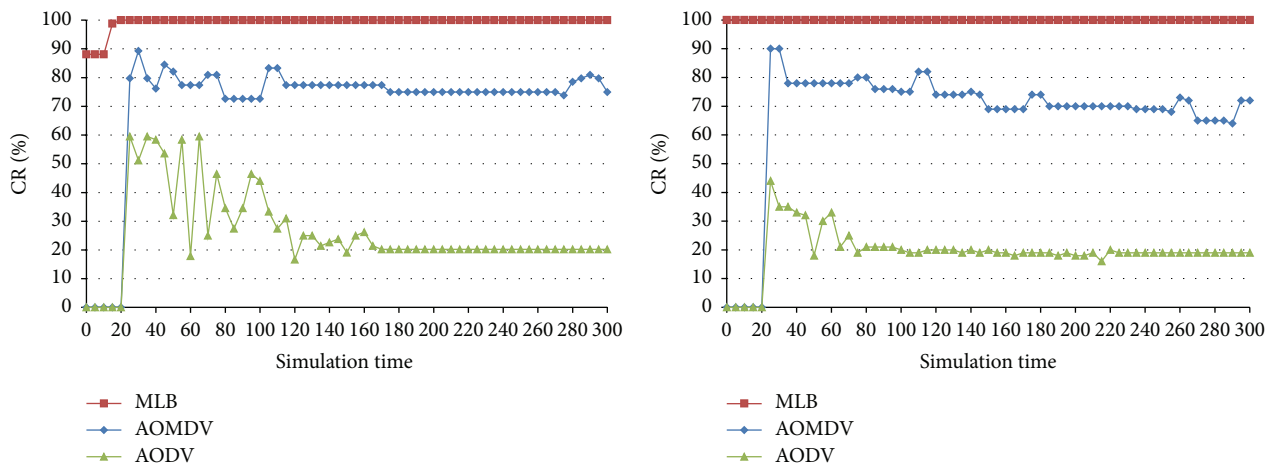
5.2. Packet Loss Rate. PLR directly reflects network performance of IoT applications. As data traffic sources increase, the network loading of the entire network increases. Since data traffic is usually accumulated at layer 1, the network loading of layer 1 can increase significantly in case of poor load balancing performance. So PLR can directly reveal the impact of load balancing performance for IoT applications.

Figure 6 shows PLR of MLB, AOMDV, and AODV versus different number of data traffic sources. Since load balancing performance of AODV is poor, PDR increases significantly as the data traffic sources increase. Because data delivery in the random topology is more challenging than the grid topology due to the node deployment policy, PLR increases much more in the random topology, which is actually a more realistic topology than the grid one. From Figure 6, AODV's PLR increases dramatically so AODV is not a scalable choice for IoT applications. In the grid topology, both AOMDV and MLB have low PLR, but only MLB can still keep low PLR in the random topology. AOMDV's PLR in the random topology can become up to 60% compared with 40% in MLB case. It shows MLB is a scalable solution compared with AOMDV in the challenging random topology. Therefore, according to the results in Figure 6, MLB is a more reliable and scalable



(a) Grid topology (b) Random topology

Figure 6: Packet loss rate.



(a) Grid topology (b) Random topology

Figure 7: Connectivity ratio.

routing solution than AOMDV and AODV for IoT applications.

5.3. Connectivity Ratio. CR is the ratio of nodes having routes toward the IoT gateway over all sensor nodes. CR illustrates the routing connectivity between nodes and the gateway in real time and thus shows the routing reliability of routing protocols for IoT applications. As the network traffic increases, the traffic jams may trigger routing errors. If the routing protocol cannot repair the errors instantly, these routing errors can propagate and trigger more routing error messages, which cause more nodes that are unable to connect the gateway. CR traces the ratio of these nodes that cannot connect to the gateway and monitors the impact of routing errors in real time.

Figure 7 shows CR of MLB, AOMDV, and AODV in 300 seconds. In the beginning, MLB takes shorter time to establish the routes from nodes to the gateway than AOMDV and AODV. Unlike AODV and AOMDV, nodes in MLB just need to connect the upper layer nodes without building a path to the gateway by flooding routing messages, and thus the routing establishing time is shorter in MLB than in AOMDV and AODV. After the routing topology has been stabilized, as half of the nodes send data traffic and the number of nodes is up to 100, CR in MLB is 100% in both topologies. This shows MLB has great routing reliability under the high traffic condition.

On the other hand, CR in AOMDV is about 80%, and this may result from the regulation of AOMDV multipath routing: the routes in AOMDV must have disjoint nodes. This rule may limit the route recovery capability in AOMDV and result in the routing vulnerability. CR in AODV is even worse due to the routing errors caused by the high traffic volume. Since AODV only supports one next-hop for each route, the routing recovery capability is the worst so the routing reliability is the worst. Therefore, the results of CR show that MLB is the most reliable routing protocol among the three routing protocols.

6. Conclusion

As IoT applications grow rapidly, reliable routing is highly desirable to allow IoT sensors delivery data packets to IoT gateway through multihop transmissions accurately. For preventing bottleneck issues in Zigbee's AODV routing services, MLB is proposed to provide a load balancing, robust and reliable routing service for IoT applications. To achieve these goals, MLB consists of LAYER_DESIGN and LOAD_BALANCE. LAYER_DESIGN provides a multipath layer routing service toward IoT gateway for IoT applications, and LOAD_BALANCE estimates load information for data sender choosing the inner-layer next-hop with the least network loading. The synergy of LAYER_DESIGN and LOAD_BALANCE eliminates the bottlenecks and thus provides a load balancing and reliable routing service. The experiment results demonstrate that MLB achieves much better load balancing than AODV and AOMDV according to LBD values. Based on PLR and CR, MLB provides more reliable routing than AODV and AOMDV. In conclusion, based on the load balancing design, MLB provides the most reliable routing service for IoT applications compared with the current famous in-use routing solutions, Zigbee' AODV and its improved multipath version, AOMDV.

Competing Interests

The author declares no competing interests.

References

[1] M. A. Feki, F. Kawsar, M. Boussard, and L. Trappeniers, "The internet of things: the next technological revolution," *Computer*, vol. 46, no. 2, pp. 24–25, 2013.

[2] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[3] L. D. Xu, W. He, and S. Li, "Internet of things in industries: a survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.

[4] Y.-W. Wang, H.-L. Yu, and Y. Li, "Internet of things technology applied in medical information," in *Proceedings of the IEEE International Conference on Consumer Electronics, Communications and Networks (CECNet '11)*, pp. 430–433, April 2011.

[5] G. Chong, L. Zhihao, and Y. Yifeng, "The research and implement of smart home system based on internet of things," in *Proceedings of the International Conference on Electronics, Communications and Control (ICECC '11)*, pp. 2944–2947, IEEE, Zhejiang, China, September 2011.

[6] P. Vlacheas, R. Giaffreda, V. Stavroulaki et al., "Enabling smart cities through a cognitive management framework for the internet of things," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 102–111, 2013.

[7] Zigbee Alliance, *ZigBee Specification Version 1.0*, 2004, http://www.zigbee.org.

[8] C. Perkins, E. Belding-Royer, and S. R. Das, "Ad hoc on-demand distance vector (AODV) routing," IEEE RFC 3561, 2003.

[9] M. K. Marina and S. R. Das, "Ad hoc on-demand multipath distance vector routing," *Wireless Communications and Mobile Computing*, vol. 6, no. 7, pp. 969–988, 2006.

[10] M. Hammoudeh, O. Aldabbas, S. Mount, S. Abuzor, M. Alfawair, and S. Alratrout, "Algorithmic construction of optimal and load balanced clusters in wireless sensor networks," in *Proceedings of the 7th International Multi-Conference on Systems Signals and Devices (SSD '10)*, pp. 1–5, IEEE, Amman, Jordan, 2003.

[11] J. Zhang and T. Yang, "Clustering model based on node local density load balancing of wireless sensor network," in *Proceedings of the 4th International Conference on Emerging Intelligent Data and Web Technologies (EIDWT '13)*, pp. 273–276, Xi'an, China, September 2013.

[12] Y. Liao, H. Qi, and W. Li, "Load-balanced clustering algorithm with distributed self-organization for wireless sensor networks," *IEEE Sensors Journal*, vol. 13, no. 5, pp. 1498–1506, 2013.

[13] Y. Wu and W. Liu, "Routing protocol based on genetic algorithm for energy harvesting-wireless sensor networks," *IET Wireless Sensor Systems*, vol. 3, no. 2, pp. 112–118, 2013.

[14] A. Bhatia and P. Kaushik, "A cluster based minimum battery cost AODV routing using multipath route for ZigBee," in *Proceedings of the IEEE 16th International Conference on Networks (ICON '08)*, pp. 1–7, New Delhi, India, December 2008.

[15] K. Kim, K. Cho, and S. Bahk, "Neighbor-aware adaptive load balancing algorithm for dense wireless sensor networks," in *Proceedings of the IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS '09)*, 2009.

[16] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics*, vol. 32, no. 1, pp. 1–12, 1990.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 7142739, 9 pages
http://dx.doi.org/10.1155/2016/7142739

Research Article

Design and Simulation Analysis for Integrated Vehicle Chassis-Network Control System Based on CAN Network

Wei Yu and Ning Sun

College of Automobile and Transport Engineering, Nanjing Forestry University, Nanjing 210037, China

Correspondence should be addressed to Wei Yu; yuwei505@163.com

Received 30 March 2016; Revised 18 May 2016; Accepted 26 June 2016

Academic Editor: Antonio Fernández-Caballero

Due to the different functions of the system used in the vehicle chassis control, the hierarchical control strategy also leads to many kinds of the network topology structure. According to the hierarchical control principle, this research puts forward the integrated control strategy of the chassis based on supervision mechanism. The purpose is to consider how the integrated control architecture affects the control performance of the system after the intervention of CAN network. Based on the principle of hierarchical control and fuzzy control, a fuzzy controller is designed, which is used to monitor and coordinate the ESP, AFS, and ARS. And the IVC system is constructed with the upper supervisory controller and three subcontrol systems on the Simulink platform. The network topology structure of IVC is proposed, and the IVC communication matrix based on CAN network communication is designed. With the common sensors and the subcontrollers as the CAN network independent nodes, the network induced delay and packet loss rate on the system control performance are studied by simulation. The results show that the simulation method can be used for designing the communication network of the vehicle.

1. Introduction

From the current development of the control system of vehicle chassis, integration and networking trend is very obvious [1]. The architecture of system control and network has different degrees of influence on the stability of chassis control. Due to the different functions of the system used in the vehicle chassis control, the hierarchical control strategy also leads to many kinds of the network topology structure and the distribution of the system computing tasks. In the 80s of last century, the researchers began to decompose the complex chassis control problem into a number of subcontrol systems and then use a mechanism to coordinate the dynamic relationship between the subsystems to meet the control requirements. Therefore, the research and discussion of the integrated control architecture of the chassis form [2–9] began to become the focus.

As far as the integrated control strategy of vehicle chassis is concerned, numerous studies have shown that the hierarchical control can effectively reduce the operation conflict between different functional subsystems,and quickly and effectively make the vehicle get the best performance. A large number of literatures [2–4] divide chassis control into different subcontrol systems according to the vertical, lateral, and normal control systems, and the integrated optimization control of the chassis is realized through the hierarchical control strategy. Li et al. put forward the integrated control structure of chassis based on the combination of the main loop and servo loop and discussed the problems of different directional force and force distribution of the chassis [6].

Chang and Gordon divided the chassis control system into three layers to achieve the active collision avoidance control [8]. Using the system architecture for the independent control units of the chassis of integrated control with upper coordinated control [10] can effectively adjust the collaborative work of control units, avoid the conflict of the controllers, and make the vehicle obtain the best running state. Through the analysis of the complex working conditions, the supervision mechanism is used to coordinate the multiple control

systems of the vehicle chassis, which can achieve a very good control effect of the system integration [11].

For these reasons, this paper firstly according to the hierarchical control principle, puts forward integrated control strategy of the chassis based on supervision mechanism. Based on the verification of the validity of this control strategy, the purpose of the study is to consider how the integrated control architecture affects the control performance of the system after the intervention of CAN network. Some exploratory simulation research is carried out. In order to facilitate the discussion, the integrated control system of network of the vehicle chassis based on network communication is abbreviated as IVC-NCS, namely, Integrated Vehicle Chassis-Network Control System.

2. Dynamic Model of the Vehicle

At present, international vehicle coordinate system mainly has two kinds [12]: one is SAE vehicle coordinate system issued by American Society of Automotive Engineers and another one is ISO vehicle coordinate system issued by International Standardization Organization. In this paper, SAE vehicle coordinate system is used for modeling, calculation, and analysis of vehicle dynamics. Based on the above assumptions, the nonlinear vehicle dynamics model has eight degrees of freedom.

There are a lot of tire models to calculate the complex nonlinear force between the road surface and the wheel. The most commonly used in the project is magic formula raised by Pacejka of Holland [13, 14] and unified tire model of overall conditions raised by Guo Konhui of China [15]. This paper uses Dugoff tire model [16], which is often used in computer simulation. It belongs to analytical model, and the parameters are small and easy to obtain.

3. Architecture of IVC-NCS Based on Supervision Mechanism

Figure 1 shows the architecture of IVC-NCS based on supervision mechanism. Three subcontrol systems are ESP, AFS, and ARS. Each subsystem can be controlled according to the calculation of local state variables. Based on the global state of the vehicle, the upper supervision controller judges the function weight by the vehicle stability for each subcontrol systems. The implementation of the execution mechanism is determined by the calculation results of each subcontroller and control weight.

3.1. ESP Subcontrol System. ESP takes the handling stability as the control target on the critical conditions of the whole vehicle. By controlling the braking intensity of four wheels, the electronic control of vehicle active safety is finished. The yaw rate tracking is the control target by applying a braking force at the right wheel to correct the unstable state of the vehicle. The system adopts sliding mode control strategy, and the tracking error of yaw rate is defined as sliding mode variable:

$$s_b = r - r_{\text{idl}}, \tag{1}$$

where r is the actual yaw rate and r_{idl} is the ideal yaw rate.

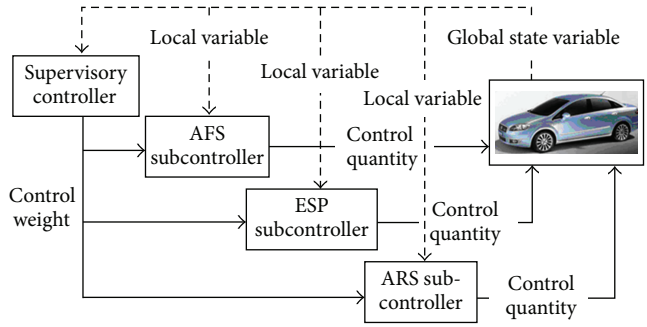


Figure 1: Architecture of IVC-NCS based on supervision mechanism.

The condition for reaching the sliding surface is defined as

$$\dot{s}_b = -\lambda_b s_b - \kappa_b \operatorname{sat}\left(\frac{s_b}{\varepsilon_b}\right), \tag{2}$$

where λ_b and κ_b are all positive constants, λ_b reflects the response speed of yaw tracking controller, κ_b shows the convergence rate of sliding mode surface of the system, S_b is the tracking error of yaw rate, and ε_b is the thickness of boundary layer.

The sliding mode controller satisfies the stability condition of Lyapunov sense.

Ignoring the inclination of vehicle and considering formula (2), when ESP control system acts on braking of single vehicle, the calculation formula of additional yaw rate torque is gotten:

$$\frac{M_{zc}}{I_{zz}} = -\frac{L_f\left(F_{y\text{fl}} + F_{y\text{fr}}\right) - L_r\left(F_{y\text{rl}} + F_{y\text{rr}}\right)}{I_{zz}} + \dot{r}_{\text{idl}} - \lambda_b s_b - \kappa_b \operatorname{sat}\left(\frac{s_b}{\varepsilon_b}\right), \tag{3}$$

where M_{zc} is the additional yaw torque generated by longitudinal driving force or braking force, I_{zz} is the moment of inertia of the vehicle body around z-axis, L_f is the vertical distance from the centroid to the front axle, L_r is the vertical distance from the centroid to the rear axle, $F_{y\text{fl}}$ is the longitudinal force of the ground on the left front tire, $F_{y\text{fr}}$ is the longitudinal force of the ground on the right front tire, $F_{y\text{rl}}$ is the longitudinal force of the ground on the left rear tire, $F_{y\text{rr}}$ is the longitudinal force of the ground on the right rear tire, r_{idl} is the ideal yaw rate, λ_b reflects the response speed of yaw tracking controller, κ_b shows the convergence rate of sliding mode surface of the system, S_b is the tracking error of yaw rate, and ε_b is the thickness of boundary layer.

In order to improve the unstable state in extreme conditions, braking force is applied to inward rear wheel when the vehicle has the understeer, or braking force is applied to outward front wheel when the vehicle has the oversteer. It can quickly and effectively improve vehicle stability. Therefore, the additional yaw torque calculated by formula (3) is converted to the equivalent braking force that can be applied to wheel with the most effective braking force.

3.2. AFS Subcontrol System. In steering system of the vehicle chassis, a relatively independent subcontrol system such as AFS is increased to adjust the front wheel angle for obtaining the optimum performance of IVC-NCS.

The system adopts sliding mode control strategy, and the tracking error of yaw rate is defined as sliding mode variable:

$$s_{sf} = r - r_{\text{idl}}, \tag{4}$$

where r is the actual yaw rate and r_{idl} is the ideal yaw rate.

The condition for reaching the sliding surface is defined as

$$\dot{s}_{sf} = -\lambda_{sf} s_{sf} - \kappa_{sf} \operatorname{sat}\left(\frac{s_{sf}}{\varepsilon_{sf}}\right), \tag{5}$$

where

$$\operatorname{sat}\left(\frac{s_{sf}}{\varepsilon_{sf}}\right) = \begin{cases} \dfrac{s_{sf}}{\varepsilon_{sf}}, & \left|s_{sf}\right| < \varepsilon_{sf} \\ \operatorname{sign}\left(\dfrac{s_{sf}}{\varepsilon_{sf}}\right), & \left|s_{sf}\right| \geq \varepsilon_{sf}, \end{cases} \tag{6}$$

where λ_{sf} and κ_{sf} are all positive constants, λ_{sf} reflects the response speed of yaw tracking controller, κ_{sf} shows the convergence rate of sliding mode surface of the system, s_{sf} is the tracking error of yaw rate, and ε_{sf} is the thickness of boundary layer.

The control law of steering angle about front wheel is

$$\delta_f = \frac{1}{b_{21}}\left[-a_{21}V_y - a_{22}r + \dot{r}_{\text{idl}} - \lambda_f\left(r - r_{\text{idl}}\right)\right] - \kappa_f \operatorname{sat}\left(\frac{s_{sf}}{\varepsilon_{sf}}\right), \tag{7}$$

where V_y is the lateral vehicle speed, r is the actual yaw rate, r_{idl} is the ideal yaw rate, λ_f reflects the response speed of yaw tracking controller, κ_f shows the convergence rate of sliding mode surface of the system, s_{sf} is the tracking error of yaw rate, and ε_{sf} is the thickness of boundary layer.

According to the vehicle model of two degree of freedom, $b_{21} = 2L_fC_f/I_{zz}$, where L_f is the vertical distance from the centroid to the front axle, C_f is the pitch damping, I_{zz} is the moment of inertia of the vehicle body around Z axis.

And $a_{21} = (2L_rC_r - 2L_fC_f)/I_{zz}V_x$, where L_r is the vertical distance from the centroid to the rear axle, C_r is the caster damping, L_f is the vertical distance from the centroid to the front axle, I_{zz} is the moment of inertia of the vehicle body around z-axis, V_x is the longitudinal speed.

And $a_{22} = -(2L_f^2C_f + 2L_r^2C_r)/I_{zz}V_x$, where L_f is the vertical distance from the centroid to the front axle, C_f is the pitch damping, L_r is the vertical distance from the centroid to the rear axle, C_r is the caster damping, I_{zz} is the moment of inertia of the vehicle body around z-axis, and V_x is the longitudinal speed.

3.3. ARS Subcontrol System. Active four-wheel steering technology can improve the handling stability of the vehicle at high speed and the controlling flexibility at low speed. The ideal yaw rate calculated by vehicle model of two degrees of freedom is the tracked target. So ARS takes round steering angle as the controlled variable.

The system adopts sliding mode control strategy, and the tracking error of yaw rate is defined as sliding mode variable:

$$s_{sr} = r - r_{\text{idl}}, \tag{8}$$

where r is the actual yaw rate and r_{idl} is the ideal yaw rate.

The condition for reaching the sliding surface is defined as

$$\dot{s}_{sr} = -\lambda_{sr} s_{sr} - \kappa_{sr} \operatorname{sat}\left(\frac{s_{sr}}{\varepsilon_{sr}}\right), \tag{9}$$

where

$$\operatorname{sat}\left(\frac{s_{sr}}{\varepsilon_{sr}}\right) = \begin{cases} \dfrac{s_{sr}}{\varepsilon_{sr}}, & \left|s_{sr}\right| < \varepsilon_{sr} \\ \operatorname{sign}\left(\dfrac{s_{sr}}{\varepsilon_{sr}}\right), & \left|s_{sr}\right| \geq \varepsilon_{sr}, \end{cases} \tag{10}$$

where λ_{sr} and κ_{sr} are all positive constants, λ_{sr} reflects the response speed of yaw tracking controller, κ_{sr} shows the convergence rate of sliding mode surface of the system, s_{sr} is the tracking error of yaw rate, and ε_{sr} is the thickness of boundary layer.

In order to restrain the shake of high frequency caused by frequent switching on the sliding surface, ε_{sr} is taken as the thickness of the boundary layer. λ_{sr} reflects the response speed of yaw tracking controller, and κ_{sr} reflects the rate how the system reaches the sliding surface.

3.4. Upper Supervisory Controller Design. The control idea of the supervisory controller is as follows: judging the steady state of the vehicle according to the stability factor, distributing the weight of the control function of three subcontrollers, and coordinating the output of each subcontroller.

Firstly, the stability factor of front and rear wheels is defined as [17]

$$\mathrm{SF}_f = \left|c_1\alpha_f + c_2\dot{\alpha}_f\right|, \tag{11}$$

where SF_f is the possibility that the front wheels come into the slipping state and α_f is the corresponding sideslip angle of the middle of the left and right wheels on the front axle:

$$\mathrm{SF}_r = \left|c_1\alpha_r + c_2\dot{\alpha}_r\right|, \tag{12}$$

where SF_r is the possibility that the rear wheels come into the slipping state and α_r is the corresponding sideslip angle of the middle of the left and right wheels on the rear axle.

c_1 and c_2 can be obtained by analyzing the relationship between the phase plane and the steering stability of the tire [18].

SF_f and SF_r show the possibility of the corresponding wheel beginning to side. The larger the value, the bigger the side slipping possibility of corresponding wheel, namely, the smaller the control margin provided by the wheel. Conversely, the smaller the value, the greater the effective strength of corresponding wheel.

Table 1: Rules of fuzzy controller of IVC.

SF_f	SF_r	W_{AFS}	W_{ARS}	W_{ESP}
S	S	B	B	S
S	MS	B	M	S
S	MB	B	S	S
S	B	B	S	MS
MS	S	M	B	S
MS	MS	M	M	MS
MS	MB	M	S	MS
MS	B	M	S	MB
MB	S	M	B	S
MB	MS	M	M	S
MB	MB	M	M	MS
MB	B	S	S	MB
B	S	S	B	MS
B	MS	S	M	MS
B	MB	S	S	MB
B	B	S	S	B

Through repeated simulation tests, when SF_f and SF_r are less than 0.7, the active steering control of the front and rear wheels can meet the requirements of vehicle stability. When SF_f or SF_r is bigger than 1.3, the use of ESP can be more effective to correct the excessive or lack steering state, which can keep the vehicle stable fast. When SF_f and SF_r are in the range from 0.7 to 1.3, the wheels with smaller stability factor provide a greater role in vehicle stability control. Based on this, the design of fuzzy logic controller is designed as follows.

The controller takes the stability factors of the front and rare wheels, such as SF_f and SF_r, as the input. The membership functions are in the same range $[0, 2]$, and the fuzzy subset is {S, MS, MB, B} as shown in Figure 2(a). The outputs of the controller are the control weights of three subcontrollers whose range is $[0, 1]$.

The membership functions of AFS and ARS are the same, and fuzzy subset is {D, M, E} as shown in Figure 2(b). The membership function of ESP subcontroller is shown in Figure 2(c), and fuzzy subset is {S, MS, MB, B}. The collection of letters is as follows: S is small, M is medium, and B is big.

Considering the actual application of the computation and real-time, all variables of the membership functions are easy to be calculated by the procedure, such as trigonometric function or trapezoidal function. Table 1 shows the inference rules of fuzzy controller of IVC.

4. Network Topology Design of IVC-NCS

According to system control strategy of IVC, combined with the control requirements of vehicle stability, the following several points are considered as the basis for the design. Actual limitations of vehicle space layout are as follows: because CAN network agreement and the corresponding international standards limit the length of the branches connecting the nodes and communication trunks, so network nodes in the actual space layout is one of the major considerations of network topology structure. Such as ARSC and AFSC, they are divided into two control units to control the system separately, which is helpful to connect the sensors and the executing agency.

Load capacity constraint of network communication is as follows: for IVC-NCS, if all sensors, controllers, and actuators exist as independent network nodes and the network works in 250 Kbps rate of regulated by vehicle high speed network of SAE, only from the theoretical calculation of CAN communication capability, its load capacity is difficult to meet the control requirements. While the communication speed is increased to 500 Kbps, the anti-interference ability of the node will be poor, so it is difficult to realize the high speed communication in the bad electromagnetic environment.

Real-time requirements of subsystems are as follows: three subsystems of IVC-NCS are the relatively independent closed-loop control system. ESP subsystem has higher request on real-time of wheel speed signals, which requires the executing agencies to react quickly according to control orders.

The sensors necessary for many systems are designed as independent network nodes. The subcontrol systems adopt traditional point-to-point connection in the controllers, sensors, and executing agencies. Its object is to obtain satisfactory real-time performance and reliability.

Based on above analysis, the network in Figure 3 is designed as IVC-NCS structure. CAN network is taken as the communication medium of the controller node, and each subsystem is connected with the traditional method of point to point. Considering that ESP system has obvious effect for vehicle stability in extreme conditions, the supervision and control tasks of the system and the control calculation of ESP are assigned to one node.

The sensor signals are the basis of the controller to judge the state of the vehicle and control instructions. When the network communication load suddenly increases, the probability of signal loss of low level sensors will be significantly increased. Therefore, in order to ensure the real-time performance of the sensor signal transmission, the message priority of the sensor nodes is set higher to avoid the message loss in the control cycle, which leads to control instability. Table 2 shows the communication matrix table of IVC-NCS. Messages Msg7 and Msg9, as the state messages of executing agencies, can help the controller nodes to understand the operation status of the system. Because they do not participate in the control calculation, so the priority is low, and the transmission cycle is relatively large.

5. Simulation and Result Analysis

According to nonlinear vehicle model with eight degrees of freedom to calculate the state of the vehicle, Simulink platform is used for simulation. Before the performance of IVC-NCS, the IVC system is simulated and tested to verify the effectiveness of the controller.

5.1. Effectiveness Verification of IVC System Control. In order to verify the effectiveness of IVC system, the sine curve and the step curve with the maximum value 5 degrees (about 0.087 rad) of the vehicle steering wheel are input to simulate

Table 2: Communication matrix table of IVC-NCS.

Message name	Message content (signal)	Transmission node	Message property			Signal description
			Priority	Cycle	Data domain	
Msg1	Interf	Node 1	1	Pending	8 bytes	Meaningless message
Msg2	S_Yaw_Acc	Node 2	2	$P = 5$ ms	6 bytes	Yaw rate and lateral acceleration
Msg3	S_StrWhAgl	Node 3	3	$P = 5$ ms	4 bytes	Steering wheel angle
Msg4	S_Vx	Node 4	4	$P = 5$ ms	4 bytes	Longitudinal speed
Msg5	Weight	Node 4	7	AP	4 bytes	Control weight of AFS and ARS
Msg6	S_ARS	Node 5	5	$P = 5$ ms	8 bytes	Speed and rotation angle of rear wheel
Msg7	D_ARS_M	Node 5	8	$P = 20$ ms	4 bytes	Rear wheel motor status
Msg8	S_AFS	Node 6	6	$P = 5$ ms	8 bytes	Speed and rotation angle of front wheel
Msg9	D_AFS_M	Node 6	9	$P = 20$ ms	4 bytes	Front wheel motor status

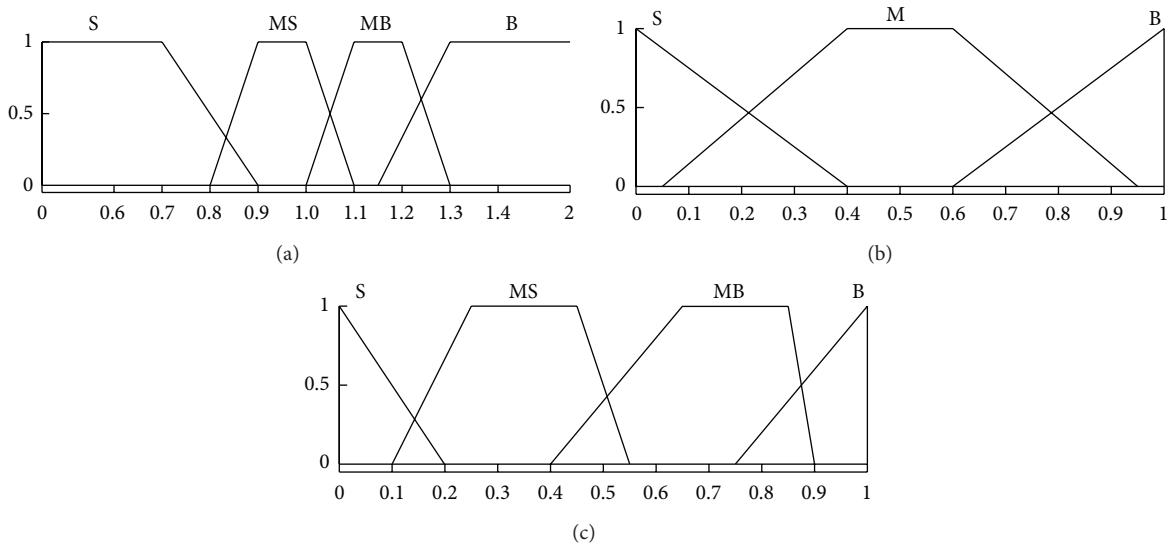


Figure 2: (a) SF_f and SF_r. (b) Control weight W_{AFS}, W_{ARS} of AFS and ARS. (c) Control weight W_{ESP} of ESP.

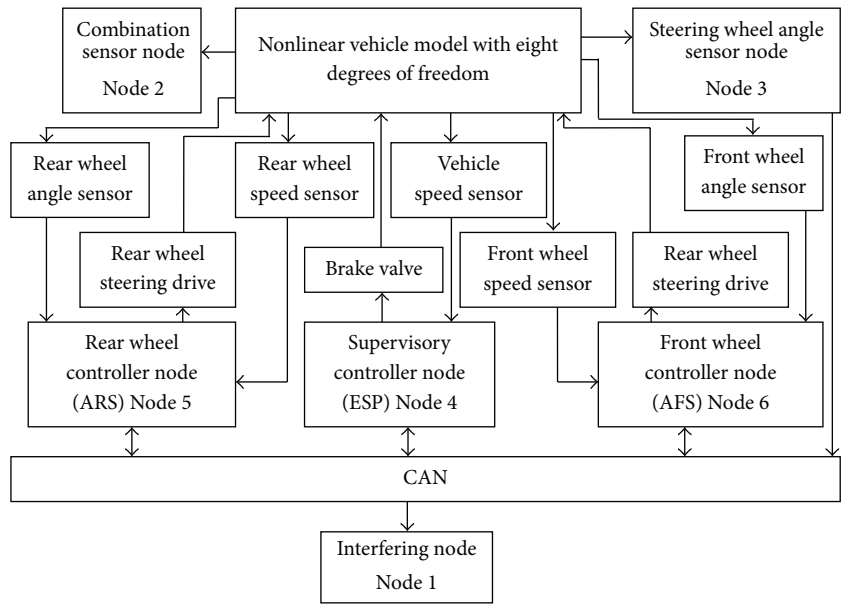


Figure 3: IVC-NCS network structure.

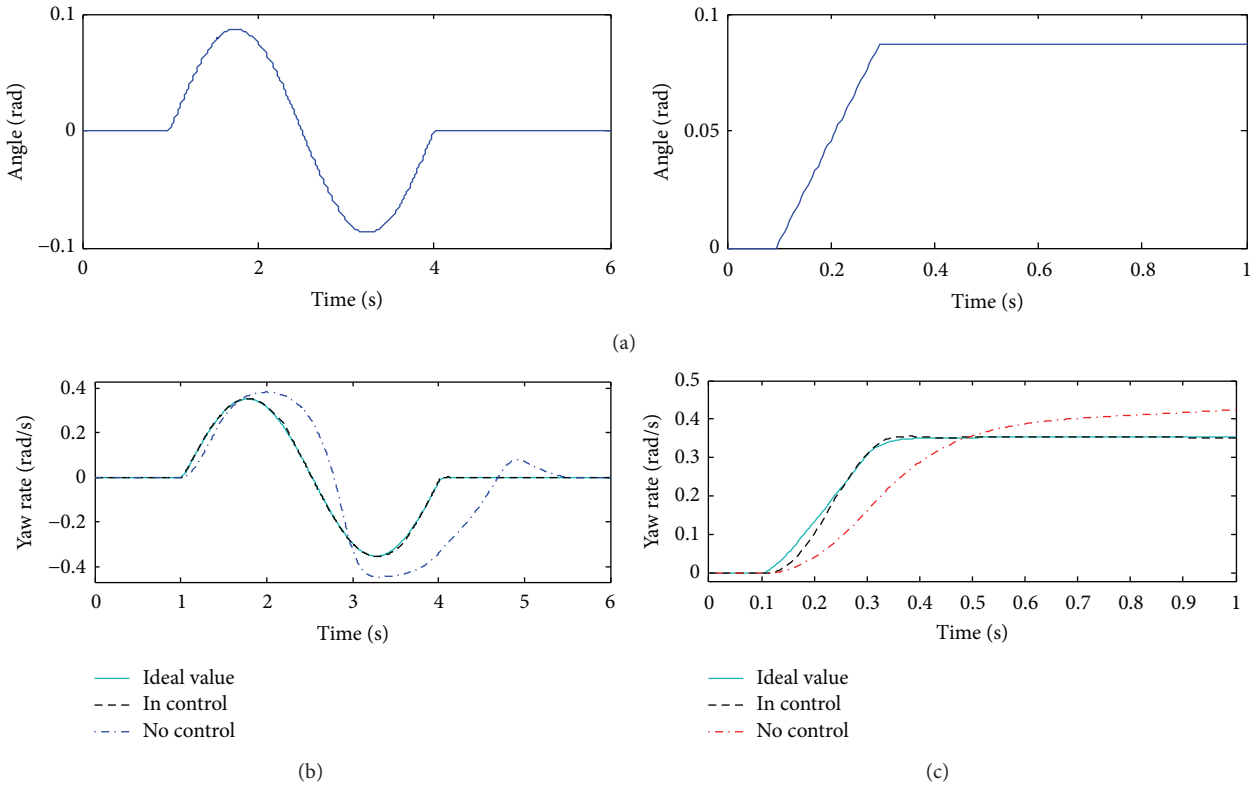


FIGURE 4: (a) Input curve of front wheel angle. (b) The response curve of yaw rate of the steering wheel with sine angle input. (c) The response curve of yaw rate of the steering wheel with step angle input.

the tracking response of the vehicle under different input yaw rates. According to the transmission ratio of the steering system, the corresponding input curve of front wheel steering angle is shown in Figure 4(a). The vehicle travels at a good road with a adhesion coefficient of 0.85, and the initial speed is 25 m/s.

Figures 4(b) and 4(c) are the response curves of vehicle yaw rate at different angle inputs. It can be seen that the yaw rate of the controlled vehicle can quickly and effectively track the ideal value when compared with the system without the control. For the sine input, the execution of the vehicle is a nonstandard single lane change test. At this time due to the correction function of angle changes of the front wheel, so after the apparent slip, the yaw rate is settled in zero value, as shown in Figure 4(b).

Under the step input of steering wheel angle in Figure 4(c), the yaw rate of the vehicle without control cannot track the ideal value, which appears as the trend of divergence. So the vehicle cannot achieve stable circular motion and rollover because of instability. The yaw rate of the vehicle with controllers is good at tracking the ideal value. Simulation results show that the IVC system can effectively improve the stability vehicle in critical conditions, which verifies the effectiveness of the designed control system.

5.2. Simulation Analysis of IVC-NCS Based on CAN. In order to investigate the performance change of the designed IVC system after the CAN network is involved in the control, the stability of the vehicle was investigated using the same step input of the steering wheel. The initial speed is 25 m/s, and the road adhesion coefficient is 0.85. Considering the practical application of CAN network with high speed, the communication rate is set to 250 Kbps. Node sends only data frames. If the interfering nodes do not send any message, the network load is about 84% when the maximum is filled. When the interference nodes send the interference message of high priority with 4 ms cycle, it can ensure that the network load is close to 1 but less than the network bandwidth, which ensures the system communication not to lose the frames.

According to the assumptions and simulation conditions, Figure 5 shows the comparison curve of yaw rate tracking about CAN network communication and point-to-point connection. Compared with point-to-point connection mode, the IVC system with CAN network connection can quickly and effectively track the ideal value under the condition of good network environment without changing the steady state of the control system. It can be clearly seen that, in the part of the amplified image, the network involves in the control system, which makes the yaw rate fluctuate with microamplitude. The overshoot of control increases from 3.1% of the point-to-point connection to 6% of the CAN network connection.

In order to investigate the influence of different network state on the control performance of the system, the tracking simulation test of vehicle yaw rate is carried out for different network load and packet loss rate.

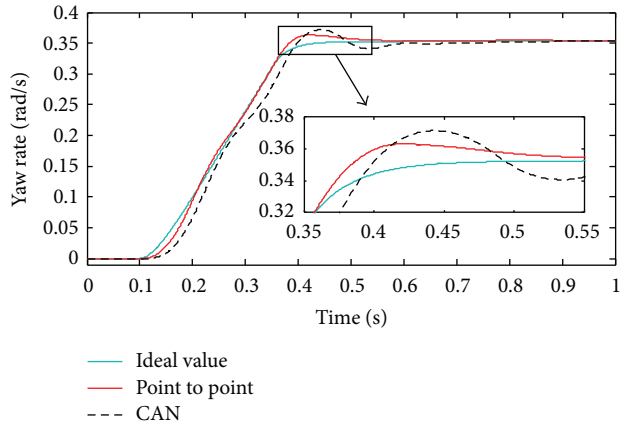


Figure 5: The response curve of IVC yaw rate of CAN network connection.

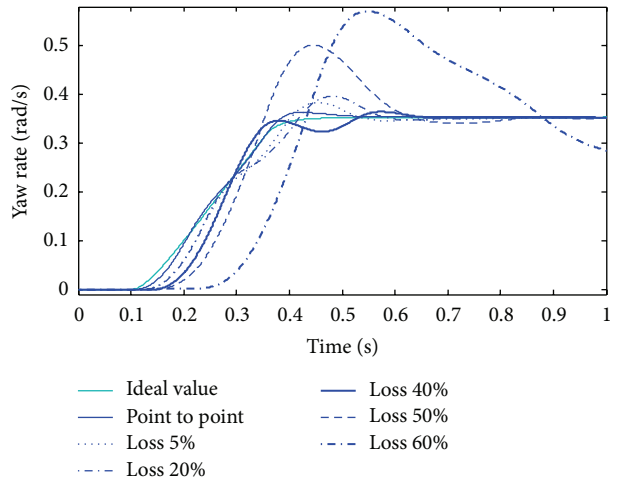


Figure 6: The response curve of different packet loss rates.

Figure 6 shows the response curve of different packet loss rates of IVC-NCS yaw rate. In the simulation process, the interference nodes do not send the messages. It can be seen that when the packet loss rate is lower than 20%, the dynamic characteristic of the system becomes bad. In the packet loss rate of 5% and 20%, the corresponding overshoots of the system are about 9% and 12.5%. In 0.3 s after the step input of front wheel ends, the vehicle yaw rate can be stable to track the ideal value. When the packet loss rate is less than 40%, the yaw rate of the vehicle can be finally stabilized at an ideal value. When the packet loss rate is more than 40%, the yaw rate is obviously fluctuated in the ideal yaw rate tracking process. At 50%, the overshoot of yaw rate increases rapidly to about 42%, the vehicle begins to sideslip.

When the packet loss rate is up to 60%, the vehicle yaw rate tracking is seriously lagging behind, which cannot achieve stable circular motion. The analysis shows that when the packet loss rate is low, the message transmission keep high success rate. The information of the sensors can be obtained by control nodes in time, so the controller works fast with little effect on the performance of system control. With the increase of packet loss rate, the control instructions cannot be timely generated and executed, which makes the control cycle become longer. The status of executing agency cannot be corrected in time. The input of executing agency will be too large or too small, which causes the control to fail.

Figure 7 shows that the interference nodes send the messages of highest priority in 4 ms cycle, and the network load is close to 1. The long dashes are the response curve of yaw rate of CAN network without the interference, when the network load is about 84%. The short dashed lines, dashed-dotted lines, and bold dashed lines are separately response curves of yaw rate at t different packet loss rates when the load is full.

Under the condition to meet the communication requirements of the control system, when the network load is close to 1, the induced delay of the system is largest. It can be calculated, when the network load increases from 84% to nearly 100% and the overshoot increases from 6% to 7%. When the network load is 1 and packet loss rate is 30%,

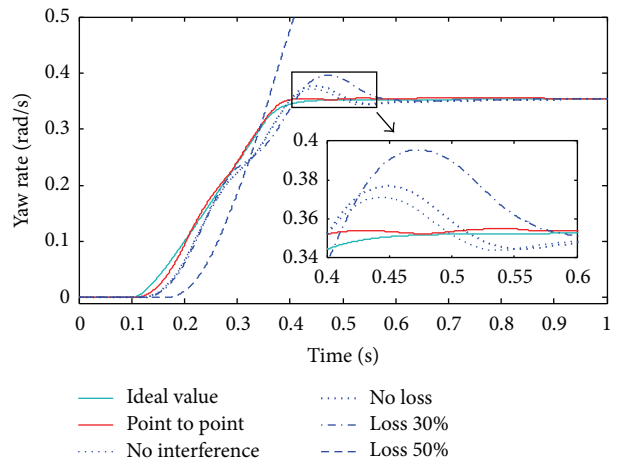


Figure 7: The response curve of yaw rate with full load of network communication.

the overshoot of yaw rate is 15.7%. Therefore, although the network load increases, as long as network load can meet the communication requirements of the control system, the network intervention only has little effect on the quality of dynamic control, which does not change the steady characteristics of the system. The vehicle can achieve the stable circular motion within 0.3 s of the yaw rare input of the front wheel.

When the communication network is fully loaded and the packet loss rate is 50%, the vehicle cannot complete the scheduled circular motion. The yaw rate of the vehicle diverges to make the vehicle out of control. The simulation results show that when the network bandwidth meets the needs of control system, the effect of the network induced delay of control system is very small and negligible. And the network packet loss will affect the performance of control system seriously. When the packet loss rate is up to 50%, the system control performance will deteriorate significantly.

5.3. Stability and Coordination Analysis. From the development of the vehicle chassis control system, the trend of integration and network is very obvious. The system control architecture and the network architecture form have different effects on the stability control of the chassis. In this paper, the design of the control system fully takes into account the stability of the chassis control performance.

Because ABS is the basis for the realization of ESP, and the latter needs to achieve the independent control of braking intensity about the four wheels, so ABS is designed as an independent four-channel mode. As one kind of the controller associated with safety and real-time, the execution and controller of ABS usually adopt directly connected manner, in order to reduce the information switching delay and ensure the safety and stability of the vehicle.

The control target of ESP system is to control the stability of the vehicle in the extreme conditions, through the control of braking strength of four wheels to achieve the active safety. In order to improve the unstable state of the vehicle in extreme conditions, applying the braking force on inward rear wheel with the understeer or on the outward front wheel with the oversteer can quickly and effectively improve the stability. Taking into account that ESP system has the obvious effect on the vehicle stability in extreme conditions, the study will assign the supervision and control tasks and the calculation of ESP control to one node.

For the performance of network control system, communication real-time performance is the most important factor affecting the control performance, which can be expressed and measured by network delay. The existence of network delay reduces the control performance of the system, which will lead to the loss of stability of the stable control system.

Especially, in extreme conditions, the change of the vehicle state is larger. When a large number of control instructions are lost, the adjustment of the new and old control instructions is bound to increase because of the large number of cycles, which will increase the action range of the actuator. Therefore, too much data packet loss is extremely unfavorable for the stability control. When the packet loss rate is less than a certain value, only the system dynamic characteristic becomes worse, and the system stability is not changed. When the packet loss rate reaches the critical value, the system control stability is close to the critical state.

In addition, through the simulation experiment, we can know that CAN network intervention did not significantly affect the stability of vehicle braking. Therefore, when CAN network communication environment is good, the network induced delay of CAN network has a little influence on the performance of the controller, which indicates that the ABS controller built in this research has strong robustness on a single road.

In the 80s of last century, the researchers began to try to decompose the complex chassis control problem into a number of subcontrol systems and then use a mechanism to coordinate the dynamic relationship between the subsystems to meet the control requirements. Using the upper coordinated control for the integrated control architecture of multiple independent control units of the vehicle chassis can effectively adjust the collaborative work of the control units, avoid the conflict between the controllers, and make the vehicle obtain optimal running state.

The supervision mechanism is based on a hierarchical control principle, combined with fuzzy control logic to design a controller to supervise and coordinate ESP, AFS, and ARS. The target of the upper supervisory controller, according to the stability factor to judge vehicle steady state, is to redistribute the control weights of three subsystems and coordinate the output of each subcontroller.

The sensors necessary for many systems are designed as independent network nodes. The subcontrol systems adopt traditional point-to-point connection in the controllers, sensors, and executing agencies. Its object is to obtain satisfactory real-time performance and coordination.

6. Conclusions

In this paper, the vehicle chassis control system is taken as the application of CAN network. The target focuses on how the network affects the control system. The ABS, ASC, and IVC are simulated. The main research contents and conclusions are as follows.

According to the control theory of sliding mode, ESP and AFS subcontrollers are designed to track the ideal yaw rate. Based on the principle of hierarchical control and fuzzy control, a fuzzy controller is designed, which is used to monitor and coordinate the ESP, AFS, and ARS. And the IVC system is constructed with the upper supervisory controller and three subcontrol systems on the Simulink platform. Compared with the point-to-point connection, the system simulation of IVC-NCS shows that the control of the integrated control system has good performance.

According to the IVC based on the supervision mechanism, combined with the function of each subsystem, the network topology structure of IVC is proposed, and the IVC communication matrix based on CAN network communication is designed. With the common sensors and the subcontrollers as the CAN network independent nodes, the network induced delay and packet loss rate on the system control performance are studied by simulation. The simulation results show that the network does not lose frame, and even if the network traffic load is close to 1, the network intervention of IVC can only show the very small change of the dynamic quality of the system. The network packet loss has a significant impact on the performance of the system control. When the packet loss rate is less than 30%, only the system dynamic performance becomes worse, and the system stability does not change. When the packet loss rate is up to 50%, the system control stability is close to the critical state, and the vehicle is unstable.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This study was funded by The Natural Science Foundation of Jiangsu Province (BK20130977).

References

[1] T. Gordon, M. Howell, and F. Brandao, "Integrated control methodologies for road vehicles," *Vehicle System Dynamics*, vol. 40, no. 1–3, pp. 157–190, 2003.

[2] H. Chou and B. D'Andréa-Novel, "Global vehicle control using differential braking torques and active suspension forces," *Vehicle System Dynamics*, vol. 43, no. 4, pp. 261–284, 2005.

[3] C. B. Chu and W. W. Chen, "Vehicle chassis system based on layered coordinated control," *Chinese Journal of Mechanical Engineering*, vol. 44, no. 2, pp. 157–162, 2008.

[4] H. Zhu and W. W. Chen, "Active control of vehicle suspension and steering system based on strategy hierarchy," *Chinese Journal of Agricultural Machinery*, vol. 39, no. 10, pp. 1–6, 2008.

[5] M. J. L. Boada, B. L. Boada, A. Munoz, and V. Diaz, "Integrated control of front-wheel steering and front braking forces on the basis of fuzzy logic," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 220, no. 3, pp. 253–267, 2006.

[6] D. Li, X. Shen, and F. Yu, "Integrated vehicle chassis control with a main/servo-loop structure," *International Journal of Automotive Technology*, vol. 7, no. 7, pp. 803–812, 2006.

[7] E. J. Bedner and H. H. Chen, "A supervisory control to manage brakes and four-wheel-steer systems," SAE Paper 2004-01-1059, 2004.

[8] S. Chang and T. J. Gordon, "A flexible hierarchical model-based control methodology for vehicle active safety systems," *Vehicle System Dynamics*, vol. 46, supplement 1, pp. 63–75, 2008.

[9] N. Kelling, "The BRAKE project—centralized versus distributed redundancy for brake-by-wire systems," SAE Paper 2002-01-0266, SAE International, 2002.

[10] J. X. Wang, *Research of Integrated Control System of Vehicle Chassis Based on Multi Agent*, Southeast University, Nanjing, China, 2010.

[11] Y. Zhang, C. L. Yin, and J. W. Zhang, "A real time estimation method for the lateral velocity of the center of mass of the vehicle," *Chinese Journal of Mechanical Engineering*, vol. 44, no. 2, pp. 219–222, 2008.

[12] D. Crolla and Y. Fan, *Vehicle Dynamics and Control*, Chinese Communications Press, Beijing, China, 2003.

[13] H. B. Pacejka and E. Bakker, "Magic formula tyre model," *Vehicle System Dynamics*, vol. 21, no. 1, pp. 1–18, 1993.

[14] H. B. Pacejka and I. J. M. Besselink, "Magic formula tyre model with transient properties," *Vehicle System Dynamics*, vol. 27, supplement 1, pp. 234–249, 1997.

[15] K. H. Guo and L. Ren, "A unifield semi-empirical tire model with higher accuracy and less parameters," SAE Technical Paper Series 1999-01-0785, SAE International, 1999.

[16] H. Dugoff, P. S. Fancher, and L. Segal, "Tyre performance characteristics affecting vehicle response to steering and braking control inputs," Final Report, US National, 1969.

[17] W. Jinxiang and C. Nan, "Research on supervisory control based integrated chassis control framework and its simulation," *Transactions of the Chinese Society of Agricultural Machinery*, vol. 40, no. 9, pp. 1–6, 2009.

[18] S. Inagaki, I. Kushiro, and M. Yamamoto, "Analysis on vehicle stability in critical cornering using phase-plane method," in *Proceedings of the International Symposium on Advanced Vehicle Control (AVEC '94)*, pp. 287–292, Tsukuba- Shi, Japan, 1994.

Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 9507213, 15 pages
http://dx.doi.org/10.1155/2016/9507213

Research Article

A Novel Online Detection System for Wheelset Size in Railway Transportation

Xiaoqing Cheng,[1,2] Yuejian Chen,[1,3] Zongyi Xing,[4] Yifan Li,[3,5] and Yong Qin[1]

[1] *State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*
[2] *School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[3] *Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada T6G 2R3*
[4] *School of Automation, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China*
[5] *School of Mechanical Engineering, Southwest Jiaotong University, Chengdu, Sichuan 610031, China*

Correspondence should be addressed to Yuejian Chen; yuejian1@ualberta.ca

Received 10 February 2016; Revised 9 March 2016; Accepted 5 April 2016

Academic Editor: Rafael Morales

The online detection of wheelset size has important implications for ensuring the safety of railway operation and decreasing the maintenance costs. Based on laser displacement sensors (LDS), a novel online detection system of the wheel size is proposed using only six two-dimensional LDS and two one-dimensional LDS. The calculation principles of tread profile and wheel diameter are given, as well as the calibration method. Errors induced by wheel-rail vibration, misalignment, sensor noise, S-shape running, and wheelset differential are also analyzed. After system implementation, field experiments were performed using both standard wheel and several real trains. It turns out that the detection uncertainty of flange width and height is 0.1 mm and wheel diameter 0.3 mm, which can meet the requirements of maintenance.

1. Introduction

Wheel and rail interact with each other by designed profiles and geometric parameters. The wear of the profile significantly influences the dynamic performance of railway vehicles and even leads to derailment in a massive stage [1]. Therefore, at the very beginning of railway transportation, measuring and ensuring of wheel-rail interactions are a fundamental issue [2]. With the continuous increasing of axle load, train speed, and higher reliability requirement, this issue attracts more attention than ever.

According to the charge-coupled device (UIC) 510-2 code, the geometric parameters of wheelset that need to be measured consist of diameter and tread profile managed by flange width and flange height [3]. Varieties of measuring techniques have appeared, such as specially designed calipers, hand-on automatic scanner, and online detection system [4].

At the earlier stage, the caliper is an effective tool for measuring wheelset size because of the advantage of the simple operation. However, it has shortages of high labor intensity and fluctuated accuracy depends on the skillfulness of workers. Meanwhile, calipers, a contact measuring tool, will inevitably undermine the measuring wheel, causing certain damage.

After that, advanced artificial caliper has emerged with applying noncontact technology. One recognized tool is MiniProf Wheel System developed by Greenwood Engineering [5]. The MiniProf Wheel System is magnetically attached to the wheel. It provides the calculation of wear parameters and is also capable of measuring the flange and taper line diameter on wheels. Due to the benefits of utilizing computer analysis, overall efficiency is increased but this system still takes more than five minutes to measure a single car, not to mention the whole train. Medianu et al. [6] also developed a hand-on scanning system for tread profile. This system uses one-dimensional LDS (1D-LDS) driven by worn gear to scan tread profile. On the whole, those hand-on systems require the train to remain static or even dismantled, facing a great challenge of detection efficiency.

A widely distinguished technology is online detection system which has the advantages of noncontact, high efficiency, and high accuracy. The high efficiency is realized by dynamic measurement; namely, a train passes the measurement system at a certain speed. There are some commercial companies, such as MERMEC Group [7], IEM, Inc. [8], and KLD Labs, Inc. [9], selling wheelset detection systems on the market. These systems mainly utilize structured laser light and charge-coupled device (CCD) image processing technology. Chen et al. [10] and Gong et al. [11] proposed a structured laser light and CCD based online detection system for tread profile and diameter, respectively. In their research, two pairs of structured laser light and CCD are adopted to recover the inner and outer profiles of each wheel and register them by the iterative closest point algorithm. In diameter detection, the cycloid constraint is utilized to obtain a wide distribution of the contact points. Even though many possible factors that cause the error are considered, the system still lacks real data validation and statistical detection uncertainty analysis. Mian et al. [12] provided an optical evaluation method for railway wheelset with installing image cameras along at least one circumference of the wheel. Such a system can be of high price with so many optical sensors. Gao et al. [13] utilized a pair of line structured laser light and CCD to obtain multiple contact points via repeated shooting. This method needed to measure the speed of the train precisely, and the space interval of the points was decided by the speed and the time interval of shooting. Zhang et al. [14] used only one CCD camera to capture the image of the light profile of the wheelset, and, in the meantime, the tread profile is illuminated by a linear laser. Overall, the structure of structured laser light and CCD based system seems to be too complex. Such a structure also brings about difficulty in the calibration of those systems. The combination of structures of structured laser light and CCD is also sensitive to the harsh environment with vibration and light.

Apart from the structured laser light and CCD sensors, LDS can provide more satisfactory results. The LDS is a special kind of structured light-vision measurement sensor where the photoelectrical detector and laser light source are assembled, providing benefits like easy installation and no need to calibrate intrinsic parameters online. Russian scientists [15] reported their innovative laser sensor, claiming that 1D-LDS can be enough for measuring tread profile. Then, they [16] further derived a mathematical calculation regarding wheel diameter detection. However, the method needs a high precise train speed and time interval of shooting. Gao et al. [17] utilized one 1D-LDS and two eddy current sensors to detect the wheel diameter. Wu and Chen [18] used high-speed CCD and 1D-LDS to measure the diameter and the accuracy was within 1.2 mm. Zhang et al. [19] used two 1D-LDS and a position sensor to detect the wheel diameter and, meanwhile, wavelet analysis is used to eliminate the signal noise. Triangle geometry was the main computational algorithm in the LDS method. These systems do not need a precise train speed and time interval of shooting anymore. However, the 1D-LDS methods mentioned above needed the dot laser to be strictly projected at the contact point on the wheel tread. That is difficult to achieve because of the S-shape motion of the wheelset.

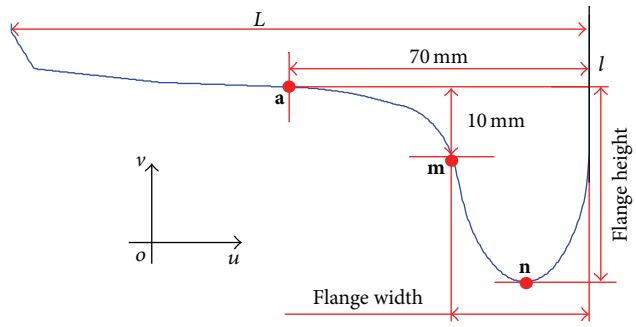


Figure 1: Schematic of measuring target.

Aiming at LDS based detection system, the authors previously proposed an online detection system using eight 2D-LDS to detect the wheel diameter and tread profile [20]. The system utilized a digital I/O card to generate digital synchronous signals which guarantee simultaneous working of all sensors. The 2D-LDS sensors are relatively of a high price and in the previous system some of the 2D-LDS are actually not frequently used. In this paper, the authors replaced several 2D-LDS sensors with 1D-LDS. Six 2D-LDS and two 1D-LDS are implemented in this new system. Working simultaneously, the data collected from all the sensors are processed, and then wheel diameter and tread profile are calculated. Errors induced by wheel-rail vibration, sensor noise, misalignment, S-shape running, and wheelset diameter differential are also analyzed. At last, after the system is implemented, the field test is carried out by standard wheelset test and real train test.

2. System Principle

2.1. Measuring Target. Figure 1 shows a typical wear tread profile [21]. We define the coordinate uov as the tread profile panel across the wheel center. The inner side of the wheel, as shown in black line l in Figure 1, has no wear-out and deformation when there is wheel-rail contact. L is the wheel hub thickness, namely, the distance from the inner side to the outer side of the wheel. The base point **a**, which is the center of the vertical wheel load and is considered to be the diameter point of the wheel, is −70 mm away from the line l along u-axis. The base point **m**, which is the center of the lateral wheel load, is −10 mm away from base point **a** along v-axis. The base point **n** is the vertex point of the wheel rim. The tread profile is somehow complex so that the condition of tread profile is usually evaluated by flange width and height. The flange width is defined by the width between base line l and base point **m** along u-axis and the flange height is defined by the distance between base point **a** and vertex point **n** along v-axis.

2.2. System Layout. The presented wheelset size online detection system depends on LDS sensors. Figure 2 shows the system layout. The system consists of six 2D-LDS and two 1D-LDS, each of which is installed below the track to measure

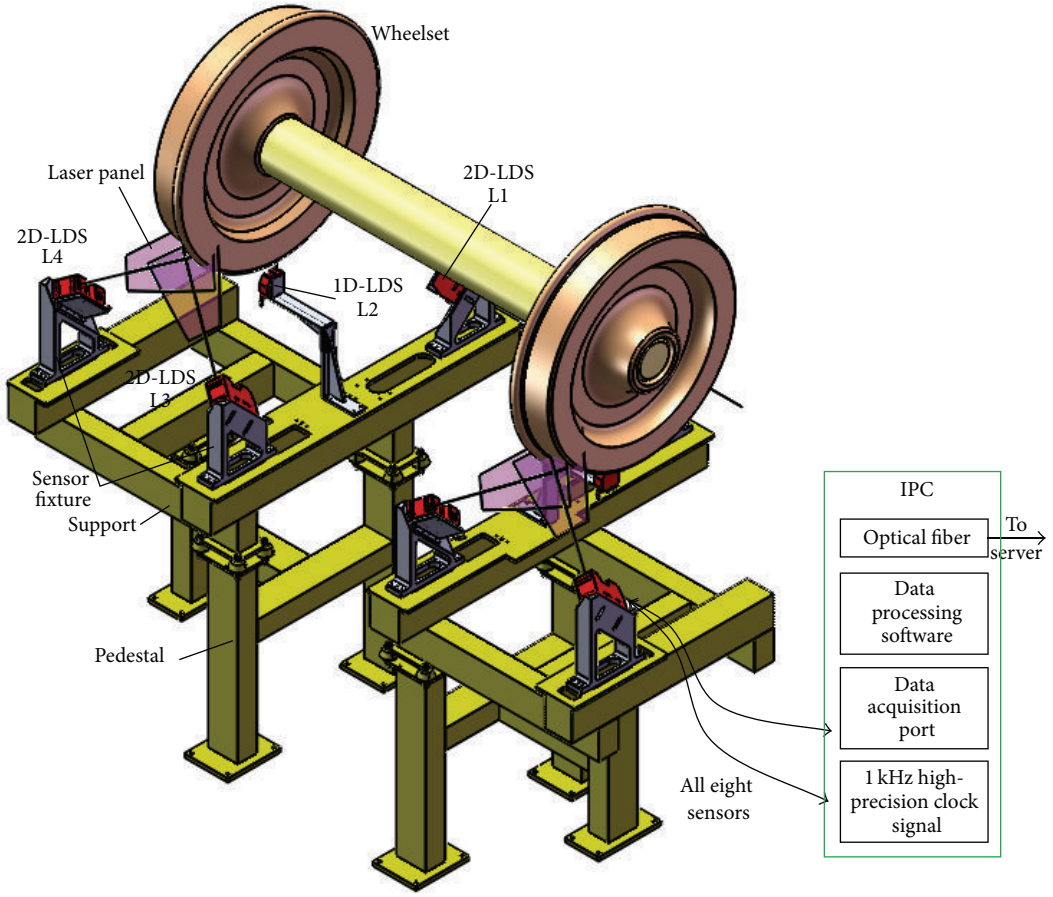


Figure 2: Schematic diagram of sensors installation.

both sides of the tread profile and diameter. They can be divided into two groups because the left side and right side are mirrored. Taking the left side LDS as an example, the 2D-LDS L1 and L3, together with 1D-LDS L2, measure the wheel diameter by three points' principle. The 2D-LDS L3 and L4 measure the tread profile.

Both the 2D and 1D laser sensors are based on laser triangulation measurement principle and are made up of laser diode and a CCD linear sensor element. The emitted laser forms a laser belt on the wheel tread and then the laser is reflected to the CCD linear inductive components. Inside the sensor, there is an integrated circuit unit to process the optical displacement data and to obtain the tread and flange profile coordinates. Based on the principle that the output points of the LDS are originated from the laser emitting source, in application, the laser emitting source should be regarded as the origin of the scanning coordinate. The signal from all the sensors is transmitted to IPC through data acquisition port. A digital I/O card is utilized to produce precisely a 1 kHz square signal in order to ensure all sensors to complete the task of acquiring the tread profile synchronously. The sensors begin to collect the data on the decline of the square wave signal and then transmit the data to the IPC through data acquisition port for the subsequent process. The sampled signal is analyzed in data processing software, and, finally, the condition of each wheel is decided. There is also an optical fiber in the IPC so that the condition of each wheel can be transmitted to distant depot office. All the sensors are fixed by special designed mechanical sensor fixture so that the sensors can be installed in certain space position. The fixtures are supported by the well manufactured mechanical structure. The whole system is finally connected with the ground by the pedestal.

In addition, the system also consists of several accessorial types of equipment that have not been shown in system layout figure, which are wheel position sensor and automatic train identification antenna. Three wheel position sensors are installed beside the outside of the rail. Along the rail, the first one is used to detect the arriving moment of the first wheel axis of a train and, hence, to trigger subsequent hardware facilities; the second one is used to trigger the scanning of all laser sensors; the last one is used to detect the leaving moment of the last wheel axis of a train and, hence, to close the subsequent hardware facilities.

2.3. Static Tread Profile Calculation Principle. Taking the left side LDS as an example, we define world coordinate frame (WCF) o-xyz and LDS scanning coordinates $o^{(1)}$-$x^{(1)}y^{(1)}z^{(1)}$, $o^{(2)}$-$x^{(2)}y^{(2)}z^{(2)}$, $o^{(3)}$-$x^{(3)}y^{(3)}z^{(3)}$, and $o^{(4)}$-$x^{(4)}y^{(4)}z^{(4)}$ for L1,

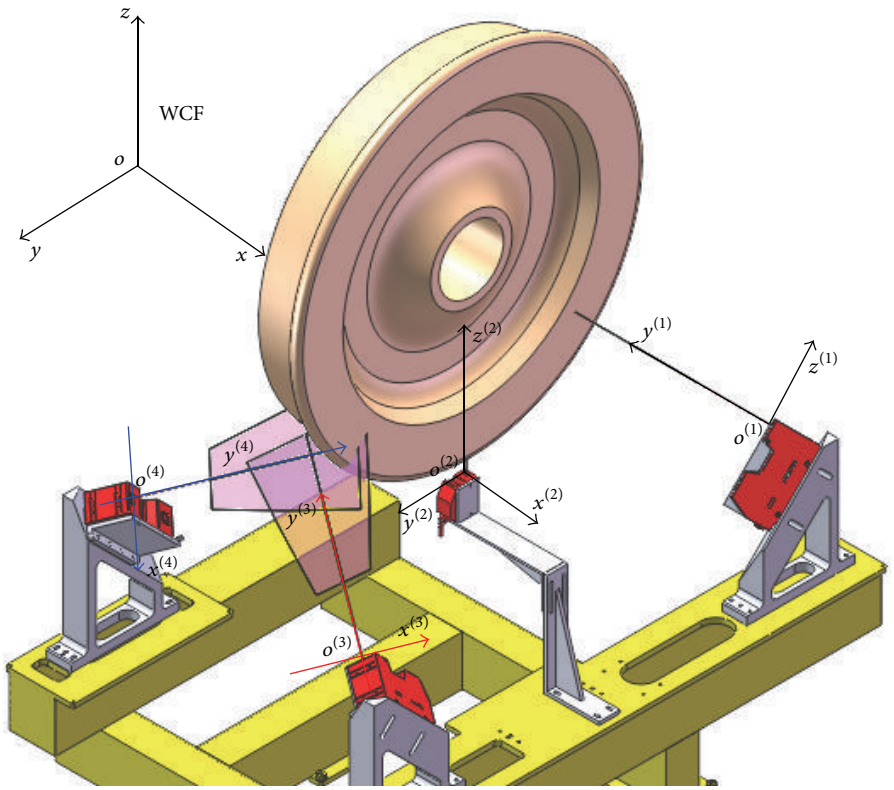


Figure 3: Coordinates set.

L2, L3, and L4, respectively. As shown in Figure 3, the scanning coordinates take the origin of laser light as the coordinate origin, the equal angle bisector of the triangle laser panel as y-axis, and the direction in triangle laser panel orthogonal with y-axis as x-axis and finally use right-hand rule to determine z-axis. The coordinates of the output points are in the scanning coordinate system of the sensor.

The wheelset is assumed to be in the right position and to remain static. The scanning coordinate of L3 and L4 is rotated by two angles from the WCF, namely, angles α and β with respect to x-axis and y-axis. The tread profile is measured by L3 and L4 and the scanning panels of L3 and L4 are the same. The tread profile can be measured in tread profile panel constituted by L3 and L4. So, as shown in Figure 4(a), only angle β is considered when measuring tread profile. Because of angle β, the output line is distorted and needs to be transformed into physical profile. According to the installation angles, β_3 and β_4, the output data is transformed by

$$\begin{aligned} u_n^{(3)} &= \sqrt{x_n^{(3)^2} + y_n^{(3)^2}} \sin\left(\theta + \beta_3\right) \\ &= x_n^{(3)} \cos\beta_3 + y_n^{(3)} \sin\beta_3, \\ v_n^{(3)} &= \sqrt{x_n^{(3)^2} + y_n^{(3)^2}} \cos\left(\theta + \beta_3\right) \\ &= y_n^{(3)} \cos\beta_3 - x_n^{(3)} \sin\beta_3, \end{aligned} \tag{1}$$

$$\begin{aligned} u_n^{(4)} &= \sqrt{x_n^{(4)^2} + y_n^{(4)^2}} \sin\left(\theta' - \beta_4\right) \\ &= x_n^{(4)} \cos\beta_2 - y_n^{(4)} \sin\beta_2, \\ v_n^{(4)} &= \sqrt{x_n^{(4)^2} + y_n^{(4)^2}} \cos\left(\theta' - \beta_4\right) \\ &= y_n^{(4)} \cos\beta_2 + x_n^{(4)} \sin\beta_2, \end{aligned} \tag{2}$$

where $(x_n^{(3)}, y_n^{(3)})$ and $(x_n^{(4)}, y_n^{(4)})$ are detected dot in LDS scanning coordinates $o^{(3)}$-$x^{(3)}y^{(3)}z^{(3)}$ and $o^{(4)}$-$x^{(4)}y^{(4)}z^{(4)}$; θ is the angle between $y^{(3)}$-axis and the line that connects origin $o^{(3)}$ and detected dot; θ' is the angle between $y^{(4)}$-axis and the line that connects origin $o^{(4)}$ and detected dot; $(u_n^{(3)}, v_n^{(3)})$ is the coordinate value of detected dot in the new coordinate $u^{(3)}o^{(3)}v^{(3)}$ and $(u_n^{(4)}, v_n^{(4)})$ is the coordinate value of detected dot in the new coordinate $u^{(4)}o^{(4)}v^{(4)}$ as well.

After transformation, the scanned lines in two different coordinates, $u^{(3)}o^{(3)}v^{(3)}$ and $u^{(4)}o^{(4)}v^{(4)}$, need to be merged into one coordinate. We define the coordinate $u^{(3)}o^{(3)}v^{(3)}$ as tread profile base coordinate uov and move all the data from $u^{(4)}o^{(4)}v^{(4)}$ into uov by (3), as shown in Figure 4(b). Hence,

$$\begin{aligned} u_n &= u_n^{(4)} + \Delta u, \\ v_n &= v_n^{(4)} + \Delta v, \end{aligned} \tag{3}$$

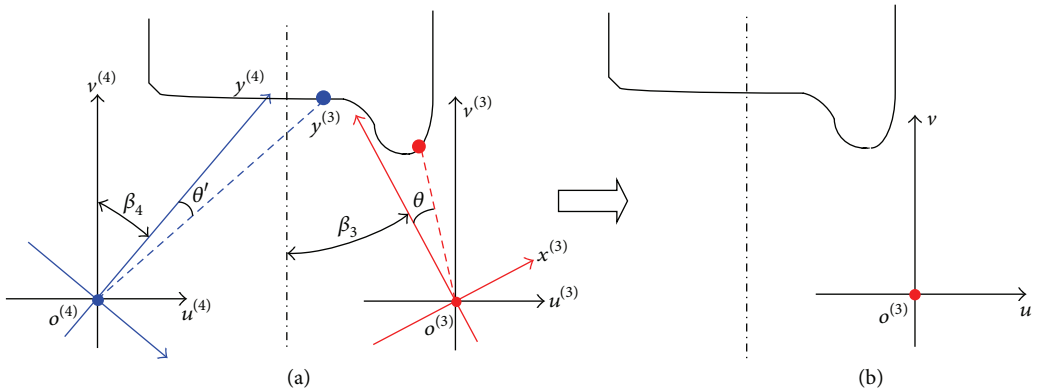


FIGURE 4: (a) Coordinate transformation of L4 and coordinate transformation of L3. (b) Moving all the data from $u^{(4)}o^{(4)}v^{(4)}$ to uov.

where (u_n, v_n) is the dot in tread profile base coordinate uov; Δu and Δv are the offset from $u^{(4)}o^{(4)}v^{(4)}$ to uov.

As we know, flange width, flange height, and wheel diameter are determined by several base points and base line. The output points from sensors are discrete, so base points **a**, **m**, and **n** are more likely not in one of the scanned points. The output points are also polluted with sensor noise, which induced more detection uncertainty when we directly regard it as the base points. Over here, curve fitting is used for extracting the base point, as well as the base line. Through this method, the coordinate value of base points can be precisely extracted and the sensor noise can also be eliminated to some extent. It is difficult to use a single curve to fit all the tread due to the complexity of tread contour. Therefore, fitting discrete points of each base point within a certain range is applied to improve the accuracy of the extracted base point coordinate value. The common method of curve fitting is the least square method [22]. The least square method uses a given set of measured data to get the functional relation $f(x, a_0, a_1, \ldots, a_n)$ between the variable x and the variable y based on the principle of least squares. Then, the weighted sum of squares' value of the residual e_k between the fitting function and the actual measured value at each point can be minimal, which means F in (4) is minimal:

$$F = \sum_{i=0}^{I} \omega(x_i)\left(f_i - y_i\right)^2, \tag{4}$$

where $\omega(x_i) \geq 0$ is the weight which reflects the notion that the data (x_i, y_i) accounts for the proportion in the experiment; I denotes the number of data points. According to the tread profile features and experimental research, fourth-order polynomial $y = \sum_{i=0}^{4} a_i x^{4-i}$ is selected to fit each subsection curve based on the least square method.

With curve fitting technique, four lines in total are fitted in order to extract the coordinate value of base points **a**, **m**, and **n**. As shown in Figure 5, at first, the inner side of the wheel has no wear-out and deformation when there is wheel-rail contact, so base line l is fitted by selecting all the data points in the inner side of the wheel. The base point **a** is 70 mm away from the base line l along u-axis. Then, the green line is fitted in order to extract base point **a** by selecting data points within

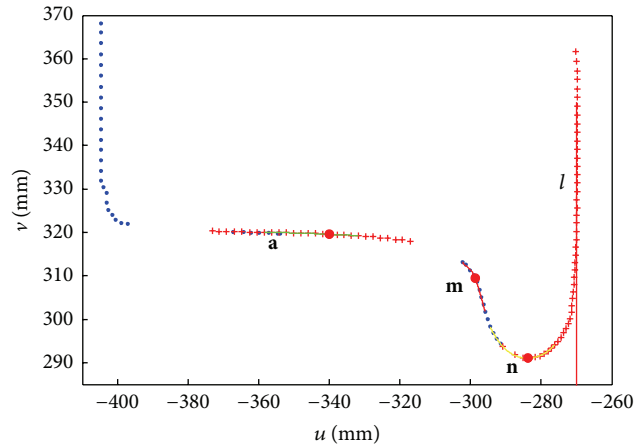


FIGURE 5: Curve fitting results.

a certain range of base point **a**. The red line and yellow line are also fitted by the same method in order to extract base points **m** and **n**, respectively.

After four lines are obtained, the precise coordinate value of all base points can be determined. To this end, the flange height and flange width are calculated as follows:

$$\begin{aligned} F_w &= u_l - u_m, \\ F_h &= v_n - v_a, \end{aligned} \tag{5}$$

where F_w is flange width; F_h is flange height; u_l is the u-axis coordinate value of base line l; u_m is the u-axis coordinate value of base point **m**; v_n is the v-axis coordinate value of base point **n**; v_a is the v-axis coordinate value of base point **a**.

2.4. Static Wheel Diameter Calculation Principle. Wheel diameter is detected by 2D-L1, 1D-L2, and 2D-L3. Each one of the LDS measures one point in the circular wheel so that the wheel diameter can be determined by three points.

The wheelset is assumed to be in the right position and to remain static. The coordinates of the output points are in the scanning coordinate system of the sensor. Similar to tread profile calculation, the coordinate transformation

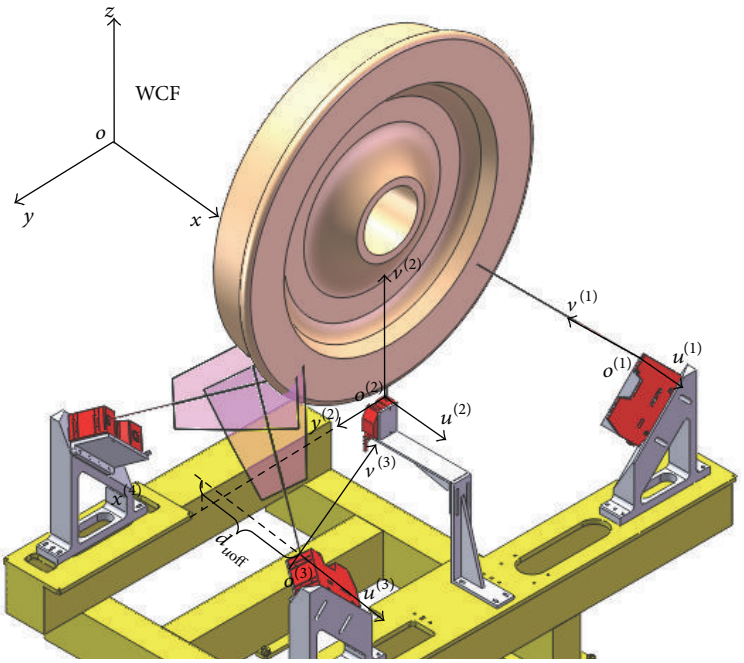


FIGURE 6: Scanning coordinates of 2D-L1, 1D-L2, and 2D-L3 after coordinate transformation.

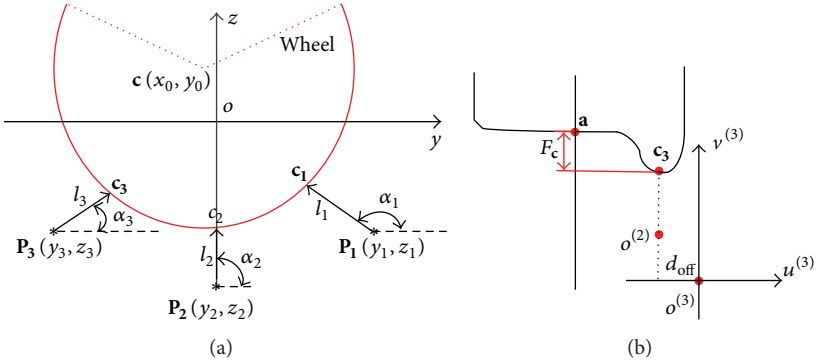


FIGURE 7: Wheel diameter calculation principle in two dimensions: (a) sight along the x-axis; scanning coordinates of 2D-L1, 1D-L2, and 2D-L3 after transformation; (b) sight in the $u^{(3)}o^{(3)}v^{(3)}$ coordinate.

was conducted and the scanning coordinates $u^{(1)}o^{(1)}v^{(1)}$, $u^{(2)}o^{(2)}v^{(2)}$, and $u^{(3)}o^{(3)}v^{(3)}$ for 2D-L1, 1D-L2, and 2D-L3, respectively, have been obtained. Figure 6 shows the scanning coordinates of 2D-L1, 1D-L2, and 2D-L3 after coordinate transformation. Notice that $u^{(2)}o^{(2)}v^{(2)}$ is still the same as $y^{(2)}o^{(2)}z^{(2)}$ because of the installation position of 1D-L2. Figure 6 also shows the offset d_{off} between the origin of the coordinate $u^{(3)}o^{(3)}v^{(3)}$ and laser scanning line of L2 in $u^{(2)}$-axis. Among three points, the two points detected by 2D-L1 and 2D-L3 are extracted from the 2D profiles. The offset d_{off} is the $u^{(2)}$-axis coordinate value to extract the points in the flange circle from two-dimensional profile. This offset d_{off} is determined by sensor installation.

Figure 7 shows the wheel diameter calculation principle in two dimensions where (a) shows the principle that three points determine a diameter in yoz WCF and (b) shows extracting the point in the flange circle among two-dimensional profile and the final wheel diameter distance subtraction by F_c. From Figure 7(a), the installation of each LDS is modeled as three parameters in yoz WCF, which are the position $\mathbf{P_i}(y_i, z_i)$ and angle α_i. They determine the position of laser origin and the direction of detection, respectively. The angle α_2 for 1D-L2 is designed as $\pi/2$. The positions $\mathbf{P_1}$ and $\mathbf{P_3}$ are designed as symmetric with respect to the scanning line of 1D-L2, as well as the angles α_1 and α_3. Even though many of the parameters are designed to be equal, for instance, $z_1 = z_3$, the real parameters will vary after engineering implementation due to errors such as installation error and manufacturing error. Thus, this nine-parameter model is proposed for diameter calculation

because it can describe all the possible errors. The real installation parameters are obtained through calibration, as later described. Moreover, the distances l_1, l_2, and l_3 are detected from three LDS sensors. The three points $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$ are in the flange of wheel detected by three sensors, respectively. The point $\mathbf{c}(y_0, z_0)$ is the origin of the detected wheel which is calculated by three points $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$.

According to Figures 6 and 7(a), the first information we can get from the LDS sensors is the laser scanned distances l_1, l_2, and l_3. l_2 is directly detected by 1D-L2. l_1 and l_3 are extracted from the 2D profiles detected by 2D-L1 and 2D-L3, respectively. To extract l_1 and l_3, we need to find the correct points in the two-dimensional tread profile. As shown in Figures 6 and 7(b), the point that determines l_1 and l_3 should be in the u-axis value of offset d_{off}. Similar to detecting tread profile, we use the same curve fitting method to obtain a curve in the contour of the wheel in the uov coordinate that is denoted by $v = f_3(u)$. When the curve line is obtained, $l_3 = v_3 = f_3(d_{\text{off}})$, namely, the v-axis value of curve $f_3(u)$ when $u = d_{\text{off}}$. Similarly, l_1 is detected by 2D-LDS L1 using the same method as deciding l_3.

Once the laser scanned distances l_1, l_2, and l_3 are determined, we get the three points $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$ in WCF coordinate yoz by

$$\begin{aligned} y_{c_1} &= y_1 + l_1 \sin\alpha_1, \\ z_{c_1} &= z_1 + l_1 \cos\alpha_1, \\ y_{c_2} &= y_2 + l_2 \sin\alpha_2, \\ z_{c_2} &= z_2 + l_2 \cos\alpha_2, \\ y_{c_3} &= y_3 + l_3 \sin\alpha_3, \\ z_{c_3} &= z_3 + l_3 \cos\alpha_3. \end{aligned} \tag{6}$$

Based on three points $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$, the wheel center $\mathbf{c}(y_0, z_0)$ is determined by

$$\begin{aligned} y_0 &= \frac{\left(z_{c_1} - z_{c_3}\right)\left(y_{c_1}^2 - y_{c_2}^2 + z_{c_1}^2 - z_{c_2}^2\right) - \left(z_{c_1} - z_{c_2}\right)\left(y_{c_1}^2 - y_{c_3}^2 + z_{c_1}^2 - z_{c_3}^2\right)}{2\left(y_{c_1} - y_{c_2}\right)\left(z_{c_1} - z_{c_3}\right) - 2\left(y_{c_1} - y_{c_3}\right)\left(z_{c_1} - z_{c_2}\right)}, \\ z_0 &= \frac{\left(y_{c_1} - y_{c_2}\right)\left(y_{c_1}^2 - y_{c_3}^2 + z_{c_1}^2 - z_{c_3}^2\right) - \left(y_{c_1} - y_{c_3}\right)\left(y_{c_1}^2 - y_{c_2}^2 + z_{c_1}^2 - z_{c_2}^2\right)}{2\left(y_{c_1} - y_{c_2}\right)\left(z_{c_1} - z_{c_3}\right) - 2\left(y_{c_1} - y_{c_3}\right)\left(z_{c_1} - z_{c_2}\right)} \end{aligned} \tag{7}$$

and the wheel diameter D_r is determined by

$$D_r = 2 \cdot \sqrt{\left(y_0\right)^2 + \left(z_0 - z_{c_2}\right)^2}. \tag{8}$$

From Figure 7(b), F_c is the distance between points $\mathbf{c_3}$ and $\mathbf{a}$ along v-axis. The wheel diameter detected by the previous three points is somewhere in the contour circle governed by the 1D-L2 only. The point $\mathbf{a}$ is considered to be the diameter point of the wheel which is −70 mm away from the inner side of the wheel. In order to obtain the final wheel diameter, we need to further subtract the distance F_c from the wheel diameter:

$$D = D_r - 2F_c, \tag{9}$$

where F_c is the distance between point $\mathbf{c_3}$ and point $\mathbf{a}$ in v-axis (as shown in Figure 7(b)); namely, $F_c = v_a - v_{c_3}$.

2.5. Dynamics Detection. The calculation principles shown above are in static case. When the train passes dynamically, multiscans can be obtained and the misalignment phenomenon caused from wheel passing will occur.

For tread profile detection, ideally, the laser light panel of L3 and L4 is assumed to include the center of the measured wheel. In dynamics detection, it is impossible to meet that assumption for all measured wheels due to the moving of wheel and the discrete sampling of LDS signal. Basically, if the laser light panel does not include the center of the measured wheel, the detected profile is horizontally stretched along v-axis. That will lead to the increase of detected flange height and flange width. This phenomenon is called the misalignment between the laser panel and the detection target [9].

Figure 8 shows the dynamics positions of the wheel center and laser panel constituted by L3 and L4 in two-dimensional WCF. The wheel moves forward with a constant speed of v. O_1, O_2, and O_i are the center of wheel diameter circle under different positions. The laser panel has an installation angle α_3 with respect to y-axis, which can be determined as $z = \tan\alpha_3 \cdot y$ in WCF. The center points of wheel diameter circle are calculated by (7). d_i denote the distance between ith wheel center point and the laser panel.

Theoretically, for every ith position of the wheel, the distance d_i from the wheel center $\mathbf{c}(y_{0_i}, z_{0_i})$ to the laser panel can be determined by point to the distance formula as follows:

$$d_i = \left| \frac{\tan\left(\alpha_3\right) y_{0_i} - z_{0_i}}{\sqrt{\tan^2\left(\alpha_3\right) + 1}} \right|. \tag{10}$$

When the distance d_i equals zero, the wheel center $\mathbf{c}(y_{0_i}, z_{0_i})$ is in the laser panel where the flange height and flange width have no stretching. On the other hand, the bigger the distance d_i is, the farther away the wheel center $\mathbf{c}(y_{0_i}, z_{0_i})$ is from the laser panel.

It is worth mentioning that the LDS works when the angle between laser light and detected surface is within a certain range and the angle is influenced by laser wavelength, surface

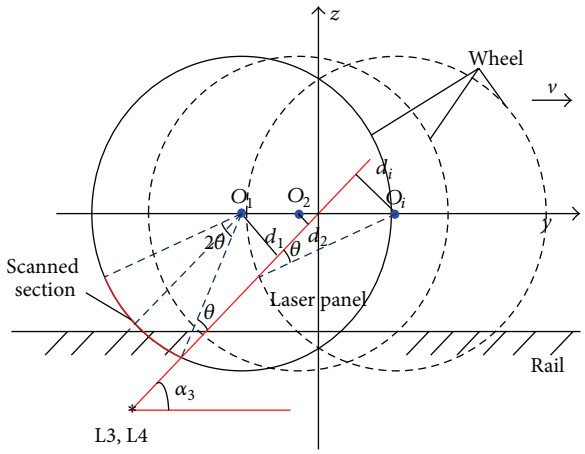


Figure 8: Dynamics position of the wheel center.

smoothness, surface material, and so forth [23]. It is assumed that the angle θ (as shown in Figure 8) is the largest angle at which the LDS can still receive effective scan. When the wheel is moving out of the detection range, the LDS will be unable to scan. Thus, the scanned section will be the arc with a central angle of 2θ, and all the tread profiles and diameters are scanned from this section. For most LDS sensors, the angle θ can reach up to 45°, so the system can measure 90° arc of the wheel. Correspondingly, the maximum value of the distance d_i is $R \sin\theta$, where R is the wheel radius.

The misalignment phenomenon will bring about certain error to the profile detection. Among all the effective scans, we must select those scans where the induced error is acceptable. In this paper, the error induced in the tread profile detection is analyzed in Section 3. As a result, the error is directly proportional to the distance d_i. So, we set up a certain threshold Kp. When the distance $d_i < Kp$, the detected tread profile can be regarded as useful profiles where the error induced by the misalignment phenomenon is negligible. The threshold Kp is firstly obtained through error analysis and also is adjustable according to the field experiment. Due to the benefits from the high sampling frequency of the LDS sensors, M times of scans can be obtained for a wheelset. Then, we can remove the bulky error first and perform average operation to get the final wheel flange and wheel width as follows:

$$F_{w_f} = \frac{1}{M}\sum_{i=1}^{M} F_{w_i},$$
$$F_{h_f} = \frac{1}{M}\sum_{i=1}^{M} F_{h_i}, \tag{11}$$

where F_{w_i} and F_{h_i} are the flange width and flange height in ith scan, respectively; F_{w_f} and F_{h_f} are the final flange width and flange height, respectively. The average operation here can reduce the final error caused by Gaussian sensor noise.

For wheel diameter detection, the three points that determine the wheel diameter are always in the contour circle. Thus, the calculation results will not be influenced by different wheel positions. However, F_{c_i} in every ith scan will still be stretched and bring about some error. Similarly, we select a set of scans by comparing whether the distance d_i is smaller than a certain threshold Kd or not. When $d_i < Kd$, the error induced in the detected F_c is negligible. The two thresholds, Kd and Kp, might be different because of the different detection error requirements for tread profile and wheel diameter. In this way, N times of scans can be obtained. Then, we can remove the bulky error first and perform average algorithm to get the final wheel diameter as follows:

$$D_f = \frac{1}{N}\sum_{i=1}^{N} D_i, \tag{12}$$

where D_i is the wheel diameter in ith scan; D_f is the final wheel diameter.

2.6. Calibration. The measuring and calculating of tread profile and wheel diameter depend on many installation parameters. Regarding tread profile calculation, they are the angle β_3 in (1), angle β_4 in (2), and the offset Δu and offset Δv in (3). For wheel diameter, they are the offset d_{off} between the origin of the coordinate uov and laser scanning line of L2 in $u^{(2)}$-axis, the angles α_1, α_2, and α_3, and the positions $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P_3}$ in (6). When the LDS are installed and fixed, it is impossible for those parameters to be the same with designed values because of the manufacture error of mechanical parts and installation accuracy. So, calibration is certainly needed.

During the calibration process for tread profile detection, a standard wheel is placed on the rail over the detection system, and then the offset and rotation angle of the coordinate transformation matrix can be determined. In terms of the angles β_3 and β_4, the calibrated accurate value is to make sure the inner and outer panels of the wheel are vertical. For the offset Δu, the calibrated accurate value is to make sure the detected wheel hub thickness equals the standard wheel hub thickness and the offset Δv is to make sure the scanned profiles from two LDS coincide with each other.

As for the calibration process for wheel diameter, a set of new ground wheelsets is used. The ground wheelset is with different diameters that are 770 mm, 790 mm, 810 mm, and 840 mm. We set the minimization function $f(x)$ as the squared summation of detected diameters subtracted by real diameter. That is,

$$\min \quad f(x) = \sum_{i=1}^{J} \left|D_i - D_r\right|^2, \tag{13}$$

where D is the detected diameter according to wheel diameter calibration principle; D_r is the real diameters; $x = [d_{\text{off}}, \alpha_1, \alpha_1, \alpha_3, y_1, z_1, y_2, z_2, y_3, z_3]$ is the variables to be calibrated; J denotes the number of ground wheelsets. MATLAB has provided such tools to solve those optimization problems. Over here, we use *fmincon* function to minimize the function and the constraints in terms of the variables are also given according to real physical ranges. Finally, the optimal values of the parameters can be obtained. These values are assumed to be the real values that the implemented system has and have been further put into use in system service.

3. Detection Error Analysis

In this section, we consider four factors which are rail vibration, sensor noise, misalignment, and wheel inclination caused by wheel S-shape running and the differential of diameters.

3.1. Wheel-Rail Vibration. Wheel-rail vibration is the first factor that we considered. In our system, all the sensors are well fixed by the mechanical support and mechanical pedestal that has no direct contact with rail. So, the wheel-rail vibration will not directly transmit to the sensors and, instead, the wheel-rail vibration has to transmit to the ground of the depot and then transmit to the sensors through mechanical support and mechanical pedestal. The vibration of the ground is on a lower level, the maximum of acceleration is only $0.4\,m^2/s$ [24] in Guangzhou metro depot, and it is also attenuated by the mechanical pedestal. We also measured the maximum of acceleration of mechanical support during train passing, which is only $0.2\,m^2/s$. So, the change of the position of the sensors due to the wheel-rail vibration in our system can be neglected. Furthermore, all the laser sensors are capturing data simultaneously and the exposure time of the LDS is within 50 microseconds. The vibration of the wheel will not cause considerable movement within such a short time. Overall, the system is assumed to be reliable against wheel-rail vibration.

3.2. Misalignment. As previously mentioned in Section 2, Dynamics Detection, if the laser light panel does not include the center of the measured wheel, the detected profile is horizontally stretched along v-axis. This phenomenon is called the misalignment between the laser panel and the detection target, which will lead to the increase of the detected flange height and flange width. Chen et al. [10] derived a geometric model regarding how many errors will be generated for flange height when wheel position varies. The error e of the flange height is

$$e = \sqrt{R_C^2 - d^2} - \sqrt{R^2 - d^2} - R_C + R, \tag{14}$$

where R is the wheel radius; R_C is the radius in the wheel rim; d is the distance from the wheel center to the laser panel, as described in Section 2.

On the basis of this geometric model, when we know how much the error of the flange height is, the errors of the flange width can be derived accordingly. For different wear wheels, the profiles, as well as the fitted line for lateral contact point **m**, are certainly different. To illustrate the massiveness of errors, here we chose the same wheel where the fitted line for lateral contact point **m** is $v = f(u)$. We obtain the inverse function $u = g(v)$ and stretch it horizontally by a factor of $(F_h + e)/F_h$. So, the stretched curve line is

$$u = g_2(v) = g\left(\frac{vF_h}{(F_h + e)}\right). \tag{15}$$

Eventually, the error of flange width is $\eta = g_2(10) - F_w$, where F_w is the original flange width.

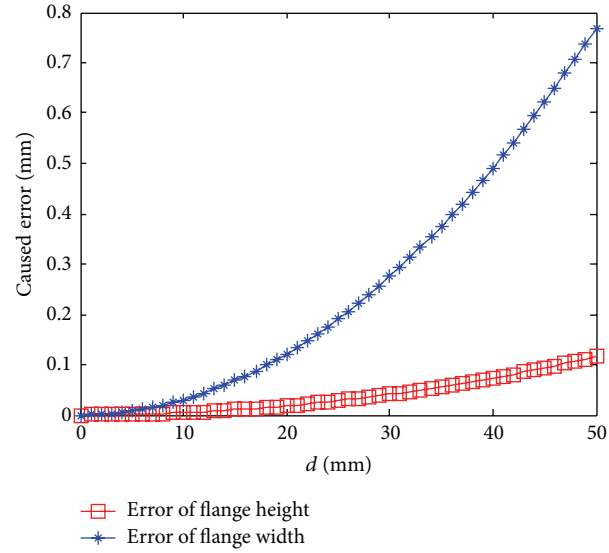


Figure 9: The error of flange width and flange height with respect to various wheel positions.

Theoretically, from (14) and (15), we know that the smaller the value of R is, the larger the error e is. So, we chose the largest standard wheelset with $R = 385$ mm and $R_C = 399$ mm. Figure 9 shows the error of flange height and flange width induced from misalignment in this case. The distance d varies from 0 mm to 50 mm with an interval of 1 mm. From Figure 9, the error of flange height is lower than the error of flange width. So, we focus on the error of flange width here.

In our system, the sampling frequency of all LDS is 1 kHz and the maximum speed of the train in the depot is 36 km/h. The maximum of sampling step size along the rail $\Delta s = 1\,\text{ms} \times 10\,\text{m/s} = 10$ mm. When we set the threshold Kp (as described in Section 2.5) as 20 mm, the total measuring distance along the rail can be 58 mm so that at least $M = 5$ times of efficient scans can be detected. The corresponding errors are less than 0.1 mm for flange width after taking the average of these 5 efficient scans. Thus, the system can perform detection normally against the misalignment error benefitting from the high sampling frequency.

3.3. Sensor Noise. The LDS cannot be ideally accurate. The measuring accuracy is influenced by temperature, the roughness of the measured surface, and so forth.

In order to obtain the quantitative influence for profile detection, we built a 3D model in SolidWorks tools and extracted ideal sensor output points of standard inner and outer tread profiles. In this model, the standard wheel is located in the position where the center of the wheel is in the laser panel. So, the misalignment phenomenon will not affect tread profile detection. The wheel is in static position so the simulated sensor output points are all from one scan. Moreover, the parameters that need to be calibrated are ideally accurate. To imitate the real situation, Gaussian noise is added to these coordinate values. The mean of noise is zero, and the standard deviation is varied from 0 to 1 mm with an

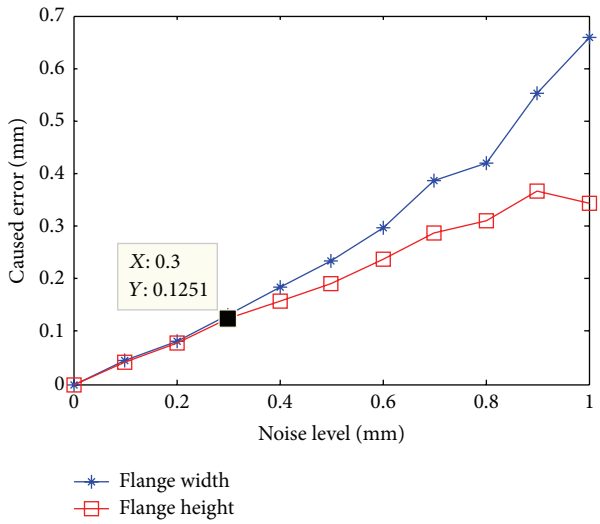


Figure 10: The RMS error of flange height and flange width caused by different sensor noise level.

interval of 0.1 mm. For each noise level, 500 experiments are carried out and the RMS error is calculated. The RMS error of flange height and flange width results caused by different noise level is shown in Figure 10. The caused error to flange height and flange width is approximately half of the sensor noise level. This can be explained by the curve line fitting method that has taken more laser points into account and thus has reduced the random noise. Because the flange height is determined by two points, the curve line fitting method has at least reduced the random error into a quarter of the original sensor noise.

The 2D-LDS we chose is LJ-V7300 from @KEYENCE which has a full-scale resolution of 0.1%/F.S and a temperature drift of 0.01% F.S/°C. The detection range in y-axis is 300 ± 145 mm and in x-axis is 110 mm to 240 mm, which formed as a trapezoid. The point in x-axis is fixed; thus, only sensor noise in y-axis needs to be considered with full scale of 290 mm. So, accordingly, the RMS error caused to the profile coordinate noise in y-axis, which is denoted by δ, is less than 0.32 mm, which only leads to an error of 0.13 mm both to flange height and to flange width. Taking dynamics detection effect into account, the final error is reduced by $\delta_f = \delta/\sqrt{5} = 0.058$ mm with at least $N = 5$ times of efficient scans. The error caused by sensor noise can be acceptable.

Regarding the error of wheel diameter, it can be theoretically derived by the theorem of error propagation [25]. The resolution of each sensor is denoted by δ_1, δ_2, and δ_3. We obtain δ_D by taking differential of (6)–(9) as follows:

$$\delta_D = \pm\sqrt{\left(\delta_1\frac{\partial D}{\partial l_1}\right)^2 + \left(\delta_2\frac{\partial D}{\partial l_2}\right)^2 + \left(\delta_3\frac{\partial D}{\partial l_3}\right)^2}. \quad (16)$$

We have chosen two 2D-LDS and one 1D-LDS to detect the wheel diameter and the two 2D-LDS are installed symmetrically. For systematic installation, we have $\delta_1(\partial D/\partial l_1) = \delta_3(\partial D/\partial l_3)$. Moreover, the analytical function of particle derivative will be too complex to derive. So, we consider a special case where

$$\begin{aligned}&[d_{\text{off}}, \alpha_1, \alpha_1, \alpha_3, y_1, z_1, y_2, z_2, y_3, z_3]\\ &= [10\text{ mm}, 45^\circ, 90^\circ, 135^\circ, -495\text{ mm},\\ &-495\text{ mm}, 0\text{ mm}, 600\text{ mm}, 495\text{ mm}, -495\text{ mm}],\end{aligned} \quad (17)$$

where the target wheel diameter is $D = 840$ mm and the origin of the wheel is located in the origin of yoz WCF. More calculation details can be found in the Appendix. Finally, we have

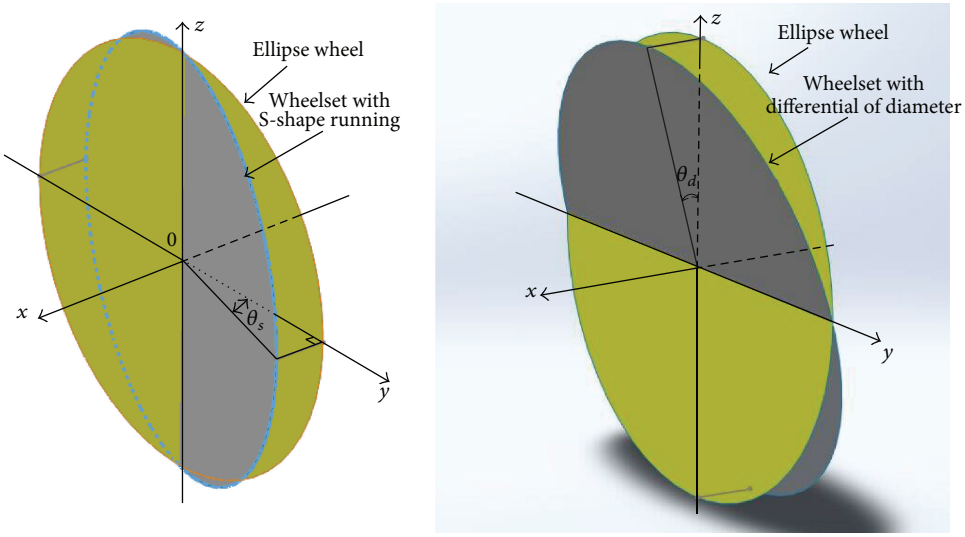
$$\begin{aligned}\frac{\partial D}{\partial l_1} &= -3.4142,\\ \frac{\partial D}{\partial l_2} &= 4.8284.\end{aligned} \quad (18)$$

The 1D-LDS we chose is LK-G8085 from @KEYENCE which has linearity of 0.05%/F.S and a temperature drift of 0.01% F.S/°C. So, according to the full scale of 30 mm, the resolution of 1D-LDS $\delta_2 = 0.018$ mm. Based upon the finding that the curve line fitting method has at least reduced the random error into a quarter of the original sensor noise, $\delta_1 = 0.075$ mm. Finally, δ_D is less than 0.372 mm. Taking dynamics detection effect into account, the final error $\delta_{D_f} = \delta_D/\sqrt{5} = 0.17$ mm. The error caused by sensor noise can be acceptable.

3.4. Wheel Inclination Caused by Wheelset S-Shape Running and Differential of Wheel Diameter. In engineering, the wheel will be inclined because of wheelset S-shape running and the differential of wheel diameter. The wheelset S-shape running is one kind of self-induced vibration due to the slope in the wheel trade. When it is S-shape running, the wheel panel will have a certain angle with respect to yoz panel in WCF, denoted by θ_s as shown in Figure 11(a). The differential of wheel diameter in a wheelset is at different wear level in the left and right wheel, mainly induced from different massiveness of wear in the circuit of wheelset turning and unbalanced loading. Similarly, it will bring a certain angle about the wheel panel with respect to the yoz panel in WCF. The angle is denoted by θ_d as shown in Figure 11(b).

For wheel diameter detection, because we only consider the calculation in two dimensions, an error will be generated when we still regard the detected three points in a circle to actually be in an ellipse. Considering the existence of angles θ_s and θ_d, we have the equation of ellipse as follows:

$$\begin{aligned}\frac{y^2}{(R\cdot\cos\theta_s)^2} + \frac{z^2}{R^2} &= 1,\\ \frac{y^2}{R^2} + \frac{z^2}{(R\cdot\cos\theta_d)^2} &= 1.\end{aligned} \quad (19)$$



(a) (b)

FIGURE 11: Mathematical illustration. (a) Wheel S-shape running and (b) differential of wheel diameter.

Similar to when we analyze sensor noise, we consider a special case as (17); the origin of the target wheel is located in the origin of *yoz* WCF. The real three points are

$$
\begin{aligned}
c_1&:\left(\frac{R\cdot\cos\theta_s}{\sqrt{(\cos\theta_s)^2+1}},-\frac{R\cdot\cos\theta_s}{\sqrt{(\cos\theta_s)^2+1}}\right),\\
c_2&:(0,-R),\\
c_3&:\left(-\frac{R\cdot\cos\theta_s}{\sqrt{(\cos\theta_s)^2+1}},-\frac{R\cdot\cos\theta_s}{\sqrt{(\cos\theta_s)^2+1}}\right),\\
c_1&:\left(\frac{R\cdot\cos\theta_d}{\sqrt{(\cos\theta_d)^2+1}},-\frac{R\cdot\cos\theta_d}{\sqrt{(\cos\theta_d)^2+1}}\right),\\
c_2&:(0,-R\cdot\cos\theta_d),\\
c_3&:\left(-\frac{R\cdot\cos\theta_d}{\sqrt{(\cos\theta_d)^2+1}},-\frac{R\cdot\cos\theta_d}{\sqrt{(\cos\theta_d)^2+1}}\right).
\end{aligned}
\tag{20}
$$

Theoretically, the larger the radius of the wheel is, the bigger the error is. So, we chose $R = 420$ mm and generated three points; then, using (7), we calculated the wheel diameter with error. Subtracting the real diameter, we have the error with respect to angle as shown in Figure 12. The effect of S-shape running caused angle has a relatively higher influence on the wheel diameter calculation.

Based on the experience from Guangzhou Metro Corporation, the differential of diameter in a wheelset should be controlled under 2 mm. Considering the track gauge of 1350 mm, the angle induced from the differential of diameter

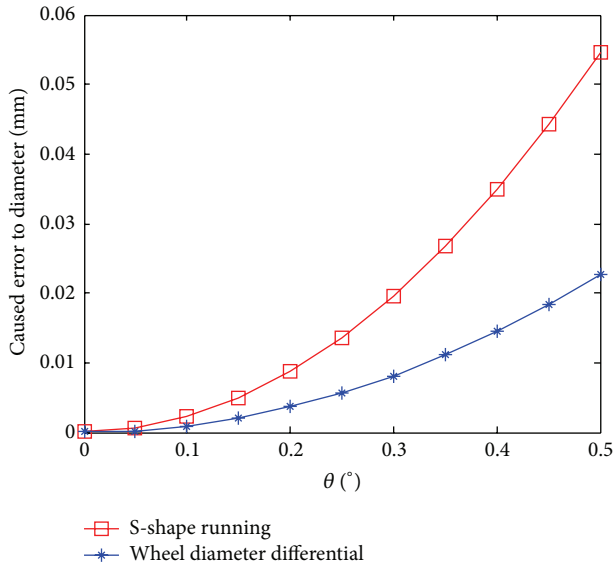


FIGURE 12: Wheel diameter errors.

in a wheelset is less than 0.001°; thus, the error can be ignored. As for wheel S-shape running, the maximum angle is 0.1° [26] when the speed of the train is under 36 km/h which will cause an error not larger than 0.005 mm.

4. Experimental Validation

4.1. System Implementation. The authors previously proposed an online detection system using eight 2D-LDS [19]. The new online detection system is installed in the same storage line of Guangzhou metro vehicle depot as the old system so that comparison can be conducted. In order to save fund, only the left side, namely, half of the system, has been

(a) (b)

Figure 13: Field test. (a) Standard wheel test and (b) real train test.

Table 1: Standard wheelset detection, the results of repeated measurements/mm.

Measurement times	Flange height		Flange width		Wheel diameter	
	Old	New	Old	New	Old	New
1	28.04	28.16	31.98	32.08	839.76	839.79
2	28.11	28.21	32.06	32.01	839.96	840.15
3	27.99	28.15	32.01	32.04	840.08	839.86
4	28.05	28.18	31.85	32.10	839.88	840.43
5	28.08	28.16	31.93	32.02	840.01	840.04
6	28.11	28.10	32.06	32.14	839.98	840.46
Mean	**28.06**	**28.16**	**31.98**	**32.06**	**839.95**	**840.12**
SD	**0.046**	**0.036**	**0.078**	**0.052**	**0.111**	**0.281**

implemented. During the system implementation, three-dimensional inclinometer and special rail gauge are used to control the position of the mechanical support. After the system is installed, the calibration described in Section 2 has been conducted to obtain the geometric parameters for tread profile calibration and diameter calculation. As shown in Figure 13, the field test is carried out by a standard wheelset and real train.

4.2. Standard Wheelset. The standard wheelset is a new produced wheelset without any wear and diameter differential. The manufacturing geometric size is as follows: wheel diameter = 840 mm, flange height = 28 mm, and flange width = 32 mm. One can also assume to have lower possibility of S-shape running because of zero external load. The standard wheelset has been placed on the rail and passed through the detection system. This test has been carried out 6 times to verify both the detection and the repeatability of the system. Comparing with the old system, the results of this system are shown in Table 1.

From Table 1, the mean values of the flange height and flange width detected by the old and the new system are very close to each other, which means the system error can be ignored. The standard deviation, which also can be denoted as detection uncertainty, of the new system measurement is slightly smaller than of the old system. That may result from the lower effect from misalignment, as described in Section 3, due to the higher sampling frequency that we used in the new system. Detection uncertainty of not greater than 0.05 mm in tread profile measurement is acceptable for the engineering requirements. As for wheel diameter detection, the mean values are also close to each other. The standard deviation of the new system measurement is slightly higher than of the old system. This may result from the replacement of 1D-LDS that has brought about higher sensor noise to the middle point among three points without curve fitting technique. However, detection uncertainty of less than 0.3 mm is also acceptable in engineering.

4.3. Real Train Detection Test. Real train test also performs 6 times of repeated detection to statistically evaluate the performance of the system. The train speed is controlled under 36 km/h. In the train we chose, there are 4 new ground wheelsets in a car of the train. Under the consideration that the ground new wheel is not out of roundness which has an effect on the analysis results, we selected the ground new wheel as our target wheel.

Table 2 shows the mean and standard deviation value of measurement. The biggest differential value of mean flange height appears in #1 wheel and for mean flange width appears in #3 wheel. The difference does not exceed 0.15 mm. As for wheel diameter, the biggest differential value 0.16 mm appears in #3 wheel. The mean value of six times of repeated detection is consistent with the standard wheelset test. In terms of standard deviation, the value is less than 0.1 mm for flange width and flange height and 0.3 mm for wheel diameter. The standard deviation of wheel diameter is relatively higher than in standard wheelset test. On the contrary, the standard deviation of flange width and flange height is relatively lower than in the old system. That is also consistent with standard

Table 2: Real train test, the mean and standard deviation value of detected measurement/mm.

Wheel number	Mean flange height		SD flange height		Mean flange width		SD flange width		Mean wheel diameter		SD wheel diameter	
	Old	New	Old	New	Old	New	Old	New	Old	New	Old	New
1	28.18	28.03	0.046	0.060	29.54	29.43	0.091	0.062	800.52	801.50	0.201	0.301
2	28.09	28.11	0.078	0.040	29.40	29.29	0.056	0.028	801.12	800.96	0.128	0.286
3	27.97	27.91	0.076	0.033	29.92	30.06	0.075	0.056	801.87	801.66	0.090	0.179
4	28.07	28.05	0.063	0.053	29.83	29.88	0.076	0.088	801.78	802.01	0.192	0.282

Table 3: Real train test, wheel #2, the result of repeated measurement/mm.

Measurement times	Flange height		Flange width		Wheel diameter	
	Old	New	Old	New	Old	New
1	27.98	28.15	29.42	29.31	801.07	801.40
2	28.14	28.11	29.36	29.31	801.06	800.97
3	28.12	28.13	29.47	29.24	801.27	800.53
4	28.10	28.12	29.31	29.29	801.06	800.87
5	28.00	28.03	29.39	29.28	800.96	800.96
6	28.18	28.11	29.39	29.32	801.28	801.10
Mean	**28.09**	**28.11**	**29.40**	**29.29**	**801.12**	**800.96**
SD	**0.078**	**0.040**	**0.056**	**0.028**	**0.128**	**0.286**

Table 4: Real train test, wheel #3, the result of repeated measurement/mm.

Measurement times	Flange height		Flange width		Wheel diameter	
	Old	New	Old	New	Old	New
1	27.90	27.90	29.78	29.99	801.95	801.59
2	28.06	27.97	29.97	30.02	801.81	801.74
3	28.02	27.90	29.91	30.05	801.96	801.48
4	27.99	27.90	29.98	30.09	801.78	801.48
5	28.00	27.87	29.94	30.15	801.95	801.94
6	27.86	27.90	29.96	30.06	801.78	801.74
Mean	**27.97**	**27.91**	**29.92**	**30.06**	**801.87**	**801.66**
SD	**0.076**	**0.033**	**0.075**	**0.056**	**0.090**	**0.179**

wheelset test. The standard deviation of wheel diameter in real train test is supposed to be higher than in the standard wheelset test because of several assumptions. One factor is the higher possibility of S-shape running because of heavy axial load. On the other hand, the wheelset that is in service is also more polluted with rust than standard wheelset, causing more detection uncertainty. However, the standard deviation from real train test also does not exceed 0.3 mm, which is consistent with standard wheelset test. This may result from the lower train speed during the test which leads to lower possibility of S-shape running. Meanwhile, the rusty wheel contour is also not in a massive stage. Tables 3 and 4 show the result of repeated measurement for wheels numbers 2 and 3, respectively. In each detection, the results remain the same and no gross error appears.

Overall, detection uncertainties for tread profile and wheel diameter are less than 0.1 mm and 0.3 mm, respectively. The results show that the detection system has a high accuracy, which can meet the requirements of maintenance operation.

5. Conclusion

This paper, based on LDS, proposed a novel on-track detection system of the wheel size using only six 2D-LDS and two 1D-LDS. Errors induced by wheel-rail vibration, sensor noise, misalignment, S-shape running, and wheelset differential are also analyzed. After the system is implemented, real data experiments including standard wheel test and real train detection test were performed. It turns out that the detection uncertainty of flange width and height is 0.1 mm and wheel diameter 0.3 mm, which can meet the requirements of maintenance. This system can be further used for different types of railway transportation, which provides a new solution for the wheel size detection technology.

Appendix

We consider a special case where

$$\begin{aligned}&\left[d_{\text{off}}, \alpha_1, \alpha_1, \alpha_3, y_1, z_1, y_2, z_2, y_3, z_3\right]\\&\quad = \left[10\text{ mm}, 45^\circ, 90^\circ, 135^\circ, -495\text{ mm},\right.\\&\quad \left.-495\text{ mm}, 0\text{ mm}, 600\text{ mm}, 495\text{ mm}, -495\text{ mm}\right].\end{aligned} \tag{A.1}$$

To provide more benefits, the target wheel diameter is $D = 840$ mm and the origin of the wheel is located in the origin of yoz WCF, as shown in Figure 14. In this special case, the relevant geometric values are $c_1(-198.02\text{ mm}, -198.02\text{ mm})$, $c_2(0\text{ mm}, 420\text{ mm})$, $c_3(198.02\text{ mm}, -198.02\text{ mm})$, $l_1 = 280$ mm, $l_2 = 180$ mm, and $l_3 = 280$ mm.

According to (8) and (9), we get the particle derivative as follows:

$$\begin{aligned}\frac{\partial D}{\partial l_1} &= \frac{\partial D}{\partial y_0}\frac{\partial y_0}{\partial l_1} + \frac{\partial D}{\partial z_0}\frac{\partial z_0}{\partial l_1},\\ \frac{\partial D}{\partial l_2} &= \frac{\partial D}{\partial y_0}\frac{\partial y_0}{\partial l_2} + \frac{\partial D}{\partial z_0}\frac{\partial z_0}{\partial l_2} + \frac{\partial D}{\partial z_{c_2}}\frac{\partial z_{c_2}}{\partial l_2}.\end{aligned} \tag{A.2}$$

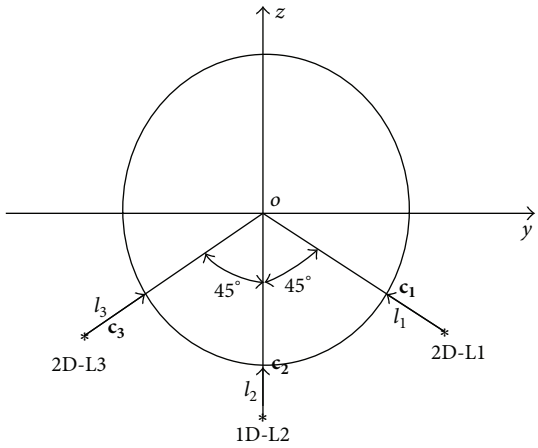


Figure 14: A special case.

Taking the derivative of diameter D with respect to y_0, z_0, and z_{c_2} according to (8) and substituting $(y_0, z_0) = (0, 0)$ and $z_{c_2} = 180$ mm, we have

$$\frac{\partial D}{\partial y_0} = \frac{2y_0}{\sqrt{(y_0)^2 + (z_0 - z_{c_2})^2}} = \frac{1}{180\,\text{mm}},$$

$$\frac{\partial D}{\partial z_0} = \frac{2(z_0 - z_{c_2})}{\sqrt{(y_0)^2 + (z_0 - z_{c_2})^2}} = \frac{1}{180\,\text{mm}}, \tag{A.3}$$

$$\frac{\partial D}{\partial z_{c_2}} = \frac{-2(z_0 - z_{c_2})}{\sqrt{(y_0)^2 + (z_0 - z_{c_2})^2}} = 2.$$

Furthermore, based upon (7), we get

$$\begin{aligned}
\frac{\partial y_0}{\partial l_1} &= \frac{\partial y_0}{\partial y_{c_1}}\frac{\partial y_{c_1}}{\partial l_1} + \frac{\partial y_0}{\partial z_{c_1}}\frac{\partial z_{c_1}}{\partial l_1},\\
\frac{\partial z_0}{\partial l_1} &= \frac{\partial z_0}{\partial y_{c_1}}\frac{\partial y_{c_1}}{\partial l_1} + \frac{\partial z_0}{\partial z_{c_1}}\frac{\partial z_{c_1}}{\partial l_1},\\
\frac{\partial y_0}{\partial l_2} &= \frac{\partial y_0}{\partial z_{c_2}}\frac{\partial z_{c_2}}{\partial l_2},\\
\frac{\partial z_0}{\partial l_2} &= \frac{\partial z_0}{\partial z_{c_2}}\frac{\partial z_{c_2}}{\partial l_2},\\
\frac{\partial y_0}{\partial l_3} &= \frac{\partial y_0}{\partial y_{c_3}}\frac{\partial y_{c_3}}{\partial l_3} + \frac{\partial y_0}{\partial z_{c_3}}\frac{\partial z_{c_3}}{\partial l_3},\\
\frac{\partial z_0}{\partial l_3} &= \frac{\partial z_0}{\partial y_{c_3}}\frac{\partial y_{c_3}}{\partial l_3} + \frac{\partial z_0}{\partial z_{c_3}}\frac{\partial z_{c_3}}{\partial l_3}.
\end{aligned} \tag{A.4}$$

When calculating particle derivative of (y_0, z_0) with respect to three points c_1, c_2, and c_3 in WCF, we assume that all parameters are with the geometric values in this special case. Then, we substitute the ideal geometric values of this variable, and we obtain

$$\begin{aligned}
\frac{\partial y_0}{\partial y_{c_1}} &= 0.5,\\
\frac{\partial y_0}{\partial z_{c_1}} &= 0.5,\\
\frac{\partial z_0}{\partial y_{c_1}} &= -1.2071,\\
\frac{\partial z_0}{\partial z_{c_1}} &= -1.2071,\\
\frac{\partial y_0}{\partial z_{c_2}} &= 0,\\
\frac{\partial z_0}{\partial z_{c_2}} &= 3.4142,\\
\frac{\partial y_0}{\partial y_{c_3}} &= 0.5,\\
\frac{\partial y_0}{\partial z_{c_3}} &= -0.5,\\
\frac{\partial z_0}{\partial y_{c_3}} &= 1.2071,\\
\frac{\partial z_0}{\partial z_{c_3}} &= -1.2071,\\
\frac{\partial y_{c_1}}{\partial l_1} &= \frac{1}{\sqrt{2}},\\
\frac{\partial z_{c_1}}{\partial l_1} &= \frac{1}{\sqrt{2}},\\
\frac{\partial z_{c_2}}{\partial l_2} &= -1,\\
\frac{\partial y_{c_3}}{\partial l_3} &= -\frac{1}{\sqrt{2}},\\
\frac{\partial z_{c_3}}{\partial l_3} &= \frac{1}{\sqrt{2}}.
\end{aligned} \tag{A.5}$$

Finally, substituting (A.5) into (A.4) and then substituting (A.4) and (A.3) into (A.2), we have

$$\begin{aligned}
\frac{\partial D}{\partial l_1} &= -3.4142,\\
\frac{\partial D}{\partial l_2} &= 4.8284.
\end{aligned} \tag{A.6}$$

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research was carried out under the National Key Research and Development Plan of China (2016YFB1200402), the Science and Technology Program of Guangzhou (201508010010), and the Fundamental Research Funds for the Central Universities (AE89454). The fund is greatly acknowledged. Special thanks are due to Mr. Jie Jiang for his help in 3D design in SolidWorks.

References

[1] Y. Chen, Z. Xing, J. Li, and Y. Qin, "The analysis of wheel-rail vibration signal based on frequency slice wavelet transform," in *Proceedings of the 17th IEEE International Conference on Intelligent Transportation Systems (ITSC '14)*, pp. 1312–1316, Qingdao, China, October 2014.

[2] R. Pohl, A. Erhard, H.-J. Montag, H.-M. Thomas, and H. Wüstenberg, "NDT techniques for railroad wheel and gauge corner inspection," *NDT & E International*, vol. 37, no. 2, pp. 89–94, 2004.

[3] The International Union of Railways, *UIC 510-2 Code. Trailing Stock: Wheels and Wheelsets. Conditions Concerning the Use of Wheels of Various Diameters*, The International Union of Railways, Paris, France, 2004.

[4] Z. Zhang, C. Lu, F. Zhang, Y. Ren, K. Yang, and Z. Su, "A novel method for non-contact measuring diameter parameters of wheelset based on wavelet analysis," *Optik*, vol. 123, no. 5, pp. 433–438, 2012.

[5] Web-1, 2016, https://www.greenwood.dk/miniprofwheel.php.

[6] S. O. Medianu, G. A. Rimbu, D. Lipcinski, I. Popovici, and D. Strambeanu, "System for diagnosis of rolling profiles of the railway vehicles," *Mechanical Systems and Signal Processing*, vol. 48, no. 1-2, pp. 153–161, 2014.

[7] Web-2, http://www.mermecgroup.com/inspection-technology/train-monitoring/87/1/wheel-profile-and-diameter.php.

[8] Web-3, http://iem.net/freight-rail-40478?id=150.

[9] Web-4, 2016, http://www.kldlabs.com/index.php?s=wheel+profile+measurement.

[10] X. Chen, J. Sun, Z. Liu, and G. Zhang, "Dynamic tread wear measurement method for train wheels against vibrations," *Applied Optics*, vol. 54, no. 17, pp. 5270–5280, 2015.

[11] Z. Gong, J. Sun, and G. Zhang, "Dynamic structured-light measurement for wheel diameter based on the cycloid constraint," *Applied Optics*, vol. 55, no. 1, pp. 198–207, 2016.

[12] Z. F. Mian, J. C. Mullaney, R. MacAllister, and T. J. Schneider, "Optical wheel evaluation," U.S. Patent No. 7,564,569, 2009.

[13] Y. Gao, S. Shao, and Q. Feng, "A new method for dynamically measuring diameters of train wheels using line structured light visual sensor," in *Proceedings of the International Symposium on Photonics and Optoelectronics (SOPO '12)*, pp. 1–4, IEEE, Shanghai, China, May 2012.

[14] Z.-F. Zhang, Z. Gao, Y.-Y. Liu et al., "Computer vision based method and system for online measurement of geometric parameters of train wheel sets," *Sensors*, vol. 12, no. 1, pp. 334–346, 2012.

[15] A. N. Baibakov, K. I. Kuchinskii, V. I. Paterikin, S. V. Plotnikov, and V. V. Sotnikov, "Experience in developing and using automated laser diagnostic equipment for the contactless monitoring of the parameters of freight car wheels," *Measurement Techniques*, vol. 53, no. 4, pp. 444–448, 2010.

[16] Yu. N. Dubnishchev, P. Y. Belousov, O. P. Belousova, and V. V. Sotnikov, "Optical control of the radius of a wheel rolling on a rail," *Optoelectronics, Instrumentation and Data Processing*, vol. 48, no. 1, pp. 75–80, 2012.

[17] Y. Gao, Q. Feng, and J. Cui, "A simple method for dynamically measuring the diameters of train wheels using a one-dimensional laser displacement transducer," *Optics and Lasers in Engineering*, vol. 53, pp. 158–163, 2014.

[18] K. Wu and J. Chen, "Dynamic measurement for wheel diameter of train based on high-speed CCD and laser displacement sensors," *Sensor Letters*, vol. 9, no. 5, pp. 2099–2103, 2011.

[19] Z. Zhang, Z. Su, Y. Su, and Z. Gao, "Denoising of sensor signals for the flange thickness measurement based on wavelet analysis," *Optik—International Journal for Light and Electron Optics*, vol. 122, no. 8, pp. 681–686, 2011.

[20] Z. Xing, Y. Chen, X. Wang, Y. Qin, and S. Chen, "Online detection system for wheel-set size of rail vehicle based on 2D laser displacement sensors," *Optik*, vol. 127, no. 4, pp. 1695–1702, 2016.

[21] CN-TB, "Tread profile for locomotive and car," 2003.

[22] A. Ravindran, K. M. Ragsdell, and G. V. Reklaitis, *Engineering Optimization: Methods and Applications*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2006.

[23] T. J. Ko, J. W. Park, H. S. Kim, and S. H. Kim, "On-machine measurement using a noncontact sensor based on a CAD model," *The International Journal of Advanced Manufacturing Technology*, vol. 32, no. 7-8, pp. 739–746, 2007.

[24] C. Zou, Y. Wang, P. Wang, and J. Guo, "Measurement of ground and nearby building vibration and noise induced by trains in a metro depot," *Science of the Total Environment*, vol. 536, pp. 761–773, 2015.

[25] A. J. Wheeler and A. R. Ganji, *Introduction to Engineering Experimentation*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 2010.

[26] A. Qin, M. Su, and Y. Yao, "Influence of hunting wave to lateral vibration of deck steel plate bridges," *Journal of Shijiazhuang Railway Institute*, vol. 20, no. 1, pp. 56–60, 2007.